EAGLES

Handbook on Spoken Language Systems

DRAFT — WORK IN PROGRESS

EAGLES DOCUMENT SLWG Phase 2, V1

Version of 18 May 1995

Contents

1 Introduction 1

2 System design 1

2.1 Introduction 1

2.2 System capability profile versus application requirement profile 2

2.3 Technical features versus system capabilities 2

2.4 Speech recognition systems 2

2.4.1 Speaker dependency 2

2.4.2 Speaker adaptive 2

2.4.3 Speaking aspects 2

2.4.3.1 Speaking mode 3

2.4.3.2 Speaking fluency 3

2.4.3.3 Speaking rate 3

2.4.4 Non-speech sounds 3

2.4.5 Vocabulary aspects 3

2.4.6 Lexicon size 3

2.4.7 Speech training material 3

2.4.8 Lexicon generation 3

2.4.9 Field speech data (in-situ recording) 3

2.4.10 Branching factor (Perplexity factor) 3

2.4.11 Rejection mode 4

2.4.12 Cut through / Anticipation 4

2.4.13 Application's vocabulary and confusion matrix 4

2.4.14 Language modelling 4

2.4.15 Channel adaptation / Environment adaptation 4

2.4.16 Task adaptation 4

2.5 Signal module 4

2.6 Speaker verification/identification 4

2.6.1 Speaker verification versus speaker identification 4

2.6.2 Text dependent / Text independent 4

2.6.3 Sentences defined by the technology provider 5
2.6.4 Sentences defined by the application developer 5
2.6.5 Training carried out by the technology provider 5
2.6.6 Training carried out by the application developer 5
2.7 Speech synthesis 5
2.7.1 Speech recording, storage, and playback 5
2.7.2 Speech concatenation 5
2.7.3 Text to speech synthesis 5
2.7.3.1 The linguistic part 5
2.7.3.2 Phonetic part 5
2.7.3.3 Acoustic module 6
2.7.4 Multi-linguality 6
2.8 Interactive voice systems 6
2.9 Software aspects 6
2.9.1 Operating systems 6
2.9.2 Drivers 6
2.9.3 Application programming interfaces 6
2.9.4 Application generators 6
2.10 Hardware aspects 6
2.10.1 Platforms 6
2.10.2 Connectivity 7
2.10.3 Characteristics 7
2.11 Multi-linguality aspects 7
2.12 Conclusions 7

3 Corpus design **1**
3.1 Introduction 1
3.1.1 About this chapter 1
3.2 Seven main differences between collections of written NL- and spoken SL-data 2
3.2.1 Durability of text, volatility of speech 3
3.2.2 Different production times for text and speech 3
3.2.3 Correcting errors in the production of text and speech 3
3.2.4 Orthographic identity and phonetic variability of lexicalised units 3
3.2.5 Printable ASCII-strings and continuously sampled speech 4
3.2.6 Size differences between NL- and SL-data 5
3.2.7 The different nature of categories and time functions 5
3.3 Applications of spoken language corpora 5
3.3.1 Speech corpora for research purposes 6

3.3.1.1 Phonetic research 6
3.3.1.2 Sociolinguistic research 6
3.3.1.3 Psycholinguistic research 7
3.3.1.4 First language acquisition 7
3.3.1.5 Second language acquisition 8
3.3.1.6 General linguistic research 8
3.3.1.7 Audiology 9
3.3.1.8 Speech pathology 9
3.3.2 Speech corpora for technological applications 9
3.3.2.1 Speech synthesis 10
3.3.2.2 Speech recognition 10
3.3.2.2.1 Knowledge-based *vs.* stochastic systems 11
3.3.2.2.2 Speaker-independent *vs.* speaker-dependent systems 11
3.3.2.2.3 Isolated words *vs.* continuous speech 11
3.3.2.2.4 Corpora for speech recognition research 12
3.3.2.3 Spoken language systems 12
3.3.2.4 Speaker recognition/verification 12
3.4 Specification of the linguistic content 13
3.4.1 Different types of speech data 13
3.4.1.1 Read aloud isolated phonemes 13
3.4.1.2 Read aloud isolated words 13
3.4.1.3 Read aloud isolated sentences 14
3.4.1.4 Read aloud text fragments 15
3.4.1.5 Semi-spontaneous speech 15
3.4.1.6 Spontaneous speech about a predetermined subject 16
3.4.1.7 The Wizard of Oz technique 17
3.4.1.8 Spontaneous speech 17
3.4.2 Factorial experiments and corpus studies 18
3.5 Specification of number and type of speakers 19
3.5.1 Corpus size in terms of speakers 19
3.5.1.1 Speech corpora with few speakers 19
3.5.1.2 Speech corpora with about 5 to 50 speakers 19
3.5.1.3 Speech corpora with more than 50 speakers 20
3.5.1.4 General remarks 20
3.5.2 Speaker characteristics 20
3.5.2.1 Stable ↔ Transient speaker characteristics 21
3.5.2.2 Demographic coverage 21

3.5.2.3 Male/Female . . . 22
3.5.2.3.1 Pitch and intensity . . . 22
3.5.2.3.2 Overall spectral slope . . . 22
3.5.2.3.3 Accuracy of pronunciation . . . 22
3.5.2.3.4 Vocabulary and syntax . . . 23
3.5.2.4 Age . . . 23
3.5.2.4.1 Voice quality . . . 23
3.5.2.4.2 Vocabulary and syntax . . . 23
3.5.2.5 Weight and height . . . 23
3.5.2.6 Smoking and drinking habits . . . 24
3.5.2.7 Pathological speech . . . 24
3.5.2.8 Professional ↔ Untrained speakers . . . 25
3.5.2.9 Geographical and sociolinguistic factors . . . 25

4 Corpus collection **1**
4.1 Introduction . . . 1
4.2 Data collection dimensions . . . 1
4.2.1 Visibility: Open vs. Secret . . . 1
4.2.2 Environment: Studio vs. On Location . . . 2
4.2.2.1 Recording in a studio . . . 2
4.2.2.2 Recording on location . . . 3
4.2.3 Communication Situation . . . 4
4.2.3.1 Face to face . . . 4
4.2.3.2 Isolated . . . 4
4.2.4 Communication Mode . . . 5
4.2.4.1 Dialogue . . . 5
4.2.4.2 Prompted Interview . . . 5
4.2.4.2.1 Text den Os . . . 5
4.2.5 Verification . . . 6
4.2.5.1 On-line verification . . . 6
4.2.5.1.1 Text den Os . . . 6
4.2.5.2 Off-line (or *Post-hoc*) verification . . . 6
4.3 Monitoring . . . 7
4.4 Procedures . . . 7
4.4.1 Equipment . . . 7
4.4.1.1 Microphone . . . 8
4.4.1.1.1 Recommendations . . . 9
4.4.1.2 Amplifier/Processor . . . 9

4.4.1.2.1 Recommendations 9
4.4.1.3 Recording Device 9
4.4.1.3.1 Recommendations 11
4.4.2 Speaker Recruitment 11
4.4.2.0.2 Recommendations 12
4.4.3 Scheduling speakers 12
4.4.3.0.3 Recommendations 13
4.4.4 Speaker Prompting 13
4.4.4.0.4 Recommendations 14
4.4.5 Cost 14
4.4.5.0.5 Recommendations 15
4.4.6 Multi-Channel Recording 16
4.4.6.0.6 Laryngography 16
4.4.6.0.7 Electropalatography 16
4.4.6.0.8 Electromagnetic Articulography 17
4.4.6.0.9 Cineradiography 17
4.4.6.0.10 Air-flow measurements 17
4.4.6.0.11 X-ray microbeam 17
4.4.6.0.12 Nuclear magnetic resonance imaging 18
4.4.6.0.13 Ultrasound imaging 18
4.4.7 Wizard of Oz 18
4.4.8 Legal Aspects 18

5 Corpus representation **1**
5.1 Introduction 1
5.2 Transcription of spoken language corpora 1
5.2.1 The transcription of read vs. the transcription of spontaneous speech 2
5.2.1.1 Transcription of dialogues 2
5.2.2 Levels of transcription 3
5.2.2.1 Orthographic transcription 3
5.2.2.1.1 Reduced word forms 4
5.2.2.1.2 Dialect forms 4
5.2.2.1.3 Numbers 4
5.2.2.1.4 Abbreviations and spelled words 4
5.2.2.1.5 Interjections 5
5.2.2.2 Orthographic transcription of read speech 5
5.2.2.3 Phonemic transcription 5
5.2.2.4 Allophonic transcription 7

5.2.2.5 Phonetic transcription 7

5.2.2.5.1 The CRIL convention 7

5.2.2.6 Prosodic transcription 8

5.2.2.6.1 ToBI transcription system 8

5.2.3 Segmentation and Labelling 9

5.2.3.1 Definition and motivation 9

5.2.3.1.1 Definition 9

5.2.3.1.2 A caveat 9

5.2.3.1.3 Use in speech and language technology research 10

5.2.3.1.4 Use in linguistic research 10

5.2.3.2 Levels of segmentation and labelling 10

5.2.3.2.1 Recording script 10

5.2.3.2.2 Orthographic 10

5.2.3.2.3 Morpho-syntactic 11

5.2.3.2.4 Citation-phonemic 11

5.2.3.2.5 Broad phonetic (or phonotypical) 12

5.2.3.2.6 Narrow phonetic 12

5.2.3.2.7 Acoustic-phonetic 13

5.2.3.2.8 Physical 13

5.2.3.2.9 Non-linguistic phenomena 13

5.2.3.2.10 General recommendations for transcribing 14

5.2.3.3 Manual segmentation 15

5.2.3.4 Automatic and semi-automatic segmentation 16

5.2.3.5 Segmentation and labelling in the VERBMOBIL project 16

5.2.4 Prosodic labelling and annotation 17

5.2.4.1 Definition and motivation 17

5.2.4.1.1 Types of approach to prosodic labelling 17

5.2.4.1.2 Examples of the two types of approach 18

5.2.4.2 The ToBI labelling system 18

5.2.4.3 The MARSEC labelling system 19

5.2.4.4 The IPO approach 19

5.2.4.5 Provisional recommendations 20

5.2.5 Annotation conventions 20

5.2.5.1 Deletions (read text) 21

5.2.5.2 Verbal deletions or corrections (implicitly or explicitly) 21

5.2.5.3 Word fragments 21

5.2.5.4 Unintelligible words 21

5.2.5.5 Hesitations (filled pauses) . . . 21
5.2.5.6 Non-speech acoustic events . . . 21
5.2.6 Recommendations . . . 22
5.3 Storage and design of the data base . . . 23
5.3.1 Data types for a speech data base . . . 23
5.3.2 Storage demands . . . 24
5.3.2.1 Storage medium . . . 24
5.3.2.2 Sampling rates . . . 24
5.3.2.3 Compression . . . 24
5.3.3 How to combine speech and speech-related data . . . 24
5.3.4 Database Management Systems . . . 26
5.3.5 Data model . . . 26
5.3.5.1 Hierarchical data model . . . 27
5.3.5.2 Network data model . . . 27
5.3.5.3 Relational data model . . . 28
5.3.5.3.1 Example . . . 28
5.3.5.3.2 Summary . . . 29
5.3.5.4 Object-oriented data model . . . 30
5.3.5.4.1 Summary . . . 30
5.3.5.5 Deductive data model . . . 30
5.3.5.5.1 Summary . . . 30
5.3.6 Safe storage of data . . . 31
5.3.7 Application-independent storage of data . . . 31
5.3.7.1 Controlled access to data . . . 31
5.3.8 Summary . . . 32

6 Spoken Language Lexica **1**
6.1 Introduction . . . 1
6.1.1 Lexica for spoken language systems . . . 1
6.1.2 Lexical information as properties of words . . . 1
6.1.3 Applications of spoken language lexica . . . 2
6.1.3.1 Types of application for spoken language lexica . . . 2
6.1.3.2 Spoken language lexical databases as a general resource . . . 3
6.1.3.3 Lexica in selected spoken language systems . . . 4
6.1.4 Recommendations on resources . . . 6
6.2 What is a spoken language lexicon? . . . 7
6.2.1 Basic features of a spoken language lexicon . . . 7
6.2.2 Lexical databases and system lexica for spoken language . . . 7

6.2.3 Spoken language and written language lexica 9
6.2.4 Basic lexicographic coverage criteria 10
6.2.5 The lexicon in spoken language recognition systems 11
6.2.6 Recommendations on defining spoken language lexica 11
6.3 Types of lexical information in spoken language lexica 12
6.3.1 Lexicon models and lexical representation 12
6.3.2 A simple sign model for lexical properties 13
6.3.3 Lexical units 15
6.3.3.1 Kinds of lexical unit 15
6.3.3.2 Fully inflected form lexica 16
6.3.3.3 Stem and morph lexica 17
6.3.3.4 The notion of 'lexical lemma' 18
6.3.4 Lexical properties and lexical relations in spoken language 18
6.3.5 Recommendations on types of lexical information 20
6.4 Lexical surface information 20
6.4.1 Orthographic information 20
6.4.2 Pronunciation information 21
6.4.3 Prosodic information 25
6.4.4 Recommendations on lexical surface information 26
6.5 Morphological information 26
6.5.1 Types of morphological information 26
6.5.2 Applications of morphology 27
6.5.3 Recommendations on morphology 30
6.6 Grammatical information 30
6.6.1 Statistical language models 30
6.6.2 Sentence syntax information 32
6.6.3 Recommendations on grammatical information 33
6.7 Lexical content information 34
6.7.1 Lexical semantic information 34
6.7.2 Pragmatic information 34
6.7.3 Idiomatic information 34
6.7.4 Recommendations on semantic information 35
6.8 Lexicon structure 35
6.8.1 Spoken language lexicon formalisms 36
6.8.2 Lexicon architecture and lexical database structure 37
6.8.3 The architecture of spoken language system lexica 38
6.8.4 The structure of lexical databases 39

6.8.4.1 A simple database type: pronunciation tables 40
6.8.4.2 A more complex lexical database 42
6.8.5 Recommendations on lexicon structure 44
6.9 Lexical knowledge acquisition for spoken language 44
6.9.1 Stages in lexical knowledge acquisition 45
6.9.2 Types of knowledge source 46
6.9.2.1 Dictionaries 46
6.9.2.2 Corpora 46
6.9.2.3 Acquisition tools 46
6.9.3 Recommendations on lexicon construction 47
6.10 Outlook 48

7 Language models **1**

8 Dialogue **1**

9 Physical characterisation **1**

10 Assessment methodologies and experimental design **1**
10.1 Introduction 1
10.1.1 How to read this chapter 1
10.1.2 Role of statistical analysis and experimentation in Language Engineering Standards 2
10.2 Statistical and experimental procedures for analysing data corpora 3
10.2.1 Statistical analysis 4
10.2.2 Population/s, samples and other terminology 4
10.2.3 Sampling 4
10.2.4 Biasses 4
10.2.5 Estimating sample means, proportions and variances 5
10.2.5.1 Estimating means 5
10.2.5.2 Estimating proportions 6
10.2.5.3 Estimating variance 7
10.2.5.4 Ratio of sample variances 7
10.2.6 Hypothesis testing 8
10.2.6.1 Simple hypothesis testing 8
10.2.6.2 Analysis of variance 9
10.2.6.3 Non-parametric tests 10
10.2.7 Experimental procedures 10
10.2.8 Experimental selection of material to employ 11

10.2.9 Segmentation 12
10.2.9.1 How to make segmentations without confounding effects of classification 12
10.2.9.2 Suggested guidelines concerning comparative analysis between human judges or between human judges and automatic algorithms . . 12
10.2.9.3 Recommendation about what judges to use 13
10.2.9.4 Sample size for assessment of performance 13
10.2.10 Classification 13
10.2.10.1 Judges 13
10.2.10.2 Procedural 14
10.2.10.2.1 Limitations of category responses and ways of circumventing 14
10.2.10.2.2 Range effects 15
10.3 Assessing recognisers 16
10.3.1 Baseline performance 16
10.3.2 Progress 16
10.3.3 Functional adequacy and user acceptance 17
10.3.4 Methodology 19
10.3.4.1 Application oriented 19
10.3.4.2 Reference oriented 19
10.3.4.3 Calibrated databases 19
10.3.4.4 Manipulated, artificial and diagnostic databases 19
10.3.5 Experimental design 19
10.3.5.1 Magnitude estimation 20
10.3.5.2 Rank order 20
10.3.5.3 Pairwise comparison 20
10.4 Assessing speaker verification and recognition systems 20
10.4.1 Sampling rare events in speaker verification and recognition systems 20
10.4.2 Employing expert judgments to augment speaker verification and assessment for forensic aspects of speaker verification and recognition 20
10.5 Interactive Dialogue systems 21
10.5.1 WOZ 21
10.5.1.1 Audio-only simulations 22
10.5.1.1.1 Requirements 22
10.5.1.1.2 Subject variables 22
10.5.1.1.3 Wizard variables 22
10.5.1.2 Multimodal 23
10.5.2 Dialogue metrics 23
10.5.2.1 Psycholinguistic metrics 24
10.5.2.2 Acoustic-based measures 25
10.5.2.3 Interruption parameters 26
10.6 Appendix 1 29
10.7 Appendix 2 32

11 Recognition assessment 1

12 Assessment of speaker verification systems 1

12.1 Presentation 1

12.1.1 Speaker classification tasks 1

12.1.2 General definitions 2

12.2 A taxonomy of speaker recognition systems 3

12.2.1 Task typology 3

12.2.1.1 Speaker identification versus speaker verification 3

12.2.1.2 Related tasks 4

12.2.1.3 Types of errors 4

12.2.2 Levels of text-dependence 4

12.2.3 Interaction mode with the user 5

12.2.4 Definitions 5

12.2.5 Examples 6

12.2.5.1 Text-dependent systems 7

12.2.5.2 Fixed-vocabulary systems 7

12.2.5.3 Unrestricted text-independent systems 8

12.3 Influencing factors 9

12.3.1 Speech quality 9

12.3.2 Temporal drift 9

12.3.3 Speech quantity and variety 10

12.3.4 Speaker population size and typology 10

12.3.5 Speaker purpose and other human factors 11

12.3.6 Recommendations 13

12.3.7 Example 14

12.4 Scoring Procedures 16

12.4.1 Notation 16

12.4.1.1 Registered speaker population 16

12.4.1.2 Test impostor population 17

12.4.2 Closed-set identification 18

12.4.2.1 Misclassification rates 18

12.4.2.2 Mistrust rates 19

12.4.2.3 Confidence ranks 21

12.4.2.4 Comments 21

12.4.2.5 Example 22

12.4.3 Verification 22

12.4.3.1 False rejection rates 23

12.4.3.2 False acceptance rates and imposture rates 24
12.4.3.3 Relative unreliability, vulnerability and imitation ability 27
12.4.3.4 Comments 28
12.4.3.5 Example 29
12.4.3.6 Expected benefit 29
12.4.3.7 Threshold setting 30
12.4.3.8 System operating characteristic 31
12.4.3.9 System characteristic modeling 32
12.4.3.10 Example 35
12.4.4 Open-set identification 35
12.4.5 Recommendations 36
12.5 Complementary assessment tools 37
12.5.1 Standard reference systems 37
12.5.2 Reference human test 38
12.5.3 Automatic / human tests 39
12.5.4 Transformation of speech databases 39
12.6 Typology and assessment of applications 40
12.6.1 Applications and products in speaker recognition 41
12.7 Forensic applications 47
12.7.1 Listener method 48
12.7.2 Spectrographic method 48
12.7.3 Semi-automatic method 49
12.7.4 Recommendations 49
12.8 Conclusions 50
12.9 References 52

13 Synthesis assessment **1**
13.1 Introduction 1
13.1.1 What are speech output systems? 1
13.1.2 Why speech output assessment? 2
13.1.3 Users of this chapter 3
13.2 Towards a taxonomy of assessment tasks and techniques 4
13.2.1 Glass box ↔ black box 4
13.2.2 Laboratory ↔ field 6
13.2.3 Linguistic ↔ acoustic 7
13.2.4 Human ↔ automated 7
13.2.5 Judgment ↔ functional testing 8
13.2.6 Global ↔ analytic assessment 8

13.3 Methodology 8
13.3.1 Subjects 9
13.3.2 Test procedures 11
13.3.3 Benchmarks 13
13.3.4 Reference conditions 14
13.3.4.1 Segmental reference conditions 14
13.3.4.2 Prosodic reference conditions 15
13.3.4.3 Voice characteristics reference conditions 16
13.3.4.4 Overall quality reference conditions 16
13.3.5 Comparability across languages 16
13.4 Black box approach 17
13.4.1 Laboratory testing 17
13.4.1.1 Functional laboratory tests 17
13.4.1.2 Judgment laboratory tests 20
13.4.2 Field testing 23
13.4.2.1 Preliminary remarks 23
13.4.2.2 Field tests 23
13.5 Glass box approach 24
13.5.1 Linguistic aspects 24
13.5.1.1 Preprocessing 25
13.5.1.2 Grapheme-phoneme conversion 26
13.5.1.3 Word stress 28
13.5.1.4 Morphological decomposition 28
13.5.1.5 Syntactic parsing 29
13.5.1.6 Sentence accent 29
13.5.2 Acoustic aspects 30
13.5.2.1 Segments 30
13.5.2.1.1 Functions of segments 30
13.5.2.1.2 Segmental tests 31
13.5.2.1.2.1 Segmental tests at the word level 31
13.5.2.1.2.2 Segmental tests at the sentence level 37
13.5.2.2 Prosody 39
13.5.2.2.1 Functions of prosody 39
13.5.2.2.2 Judgment tests of prosody 40
13.5.2.2.3 Functional tests of prosody 42
13.5.2.3 Voice characteristics 43
13.5.2.3.1 Functions of voice characteristics 43

13.5.2.3.2 Voice characteristics tests 44
13.5.2.4 Relationships among tests 45
13.6 Further developments in speech output testing 47
13.6.1 Introduction 47
13.6.2 Long-term strategy: Towards predictive tests 48
13.6.2.1 From human to automated testing 48
13.6.2.2 Predicting functional behaviour from judgment testing 49
13.6.2.3 Predicting global from analytic testing 49
13.6.2.4 Predicting field performance from laboratory testing 50
13.6.3 Linguistic testing: Creating test environments for linguistic interfaces 50
13.6.4 Acoustic testing: Developments for the near future 51
13.6.4.1 Segmental quality testing 51
13.6.4.2 Prosodic quality testing 52
13.6.4.3 Voice characteristics testing 53
13.6.4.4 Overall quality testing 53
13.7 List of recommendations 54
13.8 Appendix 1: Summary of test descriptions 57
13.9 Appendix 2: Evaluation of visual aspects of speech output 65

14 Spoken language interactive systems **1**

15 Tools **1**
15.1 Introduction 1
15.2 Signal theory 1
15.3 Speech research 2
15.4 Computer hardware and software 4
15.5 Conclusion 6
15.6 Appendix: Useful anonymous ftp sites 6

References **1**

A Computer readable alphabets **1**

B SAMPA computer readable phonetic alphabet **1**
B.1 SAMPA 1
B.1.1 Introduction 1
B.1.1.1 Notation issues 1
B.1.1.2 Transcription 1
B.1.1.3 Coding 2
B.1.1.4 Further languages 3

B.1.2 SAMPA: Present status 4
B.1.3 The phonemic notation of individual languages 7
B.1.3.1 Danish 7
B.1.3.2 Dutch 10
B.1.3.3 English 11
B.1.3.4 French 13
B.1.3.5 German 15
B.1.3.6 Greek 16
B.1.3.7 Italian 17
B.1.3.8 Norwegian 19
B.1.3.9 Portuguese 21
B.1.3.10 Spanish 21
B.1.3.11 Swedish 22
B.1.4 Levels of annotation and extension of SAMPA 24
B.1.4.1 SAMPA as a phonemic system 24
B.1.4.2 Detailed phonetic or acoustic annotation 24

C SAM label file formats 1
C.1 SAM label file formats 1
C.1.1 Introduction 1
C.1.2 Speech file and associated description file formats 1
C.1.3 Label file format 1
C.1.3.1 Label-file header format 2
C.1.3.2 Label-file body format 3
C.1.3.3 Label types 3
C.1.3.4 Label File example 3
C.1.4 Files used in a recording session using EUROPEC 5
C.1.4.1 Corpus File: CORPUS.DBF 5
C.1.4.2 Speaker File: SPEAKERS.DBF 7
C.1.4.3 Prompt file 8
C.1.5 Protocol description file 9
C.1.5.1 Codes used for presentation protocol (Protocol Description File) . . 9
C.1.5.2 Example of Protocol Description File 10
C.1.6 Recording conditions file 12
C.1.7 PRIVATE.PEC file 13

D SAM recording protocols **1**

D.1 SAM recording protocols . . . 1

D.1.1 Definition of terms . . . 1

D.1.2 Classification of general strategies for recording and prompting . . . 2

D.1.2.1 Recording mode . . . 2

D.1.3 Prompting style . . . 2

D.1.3.1 Timing strategy . . . 3

D.1.4 Recording protocol . . . 3

D.1.4.1 Microphone . . . 3

D.1.4.2 Other sensors . . . 3

D.1.4.3 Speech data capture . . . 4

D.1.4.4 Recording environment . . . 4

D.1.4.5 Recording mode and prompting style . . . 4

D.1.4.6 Recording control . . . 5

D.1.4.7 Recording procedure . . . 6

D.1.4.8 Integrity checks . . . 7

D.1.4.9 Backup procedures . . . 7

D.1.4.10 Retrieval procedures . . . 7

D.1.4.11 Calibration . . . 8

D.1.4.12 Inter site consistency and recording procedure verification . . . 8

D.1.4.13 Collation of recordings . . . 8

E SAM software tools **1**

E.1 DKISALA (Interactive Semi-Automatic Labelling Software) . . . 1

E.2 ELSA (ESPRIT Labelling System Assessment software) . . . 2

E.3 EUROPEC (European Program d'Enregistrement de Corpus) . . . 2

E.4 PTS (Progiciel de Traitement de Signal) . . . 3

E.5 RESAM . . . 4

E.6 SAMITPRO (SAM Iterative Proportional Fitting) . . . 5

E.7 SAM_REC0 Isolated Word Recogniser . . . 6

E.8 SAM_SCOR (Sam Input Assessment Scoring Software) . . . 7

E.9 SAM_SLM . . . 7

E.10 SAM_SPEX (Speech Parameter Extractor) . . . 8

E.11 SAMTRA (SAM TRanscription Analysis) . . . 9

E.12 SOAP (Speech Output Assessment Package) . . . 10

E.13 PTM (Parametric Test Manager) . . . 10

F EUROPEC recording tool **1**

G EUROM-1 database overview 1

H Polyphone project overview 1

H.1 Introduction 1

H.2 Reusable resources 2

H.3 The Dutch POLYPHONE corpus 2

H.3.1 Recording workstation 2

H.3.2 Speaker selection 3

H.3.3 The speech material 3

H.3.4 Postprocessing 4

H.4 Use of POLYPHONE in application development 4

H.4.1 Train time table information 5

H.4.1.1 Training of the recogniser 5

H.4.1.2 Building a phonemic lexicon 5

H.4.1.3 A model for *yes/no* expressions 6

H.4.1.4 Time and date expressions 6

H.4.2 Phone card services 6

H.4.2.1 Automating card services 7

H.4.3 Speaker verification 7

H.5 Requirements for future corpora 8

I Overview of speech corpora 1

I.1 Overview 1

I.2 Criteria for assessment of the situation of Spoken Language Resources 2

I.2.1 Types and specificities of corpora 2

I.2.2 Actors in speech resource production 3

I.3 Summary of the current situation on a per language basis 4

I.3.1 DANISH (Denmark) 4

I.3.1.1 Existing databases and their actors 4

I.3.1.2 Ongoing projects & new initiatives 5

I.3.1.3 Conclusion 5

I.3.2 DUTCH (The Netherlands) 5

I.3.2.1 Existing databases and their actors 5

I.3.2.2 Ongoing projects & new initiatives 5

I.3.2.3 Conclusion 5

I.3.3 ENGLISH (United Kingdom) 6

I.3.3.1 Existing databases and their actors 6

I.3.3.2 Ongoing projects & new initiatives 6

I.3.3.3 Conclusion 6

I.3.4 FRENCH (France, Belgium, Switzerland) 6

I.3.4.1 Existing databases and their actors 6

I.3.4.2 Ongoing projects & new initiatives 7

I.3.4.3 Conclusion 7

I.3.5 GERMAN (Germany) 7

I.3.5.1 Existing databases and their actors 7

I.3.5.2 Ongoing projects & new initiatives 8

I.3.5.3 Conclusion 8

I.3.6 GREEK (Greece) 8

I.3.6.1 Existing databases and their actors 8

I.3.6.2 Ongoing projects & new initiatives 8

I.3.6.3 Conclusion 8

I.3.7 ITALIAN (Italy) 9

I.3.7.1 Existing databases and their actors 9

I.3.7.2 Ongoing projects & new initiatives 9

I.3.7.3 Conclusion 9

I.3.8 NORWEGIAN (Norway) 9

I.3.8.1 Existing databases and their actors 9

I.3.8.2 Ongoing projects & new initiatives 9

I.3.8.3 Conclusion 10

I.3.9 PORTUGUESE (Portugal) 10

I.3.9.1 Existing databases and their actors 10

I.3.9.2 Ongoing projects & new initiatives 10

I.3.9.3 Conclusion 10

I.3.10 SPANISH (Spain) 10

I.3.10.1 Existing databases and their actors 10

I.3.10.2 Ongoing projects & new initiatives 11

I.3.10.3 Conclusion 11

I.3.11 SWEDISH (Sweden) 11

I.3.11.1 Existing databases and their actors 11

I.3.11.2 Ongoing projects & new initiatives 11

I.3.11.3 Conclusion 11

I.4 General conclusions 12

I.5 Production costs 13

Chapter 1

Introduction

Chapter 2

System design

Abstract: Many speech processing systems are commercially available and several applications have been successfully set up in different areas. In order to generalise their use in Man-Machine interaction applications a "predictive model of performance" of the technology to be used, needs to be derived by the application developers as a function of what they judge as relevant factors. This model of system performance, correlated to the technology specifications, has to be gauged against the ultimate target: the application requirements and user needs.

2.1 Introduction

Spoken language systems are appropriate combination of several modules including recognition of speech input, recognition of speaker identity (verification or identification), speech output generation and synthesis (including speech coding), and/or Man-Machine interaction management.

A simple use of spoken language system consists of recognising speaker utterances, interpreting them with respect to the application, deriving a meaning (or a command) and providing a consequent feedback to the user (may be a speech prompt or a system action). This is illustrated below for a speech input/output dialogue system:

Figure 1

In order to generalise the use of such systems in different Man-Machine interaction contexts, a predictive model of performance (choukri 1988)) needs to be obtained as a function of different identified relevant factors. The definition of those factors has to lead to a set of parameters that can describe a speech processing system. This description has to express two opposed points of view: one of the technology provider (designer) and the other one of the application developer (buyer). The two points of view have to be distinguished.

Designers should give proofs of the performance of their recognisers with a measure of the impact of any change. So the first contribution of this chapter is related to the technology supplier point of view, and aims at providing detailed guidelines for the specification of speech processing systems in order to explicit the operational capabilities offered by the technology. This will allow the technology providers to depict the system performance in a comprehensive way to the application developers.

Buyers need comprehensive information about how each system or device will perform in the specific conditions of their practical application. So the second contribution of this chapter is related to the application developer point of view, aims at providing detailed guidelines to express the requirement of applications that incorporate speech processing systems, in order to explicit the application requirements that should be met by the operational capabilities of the technology. This will allow the application developers to express their needs in a comprehensive way to the technology providers.

The technology specification is complex enough, and has to go beyond the sole numerical value of 99% of recognition accuracy usually announced by the equipment suppliers. This rate depends on numerous parameters. Some of them cannot be easily quantified (lea 1982), pallet 1985), choukri 1987), choukri 1988), moore 1988)). In order to focus on the most relevant parameters one need to adopt a multi-dimensional characterisation of the speech processing system. This characterisation will be called "The system capability profile" (expression firstly introduced by R. Moore).

The application requirement is also a complex phenomena, too complex to be reflected only by a transaction success rate and should also be depicted as a multi-dimensional characterisation [Ref. WP3/WP8 of sundial]. This will be referred to as "The application requirement profile".

The objective of this chapter is to list the major factors that would allow to define the above mentioned multi-dimension spaces and moreover a way to express a matching process between the two spaces. It will consists of 3 tables with keyword entries that concerns the different dimensions seen form the points of view of the technology provider as capabilities or as system features, and from the application developer point of view as requirements.

2.2 System capability profile versus application requirement profile

2.3 Technical features versus system capabilities

2.4 Speech recognition systems

2.4.1 Speaker dependency

2.4.2 Speaker adaptive

2.4.3 Speaking aspects

2.4.3.1 Speaking mode

2.4.3.2 Speaking fluency

2.4.3.3 Speaking rate

2.4.4 Non-speech sounds

2.4.5 Vocabulary aspects

2.4.6 Lexicon size

2.4.7 Speech training material

2.4.8 Lexicon generation

2.4.9 Field speech data (in-situ recording)

2.4.10 Branching factor (Perplexity factor)

2.4.11 Rejection mode

2.4.12 Cut through / Anticipation

2.4.13 Application's vocabulary and confusion matrix

2.4.14 Language modelling

2.4.15 Channel adaptation / Environment adaptation

2.4.16 Task adaptation

2.5 Signal module

2.6 Speaker verification/identification

2.6.1 Speaker verification versus speaker identification

2.6.2 Text dependent / Text independent

2.6.3 Sentences defined by the technology provider

2.6.4 Sentences defined by the application developer

2.6.5 Training carried out by the technology provider

2.6.6 Training carried out by the application developer

2.7 Speech synthesis

2.7.1 Speech recording, storage, and playback

2.7.2 Speech concatenation

2.7.3 Text to speech synthesis

2.7.3.1 The linguistic part

2.7.3.2 Phonetic part

2.7.3.3 Acoustic module

2.7.4 Multi-linguality

2.8 Interactive voice systems

2.9 Software aspects

2.9.1 Operating systems

2.9.2 Drivers

2.9.3 Application programming interfaces

2.9.4 Application generators

2.10 Hardware aspects

2.10.1 Platforms

2.10.2 Connectivity

2.10.3 Characteristics

2.11 Multi-linguality aspects

2.12 Conclusions

Chapter 3

Corpus design

3.1 Introduction

A *spoken language corpus* is defined as any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use of the data in-house, or by scientists in other organisations. This tentative definition excludes a large amount of speech recordings on analogue tapes (sometimes even on disks) and recordings without the annotation and documentation which is necessary to use the recordings effectively. For instance, it is well known that virtually all public broadcasting corporations in Europe maintain an archive of recordings of programmes, including newscasts, reports of events ranging from football matches to royal weddings and funerals. However, in most cases these recordings can only be accessed by the date of the original broadcast, and perhaps also by the type of programme. Only in very rare cases transcripts of the speech material in the recordings are available. This makes it extremely difficult and time consuming to use these data for any type of research. Of course, this does not diminish the value of these recordings for cultural and scientific purposes, but due to the inordinate amount of pre-processing necessary for any type of research they do not qualify as *spoken language corpus* under our definition. In many other respects our definition is very wide and liberal. For instance, a set of computer files containing speech signals, EMG signals and sub- and supraglottal pressure signals measured in two subjects who sustained vowels on different pitch and intensity levels would qualify as a *spoken language corpus*, provided that the files come with suitable annotation and documentation.
To demarcate the topic of spoken language corpora more precisely, we will first discuss the main differences between written language corpora and spoken language corpora.

3.1.1 About this chapter

In this chapter only the *design* of spoken language corpora and the *use* of these corpora are covered. It is *expressis verbis* not the intention of this chapter to give a comprehensive overview of corpora existing worldwide; we even do not intend to give a comprehensive list of all corpora existing in Europe. For an attempt to survey existing speech corpora the reader is referred to Fourcin et al. (1989) and to Appendix E which contains a list of existing public domain spoken language corpora. The most important reason to refrain from repeating Fourcin's exercise is our conviction that many corpora exist which cannot be identified, and the identification of which makes practically no sense, because they are company proprietary. Corpora, tools, and resources in general are not aims of their own, but means to an independently specified purpose. Thus, the eventual specification of a corpus depends in an essential way on the purpose it is intended to serve. Yet, if that purpose is not too limited and provided the corpus is properly documented and annotated, it is quite likely that it will be useful for other, perhaps unrelated research. For the moment there are few, if any, official standards for corpus development. Given the dependence on the research goal, this is

not surprising. For that reason, the recommendations in this chapter concern general aspects and factors that should be considered in designing a corpus, and guidelines for making decisions on these issues.
In the development of a speech corpus, three phases can be distinguished. In the *pre-recording phase* one has to define the application of the corpus. Furthermore, specifications of experiment design, of linguistic content, of number and type of speakers, and of the physical situation must be established. These topics will be covered in this chapter. In the *recording phase* speaker instruction and prompting, experiment and recording control, as well as storage are involved. These topics will be covered in chapter 4. In the *post-recording phase* transcription, corpus lexicon construction, labelling, and database management take place. These topics will be discussed in chapter 5.
In the remainder of this chapter we will focus on the *pre-recording phase*. More specifically, we want to discuss the following steps in the gathering of a speech corpus:

- Defining the application of the corpus.
- Specifying the linguistic content of the corpus.
- Specifying the number and type of speakers.

Before we embark on these discussions, however, it is necessary to elaborate on the differences between Spoken Language Corpora and Written Language Corpora.

3.2 Seven main differences between collections of written NL- and spoken SL-data

Traditionally, linguists and Natural Language Processing (NLP) researchers understood language corpora to consist of written material collected from text sources which already exist and often are available in published form (novels, stage and screen plays, newspapers, manuals, etc.). In this context the term “spoken language text corpora” was used to indicate that the data are not taken from existing texts but that speech had to be written down in some orthographic or non-orthographic form in order to become part of a data collection. However, the differences (and relations) between text- and speech-data are far more complex. There are at least seven important differences, which may not be ignored because they determine relevant properties of the resulting data collections. For future (technological) developments of Spoken Language Processing (SLP) they should be taken into account very seriously.

These seven differences have to do with:

1. The durability of text as opposed to the volatility of speech.
2. The different times it takes to produce text or speech.
3. The different roles errors play in written and spoken language.
4. The orthographic identity of written and the phonetic variability of spoken words.
5. The different data structure of ASCII-strings and sampled speech signals.
6. The two reasons that cause the great difference in the size of NL- and SL-data collections, and
7. The last but also most fundamental distinction (as well as relations) between symbolically specified categories and physically measured time functions.

A closer look at these seven differences between written and spoken data will reveal why the traditional term "Natural Language Processing", NLP, also could well be read as standing for "Non-spoken Language Processing". As it is our goal to call special attention to the relevant differences we will refer to the written language data as NL-data meaning non-spoken language data, and set it in opposition to the term SL-data, the acronym for spoken language data.

3.2.1 Durability of text, volatility of speech

The first distinction may seem to be rather trivial but it must nonetheless be mentioned here because it exercises an influence on specific properties of the collected NL- and SL-data. While text uses to stay on the paper when it is written down, speech is transient. It is the nature of the phonetic facts which speakers create during speech acts that they disappear just in the moment they come into existence. On the other hand, text is always produced with the intention to remain where the writer puts it, be it on paper, stone, or screen.

The first difference (which in the step from speaking to writing has helped our cultural development) explains why to collect SL-data is less trivial than to produce NL-data. The former must be recorded on a tape or a disk to make it accessible for future use.

3.2.2 Different production times for text and speech

Another difference between NL- and SL-corpora is due to the fact that speech data are time functions in a sense text data are not. Whilst a writer may consume any time he wants (or needs) to invest in producing a text, phonetic information is coded and transmitted through syllabically organised sound transitions. Speech must be running in its own natural time with a typical syllable rate of a value between 120/min and 180/min. The time for writing new text is normally much longer than it takes to read it aloud (which does not mean that silent reading and short-hand-writing can be much faster than speaking the text).

3.2.3 Correcting errors in the production of text and speech

In spontaneously spoken language the editing behaviour of the speaker is audible and remains a part of the recorded data. Interruptions, hesitations, repetitions of words (and parts of words), and especially self-repairs are a characteristic feature of naturally spoken language and must be represented in SL-data collections of spontaneous speech. On the other hand, the writer who has even more correcting and editing possibilities in producing a text document, will normally intend to produce a "clean" version. In the final version of the text all corrections which may have been carried out have disappeared; this is especially true for text intended to go into print. Thus it is extremely unusual to collect NL-data which contain all corrected errors, the replacements of words etc. that have occurred during the writing and editing of a given text document. In recording SL-data it has also not been unusual to produce similar clean speech collections. A typical example is so-called laboratory speech which is produced when a speaker who is sitting in a monitored recording room reads a list of prepared text material, and then only the proper reproductions of the individual text items are accepted to enter the data base. Examples of speech corpora collected in this way are EUROM-0 and EUROM-1. Recently, however, interest has shifted towards corpora comprising "real-world" speech, including hesitations, corrections, background noise, etc.

3.2.4 Orthographic identity and phonetic variability of lexicalised units

In correctly written texts any lexical item has just one and only one distinct orthographic form. Thus the words of European languages are easily identified and, with the exception of a few homographs, also well distinguished from each other, and there is usually also only exactly one version of each possible orthographic contextual form of any given word. This is the source of the fourth important difference between NL- and SL- data which often has been neglected because it is rather inconspicuous to the naive speakers and listeners. The spoken versions of orthographically identical word forms show a great phonetic variation in their segmental and prosodic realisation. In most European languages the phonetic form of a given word is in fact extremely variable depending on the context and other well defined intervening variables (such as speaking style and context of situation, strong and weak Lombard effects, etc.). A given word can even totally disappear phonetically, and possibly be reduced to — and only signalised by — some reflection of segmental features in the prosody of the utterance. Most of these inconspicuous variations appear in a narrow phonetic transcription of a given pronunciation.

It makes a great difference whether a word has been uttered in isolation or in a chain of connected speech. It is only in the first case that words show the phonetic form that is described in pronunciation dictionaries. Only if a word is consciously and very carefully produced in isolation we observe the explicit version of its segmental structure. These phonetically explicit forms produced in a careful speaking style are called citation or canonical forms. The segmental structure of so-called citation forms is modified (probably systematically, although very little of the system is understood) as soon as it is integrated into connected speech. Since this fourth difference between NL- and SL-data is so relevant for the design of spoken language corpora, it has also been taken account for in the conventions of the IPA proposed for Computer Representation of Individual Languages (CRIL).

In dealing with SL-data one must be able to know which words the speaker intended to express in a given utterance. This is reflected in the CRIL-convention of the IPA, which is formally described in Appendix A. Here it should be mentioned that an SL-data collection should ideally have at least two different symbolically specified levels which are related to the acoustic speech signal. On the first level the words of the given utterance are identified as lexical units in their orthographic form, and on the second level a broad phonetic transcription of the citation form should be given. This second level may be the result of automatic grapheme-to-phoneme conversion. For large SL-corpora it would cost too much time and too much money to make broad phonetic transcriptions by hand. How the given words have been actually pronounced in a given speech signal must be specified in terms of a narrower phonetic transcription of each individual utterance on a third, optional CRIL-level. This third level can then be directly aligned to the segments or acoustic features of the digital speech signal in the data base. And this information is especially relevant if also multi-sensor data are to be incorporated in SL-data bases.

It will be clear that detailed phonetic transcriptions are extremely expensive, to the extent that they are likely to be prohibitive in large corpora. However, recents attempts using large vocabulary speech recognisers for acoustic decoding of speech show some promise that the process can be automated, at least to the extent that pronunciation variation can be predicted by means of general phonological and phonetic rules.

In addition to phonetic detail on the segmental level, several uses of spoken language corpora may also require prosodic annotation. In this area much work remains to be done to develop commonly agreed annotation systems. Once such systems exist, one may attempt to support annotation by means of automatic recognition procedures.

3.2.5 Printable ASCII-strings and continuously sampled speech

Taken as pure data, any written text consists of strings of printable alphanumerical and other elements coded in 7- or 8-bit ASCII-Bytes. But the resulting NL-strings possess already a characteristic information structure which is not available in the case of primary SL-data. Separated by blanks ASCII-strings are grouped into lexical substrings, and also the proper punctuation of phrases and sentences is an important property of NL-data. Nothing of this type of information can be found in the recordings of primary SL-data since in natural speech there are no ASCII-elements representing full stops, commas, colons, quotation, question, exclamation marks. Recorded SL- data are primarily nothing but digitalised time functions.

Such recorded data are of no use, unless at least the sequence of words is known which is expressed in the recorded speech signal. As long as reliable automatic recognition of unconstrained speech remains an unsolved problem, much additional work has to be invested for producing the required symbolic annotation of the speech data. This explains why the preparation of useful SL-data collection takes so much effort and expense.

3.2.6 Size differences between NL- and SL-data

Comparing the pure extent of storing collected NL- and SL-data gives a great numerical difference. There are two reasons why SL-data require orders of magnitude more storage space than written language corpora. The first one is simply the difference in coding between text and speech. Whereas the ASCII-string of a word like *'and'* needs only three bytes, many more bytes are required as soon as the phonemes of this word are transformed into an acoustic output for storing the AD-converted data. If in the given example we assume that in clear speech the utterance of a three-phoneme-syllable takes about half a second and if we apply an amplitude quantization of 16 bits and a non-stereo hi-fi sampling rate of 48 kHz, the NL/SL-ratio amounts to approximately 1:16000.
The second reason follows from the great variability in the phonetic forms of spoken words. As pointed out above in difference #4, any written text must be reproduced by many speakers in more than one speaking style (at least in a slow, normal and fast speed with low, normal, high voice etc.), if the corpus is intended to reflect all possible sources of variability.
It should be clear that all these severe modifications of the phonetic forms of words in connected speech play no role in NL data collections.

3.2.7 The different nature of categories and time functions

The last difference, and the most important one, must be looked at from two different angles. The first thing to understand is that the relevant category of the data (that determines its collection) is already inherently given in the case of NL, but totally unknown in the case of physically recorded speech. The ASCII-symbols of a given text are elementary categories by themselves, they are directly used to form syntactically analysable expressions for the representation of all the different linguistically relevant categories. Thus relevant categorical information can be directly inferred from categorically given data and their ASCII-representations. In contrast to this NL-situation, the data of a digital speech signal do not signalise any such categories, because they only represent a measured time function without any inherent categorical interpretation. At the present stage in the development of SLP it is even not yet possible to automatically decide whether a given digital signal is a speech signal or not. Therefore, the necessary categorical annotations for SL-data must still be produced by human workers (with the increasing support of semi-automatic procedures).
The second matter that must be considered in judging the different role of categories and time functions in speech technology is that speech signals contain far more relevant information than can be expressed by the pure text which is pronounced within a given utterance. As long as NLP can be restricted to non-spoken language processing the restriction to NL-data does not pose severe problems. But as soon as real speech utterances are to be processed in an information technology application, the other — non- linguistic, but communicatively extremely relevant categories — cannot be ignored. They must be represented in future SL-data collections, and much effort has still to be invested by the international scientific community to deal with all these information-bearing aspects of any given speech utterance.

3.3 Applications of spoken language corpora

As mentioned earlier, speech corpora are always designed for specific purposes. These purposes determine the content and make-up of a corpus. Thus, a speech therapist interested in pathological

speech will collect a completely different corpus than a designer of a telephone response application. For example, in the first case hi-fi-speech recordings are most probably needed in order to study properties of voice quality, whereas in the latter case one should collect speech over the phone which will result in a rather poor quality of the speech.
In this section we will present a — non-exhaustive — list of possible users of speech corpora together with the specific speech corpora they would need. A distinction will be made between corpora for research purposes and those meant for technological applications. Of course, this does not mean that corpora gathered in the one field cannot be used in the other, although there will be differences in the exchangeability of corpora depending on the corpus. It must be clear that we cannot handle all the details of specific corpora, and that we will indicate only some general properties.

3.3.1 Speech corpora for research purposes

The speech corpora needed for scientific purposes can be very diverse. Some researchers may need carefully pronounced lists of words to study a specific hypothesis about speech production; others may want to study samples of the vernacular, the way people speak in their everyday life. In the following paragraphs some of the major scientific fields with interest in spoken language corpora are mentioned.

3.3.1.1 Phonetic research

In phonetic research all aspects of speech are studied. Phonetic experiments often require carefully controlled speech data, especially when basic phenomena, such as coarticulation, have to be studied in a systematic way. In this type of research more often than not the researcher will have no alternative for collecting new data, specifically designed for the investigation at hand. However, in recent years more and more attention is being paid to uncontrolled (or less controlled) forms of speech as well, because one has begun to realise that results obtained for carefully pronounced speech cannot simply be generalised to more casual speech. This type of research, which requires other experimental designs and other statistical test procedures as well, will profit considerably from existing corpora. Moreover, since the corpora that can support this type of research must by necessity be very large, it will be very unlikely that a researcher will have the opportunity to collect new, project-specific corpora.

3.3.1.2 Sociolinguistic research

In sociolinguistic research variation in language use is studied in heterogeneous communities, especially urban ones. Variables of interest are among others age, sex, and social status. Three common methods to gather data in this research field are:

- *By means of written questionnaires*
  Members of the communities of interest might, for instance, have to indicate on a questionnaire how they pronounce specific words, or whether they use certain sociolect variants of words. A large drawback of this method is that many people are not aware of their pronunciation habits, or the sociolect variants they use. Furthermore, people might regard their actual language use as undesirable and vulgar, and pretend they use a more prestigious form of language.
- *By observations of the investigator*
  This strategy was, for instance, used by William Labov to investigate the occurrence of /r/ deletions in New York English (Labov (1972)). He simply wrote down whether his informants pronounced an /r/ or not in specific words. The major drawback of this method is that the data collection is based on the subjective (and possibly biased) observations of a single person. In addition, the phenomenon of interest is only heard once at a possibly unexpected moment

and in a possibly noisy environment (Labov, for instance, did his investigation in department stores).

- *By collecting speech corpora*
  The gathering of speech corpora offers sociolinguists the opportunity to make detailed and reliable analyses of various phenomena of interest. Perceptual evaluation of pronunciation phenomena could be supplemented or replaced with acoustic measurements. This is especially useful when the differences between pronunciation variants are very subtle, as for instance in the case of a slightly varying vowel colour (Labov (1994)).

Dialect research is closely related to sociolect research. In dialect studies variation in language use due to differences in geographical background of speakers is investigated. Since the methods of data collection are similar to the ones used in sociolect research, the remarks made above also apply to dialect research.

3.3.1.3 Psycholinguistic research

Psycholinguistics is a very broad scientific field in which the psychology of language is studied, including language acquisition by children, the mental processes underlying adult comprehension and production of speech, language disorders, etc. Psycholinguistic experiments sometimes involve carefully controlled speech material, for instance in phoneme monitoring or gating experiments. In phoneme monitoring experiments subjects are asked to spot the first occurrence of a specific phoneme in a spoken utterance, and press a button as soon as they have spotted it. The reaction time between the actual occurrence of the phoneme and the subject's response is used to form hypotheses about underlying mental processes. In gating experiments a progressively larger portion of words is presented to listeners, who are asked to predict what the ending will be. Both techniques can be useful to get more insight in the organisation of the mental lexicon (Aitchison (1994)).
Another way to obtain information about the mental lexicon and speech production processes is to study the disfluencies in spontaneous speech. For example, false starts tell us something about the way in which speech is planned and articulated. Also repetitions of words or word fragments give information about the production and representation of speech. For this type of research, spontaneous speech corpora are very interesting. For more information about planning processes of speech see Levelt (1989).
Yet another way to gather cues about the mental lexicon is to study 'slips of the tongue'. Many tongue-slip collectors carry round a small notebook in which they write down errors whenever they hear them, on a bus, at parties, etc. As mentioned in the former section, data acquired in this way is subjective and unreliable. The use of speech corpora containing spontaneous speech samples would be the answer to this problem, but investigations in this research area would only benefit from extremely large spontaneous speech corpora, because the number of slips of the tongue produced in any one hour of spoken speech is fairly small.

3.3.1.4 First language acquisition

Language acquisition by children is subject of investigation in many (sub-)disciplines of a.o. linguistics and psychology. For example, speech of (young) children can be used to investigate (ir)regularities in language (*linguistics*); it can also be used to learn more about the mental organisation of language (*psycholinguistics*); it can be studied in relation with the *sociolinguistic* background of children; or it can be used to gain more insight in basic *phonetic* processes. All these scientific fields would thus benefit from extensive corpora containing speech of children.
Collecting language acquisition corpora is extremely time consuming and expensive, because of the difficulty in transcribing the speech, especially speech of very young children. In (psycho-)linguistics a considerable amount of work has been done to collect and transcribe corpora, and to make them

available to the research community. Presently, only transcriptions are readily accessible (CHILDES, **** add reference ****)
In the case of toddlers only 'spontaneous' speech samples can be obtained. As soon as children get somewhat older, more controlled forms of speech can also be obtained, such as naming pictures or reading texts. Game playing is another way of eliciting quasi-controlled speech.
Speech acquisition corpora must be longitudinal, i.e. the same person must be recorded repeatedly at subsequent stages in his acquisition process.

3.3.1.5 Second language acquisition

Migration between language areas is as old as history, and probably much older. Until recently, migrants were not hindered too much by their lack of (adequate) knowledge and command of the majority language in their new home countries. Now that low-education jobs are becoming increasingly rare in First World countries this situation is changing. Because command of the language is a prerequisite to education, the study of how immigrants learn to control the language of the host country (the 'second' language) has become an important topic in sociolinguistic research. The European Science Foundation, for instance, has sponsored a large scale project on second language learning in several Western European countries. The research was corpus based: large numbers of migrants have been recorded each fortnight during more than one year. Transcripts and audio tapes comprising this corpus are maintained by the Max Planck Institute of Psycholinguistics in Nijmegen.
It is especially important to study second language acquisition of immigrant children in order to find out how this might influence their education progress. In a similar vein, research into the acquisition of the majority language is needed in "second generation children" who grow up in families which still use the language of their country of origin.

Since immigrants form a minority group in the country they reside in, their native language can be strongly influenced by the second language. For the investigation of these so-called language attrition processes special purpose corpora must be (and have been) collected. In this context one must not only think of African and Asian migrants who are living in the U.S.A. or Western Europe, but also of non-Anglo Europeans who moved to the U.S.A., Canada or Australia.

From a psycholinguistic point of view, it is interesting to study how the different lexicons are organised in the minds of bilingual (and multilingual) speakers. For example, the occurrence of 'blends' (combinations of two words, in this case from different languages) shows that words are subconsciously activated in both languages (Green (1986)). Up to now, most of the research into bilingual lexicons has taken the form of controlled experiments (e.g. cross language priming in lexical decision tasks). It is conceivable, though, that large corpora of spontaneous speech of bilinguals could be used to study lexical and syntactic interferences between the languages.

3.3.1.6 General linguistic research

A substantial part of modern linguistic research is based on Chomsky's 'generative paradigm'. The goal of this (so called *mentalistic*) research programme is to eventually understand the *competence* of language users, i.e. their abstract knowledge of the language system. What speakers and hearers actually do, i.e. their *performance*, is usually of less interest to linguists in Chomsky's tradition. The construction of competence models is generally based on introspection and impressionistic ideas about language use. So, in its strictest form mentalistic linguistic research cannot benefit much from speech corpora that contain samples of the performance of language users.
However, most linguist no longer believe that performance can be neglected completely. For one thing, it has been noted that spontaneous speech corpora often contain utterances which would seem implausible (if not impossible) from introspection, but which are perfectly natural and acceptable in context. And conversely, sentences invented to illustrate grammatical points may be implausible

as actual utterances, because it is extremely difficult to imagine a situation in which they would not violate discourse constraints, aspectual perspectives taken on events, etc. (Chafe (1992)). Moreover, only an *integrated* theory of competence and performance would ultimately be able to account for actual language phenomena. In this respect speech corpora are indispensable to fill the gap between a competence grammar and actual language use.
Presently, more and more linguists are starting to realise the importance of linguistic analysis of constructs of larger size than isolated sentences or utterances. *Discourse Analysis* is the branch of Linguistics which is concerned with the analysis of naturally occurring connected spoken or written discourse (Stubbs (1984)). Obviously, Discourse Analysis will profit very much from large corpora of meaningful speech, whether it is conversational or more formal, e.g. in information seeking dialogues.
In Edwards & Lampert (1993) a comprehensive methodology is presented for transcription and coding of discourse data from various perspectives. This book also contains a list of language corpora that might be useful in discourse research.

3.3.1.7 Audiology

Audiology is the scientific study of *hearing*, often including the treatment of persons with hearing defects. A conventional *audiometer* can be used to test the intensity and frequency range of pure tones that the human ear can detect. This instrument can give a rough indication of the degree of hearing loss in hearing-impaired persons. Present day evaluation of hearing includes the use of controlled speech samples to assist in the determination of a patient's communicative capabilities. Interest in the use of speech to measure hearing has been centered around a research orientation and a practical-clinical orientation. The first orientation has resulted in research areas such as experimental phonetics, the effects of various types of distortion on the human speech recognition and speaker identification, etc. The second orientation has led to research in areas such as the effects of hearing loss on the reception of speech, auditory processing, and the effects of modifications in the range of reception of speech. The second area more or less grew out of the research in the first area (O'Neill (1975)).
For speech corpora to be useful in audiology they must be carefully calibrated, establishing performance (e.g. in terms of recognition scores) of non-hearing impaired reference subjects. Audiological test corpora may contain various types of speech stimuli to evaluate normal and disordered hearing acuity. The speech stimuli can consist of isolated phonemes, nonsense words or real words, and also of connected forms of speech. [Reference must be made here to Harvard sentences, Haskins sentences, as well as to existing CD-ROMs containing calibrated speech and audio material (e.g. the CD produced by TNO in Soesterberg).]

3.3.1.8 Speech pathology

In this scientific field various types of pathological speech are studied, ranging from mild disorders such as hoarseness to severe disorders such as aphasia. The aim of most studies of pathological speech is to find therapies that can alleviate or cure the speech disorder of interest. However, phenomena like aphasia can also be subject of psycholinguistic studies, because such language disorders can shed some light on underlying mental processes (Aitchison (1994)). Corpora of pathological speech can thus be very useful.

3.3.2 Speech corpora for technological applications

Technological applications for which speech corpora are needed can be roughly divided in four major classes: speech synthesis, speech recognition, spoken language systems, and speaker recognition/verification. Depending on the specific application, the speech corpora can be very diverse. For example, speech synthesis usually requires a large amount of speech data from one or two speakers, whereas speech recognition often requires a smaller amount of speech data from many speakers. In the following paragraphs the four domains of speech research for technological applications and the speech corpora they need are discussed.

3.3.2.1 Speech synthesis

The seemingly most natural way to synthesise speech is to model human speech production directly by simulating lung pressure, vocal fold vibration, articulatory gestures, etc. However, it turns out to be extremely difficult to determine and control the details of the model parameters in computer simulations. This is the reason that articulatory synthesisers have only been moderately successful in generating perceptually important acoustic features. Yet, modern measurement techniques have allowed the collection of substantial amounts of measurement data. Most of these data are now being made available to the research community. (ACCOR, *** add reference ***)

A relatively simple way to build a speech synthesiser is through concatenation of stored human speech components. In order to achieve natural coarticulation in the synthesised speech, it is necessary to include transition regions in the building blocks. Often-used transition units are *diphones*, which represent the transition from one phone to another. Since diphone inventories are derived directly from human utterances, diphone synthesis might be expected to be inherently natural sounding. However, this is not completely true, because the diphones have to be concatenated and in practice there will be many diphone junctions that do not fit properly together. In order to be able to smoothe these discontinuities the waveform segments have to be converted to a convenient form, such as some form of LPC parameters, often with some inherent loss of auditory quality. Until recently it was believed that a parametric representation was mandatory to be able to change the pitch and timing of utterances without disturbing the spectral envelope pattern. Since the invention of PSOLA-like techniques high quality pitch and time changes can be effected in the time domain.

Another important means to generate computerised speech is through *synthesis by rule*. The usual approach is to input a string of allophones to some form of formant synthesiser. Target formant values for each allophone are derived from human utterances and these values are stored in large tables. With an additional set of rules these target values can be adapted to account for all kinds of phonological and phonetic phenomena and to generate a proper prosody.
More detailed accounts of speech synthesis systems can be found in, for instance, Klatt (1987) and Holmes (1988).

For all three types of speech synthesis systems corpora are needed to determine the model parameters. If the user wants many different types of voice, the speech corpus should contain various speakers for the extraction of speaker specific model parameters. In particular, the user might want to be able to generate both male and female speech. Transformations to convert rule systems between male and female speech have had only limited success, so it seems more convenient to include both sexes in the speech corpus. Application specific corpora are needed to investigate issues related to prosody.

3.3.2.2 Speech recognition

There are several types of speech recognition systems, which can differ in three important ways:

- The recognition strategies they use.
- The speakers they have to recognise.
- The speech they have to recognise.

These different aspects will be discussed below.

3.3.2.2.1 Knowledge-based *vs.* stochastic systems

With respect to the strategies they use, speech recognition systems can be roughly divided in two classes: *knowledge-based* systems and *stochastic* systems. All state-of-the-art systems belong to the second category. In the — now essentially abandoned — knowledge-based approach an attempt was made to specify explicit acoustic-phonetic rules that are robust enough to allow recognition of linguistically meaningful units and that ignore irrelevant variation in these units. Stochastic systems, such as Hidden Markov Models (HMMs) or neural networks, do not use explicit rules for speech recognition. On the contrary, they rely on stochastic models which are estimated or *trained* with (very) large amounts of speech, using some statistical optimalisation procedure (e.g. the Estimate-Maximise or the Baum-Welch algorithm).
Higher level linguistic knowledge can be used to constrain the recognition hypotheses generated at the acoustic-phonetic level. Higher level knowledge can be represented by knowledge-based explicit rules, for example syntactic constraints on word order. More often it is represented by stochastic language models, for example *bigrams* or *trigrams* that reflect the likelihood of a sequence of two or three words, respectively (see also Ney's chapter).

3.3.2.2.2 Speaker-independent *vs.* speaker-dependent systems

Speech recognition systems can be either *speaker-dependent* or *speaker-independent*. In the former case the recognition system is designed to recognise the speech of just a single person, and in the latter case the recognition system should be able to recognise the speech of a variety of speakers. All other things being equal the performance of speaker-independent systems is likely to be worse than in speaker-dependent systems, because speaker-independent systems have to deal with a considerable amount of inter-speaker variability. It is often sensible to train separate recognition models for specific subgroups of speakers, such as men and women, or speakers with different dialects (Van Compernolle et al. (1991)).
Some systems can to some extent adapt to new speakers by adjusting the parameters of their models. This can be done in a separate training session with a set of predetermined utterances of the new speaker, or it can be done on-line as the recognition of the new speaker's utterances gradually proceeds.

Most recognition systems are very sensitive to the recording environment. In the past, speakers used to train and develop a system were often recorded under 'laboratory' conditions, for instance in an anechoic room. It appears that the performance of speech recognisers, which are trained with such high quality recordings, severely degrades if they are tested with some form of 'noisy' speech (Gong (1995)). Also the use of different microphones during training sessions and test sessions has a considerable impact on the recognition performance.

3.3.2.2.3 Isolated words *vs.* continuous speech

The third possible distinction between speech recognition systems is based on the type of speech they have to recognise. The system can be designed for *isolated word recognition* or for *continuous speech recognition*. In the latter case word boundaries have to be established, which can be extremely difficult. Nevertheless, continuous speech recognition systems are nowadays reasonably successful, although their performance of course strongly depends on the size of their vocabulary.
Word Spotting can be regarded as a special form of isolated word recognition: the recogniser is 'listening' for a limited number of words, these words may come embedded in background noise, possibly consisting of speech (of competing speakers, but also of the target speaker who is producing the word embedded in extraneous speech).

3.3.2.2.4 Corpora for speech recognition research

In general, two speech corpora are needed for the development of speech recognition systems: one for the training phase and one for the testing phase. The training material is used to determine the model parameters of the recognition system. The testing material is used to determine the performance of the trained system. It is necessary to use different speech data for training and testing in order to get a fair evaluation of the system performance.
For speaker-dependent systems, obviously the same speaker is used for the training and testing phase. For speaker-independent systems, the corpora for training and testing could contain the same speakers (but different speech data), or they could contain different speakers to determine the system's robustness for new speakers.
When a system is designed for isolated word recognition, it should be trained and tested with isolated words. And similarly, when a system is designed for telephone speech, it should be trained and tested with telephone speech. The design of corpora for speech recognition research thus strongly depends on the type of recognition system that one wants to develop. Several large corpora for isolated word (e.g. TIDIGITS) and continuous speech recognition (e.g. Resource Management ATIS, and Wall Street Journal) have been collected and made available, especially in the American (D)ARPA programmes.

3.3.2.3 Spoken language systems

Speech synthesis and speech recognition systems can be combined with Natural Language Processing and Dialogue Management systems to form a Spoken Language System (SLS) that allows an interactive communication between man and machine. A spoken language system should be able to recognise a person's speech, interpret the sequence of words to obtain a meaning in terms of the application, and provide an appropriate response to the user.
Apart from speech corpora needed to design the speech synthesis and the speech recognition part of the spoken language system, speech corpora are also needed to model relevant features of spontaneous speech (pauses, hesitations, turn-taking behaviour, etc.) and to model dialogue structures for a proper man-machine interaction.
An excellent overview of spoken language systems and their problems is given in Cole (1995). The ATIS corpora mentioned above exemplify the type of corpus needed for the development of SLS.

3.3.2.4 Speaker recognition/verification

The task of automatic speaker recognition is to determine the identity of a speaker by machine. Speaker recognition (usually called speaker *identification* can be divided into two categories: *closed-set* and *open-set* problems. The closed-set problem is to identify a speaker from a group of known speakers, whereas the open-set problem is to decide whether a speaker belongs to a group of known speakers. Speaker *verification* is a special case of the open-set problem and refers to the task of deciding whether a speaker is who he claims to be.
Speaker recognition can be *text-dependent* or it can be *text-independent*. In the former case the text in both the training phase and the testing phase is known, i.e., the system employs a sort of password procedure. Knowledge of the text enables the use of systems which combine speech and speaker recognition. However, password systems are volatile to fraud using recordings of the passwords spoken by a customer. One way of making fraud with recordings much more difficult is by the use of *text prompted* techniques, whereby the customer is asked to repeat one or more sentences randomly drawn from a very large set of possible sentences. In the case of text-independent speaker verification the acceptation procedure should work for any text in both the training or the testing phase.
There are various possible applications of speaker recognition, for instance helping to identify suspects in forensic cases, or controlling access to buildings or bank accounts. As with speech

recognition, the corpora needed for speaker recognition or speaker verification are dependent on the specific application. In any case a speech corpus for the training and testing of speaker recognition/verification systems has to contain some stretch of speech from a number of speakers. For a text-dependent system a fixed text is read out by the speakers; for a text-independent system the speech corpus can contain any kind of speech, including samples of spontaneous conversational speech. Furthermore, it is very important that speech is recorded in different types of environments (quiet ones and noisy ones) and that different types of microphones are used. Especially in the case of speaker verification one has to be sure that the speaker is recognised and not the environment. More detailed accounts of speaker recognition can be found in O'Shaughnessy (1986) and Gish & Schmidt (1994). LDC offers a number of corpora designed with speaker recognition research in mind, e.g. the very large *Switchboard* corpus.

3.4 Specification of the linguistic content

3.4.1 Different types of speech data

The speech material in a corpus can vary from isolated sounds to complete conversations. In general, the extent to which the experimenter has control over the speech material decreases as it becomes more and more spontaneous and natural. The term *natural* refers to a rather intuitive concept that can be interpreted in different ways. We regard speech to be maximally natural when two or more speakers have a conversation in a familiar environment about a subject they themselves choose to talk about, since this is the situation for which speech was 'invented'. Although read aloud speech is a commonly used speaking style (and may be regarded as a natural speaking style from a sociolinguistic point of view), we regard this style as derived from the most natural style mentioned above. When reading a text, people have the tendency to speak more formally and to articulate more carefully than when they are involved in free conversation. Thus, in our opinion the naturalness of speech should be judged on a gradual scale. It should be noted that *control* over the speech material is not always necessary and may even be counterproductive, especially when one wants to study the variation of speech as a function of communicative context. However, strict control over the speech material is required for some applications, such as the development of speech synthesis systems.
In the following paragraphs eight types of speech data will be distinguished.

3.4.1.1 Read aloud isolated phonemes

Vowels pronounced in isolation (or in a 'neutral' context, such as /ht/) are often used as frame of reference for experiments in which vowels from connected speech are investigated. Continuant consonants, e.g. /l, r, w, j, n, m, s, f/, can also be pronounced in isolation. Non-continuants, e.g. /p, t, k, b, d, g/, must be followed or preceded by a vowel, e.g. the 'neutral' schwa.

3.4.1.2 Read aloud isolated words

Isolated words can be either 'nonsense' words or existing words. In the case of nonsense words the experimenter can create all possible kinds of phonotactically correct sound sequences. This gives the opportunity to study coarticulation in a systematic way. Nonsense words are also used to extract models for a dictionary of phonetic elements when a synthesis system is developed. When existing words are used, the number of possible sound sequences is restricted to what is phonotactically appropriate in the lexicon of a given language. It must be realised that control over the sounds produced by the speakers may not be perfect, because the pronunciation of polysyllabic words can be influenced by the stress pattern, which may be ambiguous (cf. words like *record* in English).
When speakers have to read aloud a list of isolated words, their pronunciation may be influenced

by the orthographic representation of the words, a phenomenon known as *spelling pronunciation*. Spelling pronunciation is especially apparent in languages which form nominal compounds; if sound sequences occur across the morpheme boundaries that would lead to assimilation and degemination in connected speech, one should still anticipate that in reading aloud all sounds are realised. This phenomenon can be circumvented by having the speakers name the words through the presentation of pictures, but this method can only be applied to a very limited number of words. It is, for instance, not suitable for abstract concepts.

3.4.1.3 Read aloud isolated sentences

The carrier sentence is one type of an isolated sentence. Carrier sentences are often used when one wants to get a somewhat more natural pronunciation of (nonsense) words in comparison with words spoken in isolation, especially with respect to speech rate. The test words are embedded in the carrier sentence, as illustrated by the example "I will say – *a test word* – again". The same carrier sentence is used repeatedly for all occurring test words, so that the influence of the acoustic and linguistic context on the test words is controlled.

More natural speech material can be obtained when 'normal' (linguistically meaningful) sentences are constructed by the experimenter. Such sentences can be used to train phoneme based recognisers and to study, for instance, word stress or coarticulation in a relatively natural linguistic context. One should note that an isolated sentence may be interpreted by a speaker in a wider semantic context, which can influence the pronunciation of the sentence, especially with respect to the position of sentence accent(s). Sometimes a semantic relation between subsequent 'isolated' sentences may arise as a result of the specific ordering of the speech material. Since individual speakers may imagine a different semantic context for a specific sentence, variability in the suprasegmental features of the test sentences can occur. If desired, this variability can be reduced by using punctuation and other typographical means (for instance, capitals or boldface characters) to indicate words that should have a sentence accent. A more natural way of doing this, is to let each sentence be preceded by a question that evokes sentence accents at the desired positions. It should be clear, however, that neither practice can be recommended in the collection of large corpora of telephone speech.

For many purposes, such as the development of a phoneme-based recogniser, it is crucial that all phonemes are represented in the speech corpus in sufficiently high numbers. Due to the large differences in frequency of occurrence of the phonemes in the language in general, uniform phoneme frequencies will not obtain in randomly chosen sentence material: such material will, instead, reflect the differences in phoneme frequencies. It is proposed to reserve the term *phonetically balanced* for speech material containing phonemes according to their frequency of occurrence in the general language. *Phonetically balanced* sentences may be used for speech audiometry and for testing the transmission characteristics of communication channels or public address systems.

Approximately uniform phoneme frequency distributions can be achieved by using *phonetically rich* sentences. For that purpose *greedy algorithms* (Van Santen (1992)) can be used. Suppose you want to have a set of sentences in which each phoneme of the language of interest occurs at least once. Of course, you could try to create this set of sentences yourself, but this would be difficult and time-consuming. Furthermore, you might end up with sentences that look rather 'constructed'. An alternative would be to search for an appropriate set of sentences in a sufficient large text corpus, for instance, a large amount of newspaper data on CD-ROM. A greedy algorithm can be used to obtain the minimum number of sentences containing all phonemes. The following steps have to be taken to get the desired test set:

1. Use a grapheme-to-phoneme converter in order to be able to search for phonemes instead of characters.

2. Select the sentence in the corpus with the largest number of phonemes, not counting phonemes that are repeated within the same sentence.

3. Select each next sentence as the one with the largest number of phonemes that have not yet been covered. Stop this procedure when the entire set of phonemes has been covered.

To obtain more occurrences of each phoneme, the procedure described above can be repeated any number of times with the remaining sentences in the text corpus. Of course, the greedy algorithm can also be used for other basic units than phonemes, for instance, characters, diphones, vowels in specific consonantal contexts, subsets of words, or specific discourse units. The greedy algorithm can be amended in various ways. For example, one can maximise coverage of high frequency units by using the sum of the frequencies of the units in a sentence as criterion. This may be important when complete coverage of all units is impossible, in which case one likes to cover the most frequent units first. Furthermore, all kinds of constraints can be imposed on the sentences that are selected, for instance with respect to their length, word material, syntactic structure, etc. Note that you can also choose other contexts for the basic units than sentences, such as clauses, or words. For example, you might want to search the text corpus for the minimal set of words in which each phoneme occurs at least once.
It should be clear that very large text corpora may be needed to obtain a sufficient number of phonetically rich sentences. In some corpora *phonetically rich* does not only mean that an attempt has been made to obtain uniform phoneme frequencies, but also uniform diphone or triphone frequencies. When designing phonetically rich sentences intended to be read by members of the general public (e.g. in the POLYPHONE corpora) care must be taken to avoid very long sentences, because these are extremely difficult to read aloud. Moreover, all sentences must be checked for very rare words (which are likely to cause reading problems) and for contents which are potentially insulting. In POLYPHONE, candidate sentences had to contain at least four words and a maximum of 80 characters, including spaces and punctuation marks. The latter condition guarantees that the prompting text will not span more than two lines on the prompting sheet. Selection of sentences on the basis of length and phonemic contents can be done automatically. However, checking for insulting contents must be done manually.

3.4.1.4 Read aloud text fragments

The naturalness of the produced speech may increase even more when speakers read aloud a series of sentences that are semantically related, *provided that the subject is able and used to read aloud paragraph length material.* The prompting material can consist of a text fragment taken from, for instance, a newspaper or a book. But the text fragment can also be created by the experimenter, when it is necessary to impose some specific restrictions on the speech material, for instance with respect to phonemic structure, word structure, or syntactic structure. Reading aloud a text fragment is more difficult than reading aloud a list of isolated sentences. It is very likely that the speech produced by different speakers, who are asked to read a text fragment, will vary considerably, especially with regard to aspects like vividness, speech rate, omitted speech segments, prosody, etc. The preferred position of sentence accents in a text fragment can be indicated with capitals or boldface characters. This is not recommended if one is interested in more natural speech.

3.4.1.5 Semi-spontaneous speech

When speech corpora are gathered for commercial applications, a common task of speakers is to read numbers or alpha-numerical expressions, such as ZIP-codes. Speakers have to some extent the freedom to pronounce these numbers or alpha-numerical expressions as they like. For example, there appear to be substantial differences between the ways in which subjects express telephone numbers. Some may read the telephone number as a string of digits, whereas others may read it as a string of numbers containing two or more digits. In addition, it may make a difference whether the telephone number is familiar (for instance, a friend's number), or unfamiliar. The POLYPHONE corpora are good examples of corpora that contain such semi-spontaneous speech.

3.4.1.6 Spontaneous speech about a predetermined subject

The previous types of speech material were all concerned with the *reading aloud* of some piece of text by *one* speaker at the time (disregarding the naming of words through the presentation of pictures). In the present section we will discuss *spontaneous* speech from *one or more* speakers. The major difference between read speech and spontaneous speech is that the former fixes vocabulary and syntax, whereas the latter leaves speakers free to choose their own vocabulary and syntax. The naturalness of the produced speech increases when speakers are allowed to choose their own words. In order to keep some control over the speech material, the experimenter can determine the subject the speaker has to talk about.

The subject of conversation is relatively fixed when speakers are asked to retell a story that they heard or read shortly before. Since it is likely that speakers will use at least some of the words that occurred in the story, this method allows the experimenter to gather 'spontaneously' spoken versions of specific words of interest. In a variant of this method, speakers can be asked to invent a story based on a cartoon (without text balloons), or on some complex picture that is bound to evoke the words of interest. In all these designs, monologues are involved, although a session manager may try to guide the discourse in the desired direction. However, one should be aware that many naive subjects do not feel at ease in a situation in which they must maintain a monologue for an extended period of time. Most people feel much more comfortable in a dialogue situation. Moreover, interview situations provide some additional control over subjects' speech, because the interviewer determines the subject of conversation, and subsequently guides the conversation in the desired direction.

Another kind of guided spontaneous speech is an information dialogue: people who attempt to obtain information about, for instance, train or plane schedules. Speakers ask information from an information agent or a computer system about time and place of departure, destination, etc. In this way spontaneous speech can be obtained, even if it concerns a very restricted subject. This paradigm is used in the (D)ARPA Air Line Travel Information System (ATIS) task. Train time table information dialogues are now being recorded in several languages, e.g. German, Dutch, Italian, etc. in the MLAP projects MAIS and RAILTEL.

Although a speech situation with two or more people is more natural than a monologue, overlapping acoustic material may result from several people speaking simultaneously. For some applications, such as research on basic speech processes, overlapping acoustic material is difficult or impossible to use. Of course, when dialogues were recorded one could try to extract speech fragments in which only a single speaker is talking. The study of simultaneous speech from two or more speakers is important for research on dialogue/discourse analysis, intention analysis, and spoken language understanding. The gathering of multiple simultaneous speaker corpora is still in its infancy. Such corpora seem indispensable to study speech in all its relevant aspects. In addition, speech recognisers, which are up to now only able to deal with one speaker at a time, would eventually also have to be able to deal with different speakers talking simultaneously. Speech corpora containing dialogues could supply the training and testing data for such advanced recognisers. To make such corpora useful for research and development purposes each individual speaker should be recorded on a separate track, using a microphone array with very high directional sensitivity. Additional tracks could then be synthesised, simulating less perfect directional sensitivity. Alternatively, subjects could be recorded in a teleconference, although such distributed recordings would require extensive precautions to allow one to synchronise the tracks originating from completely independent recorders.

A special type of information seeking dialogue, which is becoming increasingly important, is the one between a human and a computer. In order to gain a clear insight into the way people behave when they have to interact with computers, in the absence of computers that can entertain such a conversation, the *Wizard of Oz technique* was invented. This technique will be briefly described in the next paragraph.

3.4.1.7 The Wizard of Oz technique

In the children's novel The Wizard of Oz (Baum (1900)) a young boy is bullied by an oracle called the Wizard of Oz. The crux of the story is that the Wizard of Oz turns out to be nothing more than a device operated by a man. In the Wizard of Oz technique a human plays the role of the computer in a simulated human-computer interaction. Of course, the easiest way to learn about the way humans behave when they have to interact with computers would be to actually have them interact with a computer. However, in order to be able to build a computer system that can participate in a dialogue with a human, one has to know how a human-computer interaction is likely to proceed. The Wizard of Oz technique can be seen as an intermediate step in the design of such a computer system. Because the subjects who participate in a Wizard of Oz experiment have to be convinced that they are actually talking to a computer, some precautions must be taken. For example, the wizard simulating the computer should be talking with a 'computer voice' (in the case that spoken output is required), and the wizard should also make deliberate errors similar to the ones that a computer could be expected to make in the application of interest.
As Spoken Language Systems are rapidly approaching a performance level that is acceptable for an increasing range of applications, it seems likely that man-machine dialogue systems will be used more and more in the near future. For the development of such systems speech data gathered in Wizard of Oz experiments will be indispensable, as long as at least one part of the system is not yet good enough for experiments with large groups of users. A more comprehensive discussion of the Wizard of Oz technique is given in chapter 14.
As the performance of SLSs improves, the development of new applications will be increasingly based on pilot experiments with *a system in the loop*, i.e., with test versions of the application in which the wizard is replaced by a computer system which has enough functionality to support the human-machine interaction.

3.4.1.8 Spontaneous speech

This type of speech material, in which speakers are allowed to freely choose their own words and their own subject of conversation is most natural, especially in a dialogue situation. Most remarks that were made in the previous paragraph, also apply to the present one. As with all natural processes, the observer's paradox can play a role in the recording of spontaneous speech. That is, in order to obtain speech that is as natural as possible, the researcher has to observe how people speak when they are not being observed (Labov (1972)). To overcome this methodological paradox, several techniques have been proposed throughout the development of sociolinguistic research (Argente (1991)):

- *Trying to give subjects the impression that observation is absent*
  Subjects are usually unaware of being observed when they are (surreptitiously) recorded before or after the actual experimental procedure (reading sessions, interviews, or whatever). Also (planned) interrupting phone calls, or interruption by a third person during the recording session can give a subject the impression that observation has temporarily stopped.

- *Asking subjects about emotional events in their life*
  The investigator can, for instance, ask subjects to remember occasions when they were in physical danger. Speech data available from radio or television interviews with witnesses of an actual disaster (a plane crash, an earthquake, etc.) could contain realistic samples of emotional speech. This technique is based on the belief that emotional involvement triggers unattended forms of speech.

- *Using a member of the subject's own community for interviews or conversations*
  This technique is based on the belief that people tend to talk in a more casual way to people with a similar social and/or geographical background. Especially recordings of people who are friends in everyday life, are likely to contain samples of the vernacular.

- *Focussing research on adolescents*
  This strategy is based on the belief that adolescents have not yet developed the full-fledged range of speech behaviour and communicative competence characteristic of sophisticated adults and that adolescents therefore may be inclined to use a more casual speech style even in the presence of a recording device.

3.4.2 Factorial experiments and corpus studies

Experimental speech research has traditionally been focussed on factorial experiments, that is, experiments in which a number of factors are defined that are hypothesised to influence some aspects of the speech behaviour, in production or in perception. The amount of speech in these experiments has typically been small, if only because it was practically impossible to record large amounts of speech in production experiments or to generate large amounts for perception experiments. The major causes of the limitations were in the tight control of the speech needed for well designed factorial experiments and in the time required from the subjects. Tight control is necessary to prevent the outcome of factorial experiments of being meaningless: this type of experiment requires that all conceivable factors different from the small number under study be kept constant, whereas the experimental factors are being varied over a limited range. It is not our intention to criticise factorial experiments, if only because they have contributed to virtually all the knowledge we have about speech and because until recently there was hardly an alternative. Yet, it must be acknowledged that, exactly due to the tight control, the speech used in the older experiments may not have been exactly 'communicative'. In the majority of the cases the subjects performed in situations which are quite remote from normal communicative behaviour; therefore, some caution should be exercised in generalising the results of controlled experiments to 'normal communicative' speech.

Another reason to be careful in interpreting results of factorial experiments is the possibility that the experimenter did not completely succeed in keeping all non-experimental factors constant: it may be the case that nonexperimental factors did co-vary with experimental ones, thereby being responsible for at least part of the effects attributed to the experimental factor(s). One case in point is intonation research, that has been pretty much focussed on pitch and on duration effects. There is, however, increasing evidence that other factors like spectral structure, spectral slope, spectral dynamics, etc. also play a role, and perhaps one that is quite important. In short: there is a danger that factorial experiments lead to overestimating the impact of the factors under investigation, at the cost of factors that were supposed to be constant, but that actually co-varied so as to enforce the effects of the experimental factors.

Now that very large corpora become available it is possible to set up another type of experiment, in which the behaviour of one or more specific factors is investigated in a very large, perhaps comprehensive number of different contexts. Instead of trying to neutralise the effect of concomitant factors by trying to keep them constant (which will normally mean that one of the many different levels of such factors are selected, e.g. a voiceless stop as the right neighbour of the phonemes under study, or only syllables which have a prominence lending High-Low pitch contour), one may try instead to sample many different contexts. Of course, in order to make this type of research feasible, one has to assume that subject effects can be treated in exactly the same way as context effects, because it will still be extremely difficult to have subjects perform for very long periods of time. In designing corpus based experiments one must be aware of the extreme skewing of many frequency distributions observed in spoken language. For instance, in all languages for which data on phoneme frequencies are available it has appeared that within a system some phonemes occur much more often than other phonemes. Random sampling would leave one with a very high likelihood of missing infrequent phonemes and of missing possible contexts, unless the total corpus is made excessively large. *Greedy algorithms* (cf. Van Santen (1992)) can be used to find the minimum amount of linguistic material that covers a maximum number of phenomena, but even with the use of greedy algorithms it cannot be guaranteed that all possibly relevant conditions are indeed covered: conditions which are not formulated as targets for the search will only be present by chance. Since complete coverage is not practically attainable, corpus research must deal with missing data in one way or another. Attempts have been made to handle missing data by means of knowledge-based arithmetic models, including all relevant parameters; alternatively, 'blind' statistical modelling techniques like CART can be used. There seems to be some preference for arithmetic models, unless one can guarantee that the missing data are not concentrated in a few subspaces (cf. Van Santen (1994)).

3.5 Specification of number and type of speakers

In addition to specifying the linguistic content of a corpus, the number and type of speakers in a corpus is the second major factor in specifying and classifying corpora. Due to their idiosyncratic characteristics speakers add substantially to the amount of variation present in a corpus.

3.5.1 Corpus size in terms of speakers

The number of speakers who are represented is one of the most important characteristics of a spoken language corpus. Based on the *number* of speakers in the corpus, speech corpora can be roughly divided in the following three classes:

3.5.1.1 Speech corpora with few speakers

Such corpora are often used in the development of speech synthesis systems. In most cases the speech of one or two persons (typically one man and one woman) is recorded. The corpus is used to prepare dictionaries of phonetic elements (allophones, diphones, etc.), and to design prosodic models. The speech material may consist of nonsense words in which sequences of phonetic elements are systematically varied, and a series of sentences to extract prosodic rules. For developing synthesis systems it is recommended to use experienced speakers. Especially when recording the material that serves for building the segment inventory it is extremely important that the speakers can keep pitch, loudness, voice quality and tempo constant.
Corpora comprising very few speakers are also common in basic speech research, especially where invasive measurements must be made. Corpora in this domain typically contain several additional signals recorded simultaneously with the acoustic speech signal, see e.g. the ESPRIT Basic Research Project on Articulatory Phonetics. The additional signals can range from the Electroglottogram (which was also recorded in part of the EUROM.1 corpus and the Transnational English Corpus) to subglottal pressure recorded via tracheal puncture and EMG activity of intrinsic laryngeal muscles. It should be emphasised that *very few speakers* does not necessarily imply a *small corpus*. For instance, when one needs to record one speaker producing all three-consonant clusters in languages like Dutch, English or German, in all possible phonetic contexts, within syllables, across syllable boundaries, across word boundaries, in stressed and unstressed syllables, at several positions in a sentence or in a prosodic contour, the amount of speech required is formidable, even when *greedy search* algorithms (cf. Van Santen (1992)) are used to find the smallest possible number of sentences which comprise all contexts.

Similar remarks apply to intonation and prosody in general. If a text-to-speech system is developed that must be employed in many different applications (reading factual information in e.g. a train time table information system, or reading popular daily newspapers to blind subscribers), enormous amounts of speech are needed to capture all relevant prosodic phenomena.

3.5.1.2 Speech corpora with about 5 to 50 speakers

Speech corpora of this size are often used in experimental factorial research. The speech material can range from isolated nonsense words to a complete discourse, dependent on the specific application. Studies on prosody, for instance, would require linguistic units that exceed the word level. Speakers can be men, women, or children. The speech can be recorded under high quality laboratory conditions, but also 'in the field'. In general, the number of speakers and the number of repetitions of the speech phenomena that are investigated should be large enough for a meaningful statistical processing if factorial experimental designs are planned. The power of a statistical test depends on the number of independent observations. If a corpus is developed for a factorial experiment, standard procedures are available and should be adhered to for determining the minimum number of speakers and/or the minimum number of utterances per speaker to allow planned statistical tests to reach a pre-specified power. These standard procedures can be found in most textbooks on statistics, such as Hayes (1963), Ferguson (1976), or Marascuilo & Serlin (1988)). In designing very large vocabulary speech recognition systems, on the other hand, one will strive for a maximally broad coverage of relevant phenomena, probably at the expense of high numbers of exact replications of specific (relatively rare) phenomena.

3.5.1.3 Speech corpora with more than 50 speakers

Speech corpora of this composition are necessary to adequately train and test speaker-independent recognition systems. Speakers can be men, women, or children, dependent on the application. The speech material can be limited to a list of isolated words or numbers, but it can also contain read aloud sentences and paragraphs or even spontaneous speech in the case of interactive dialogues. Speech may be recorded under laboratory conditions or in (quiet) offices, but if a telephone recognition system is involved, the speech corpus should, of course, consist of telephone speech both for the training phase and the testing phase.

3.5.1.4 General remarks

Of course, possible applications of corpora may be quite different from the typical ones listed above; some fundamental research may, for instance, require a very large speech corpus, whereas a simple recognition system may be developed with a rather small speech corpus. Furthermore, the list of applications of speech corpora given above is not meant to be exhaustive, but it should help to illustrate the large differences between speech corpora, depending on their research goal. *Speaker Recognition* is a branch of speech (technology) research which has received little attention in the past decade. This is reflected in the lack of publicly available corpora to support speaker recognition research. However, it is not completely true to say that there are NO corpora which are suitable for speaker recognition research; notable exceptions are the King corpus and the Switchboard corpus, both available though the LDC (cf. also the Proceedings of the ESCA workshop on Speaker Recognition in Martigny, April 1994). For a corpus to be suitable for speaker recognition research it is essential that speakers are recorded more than once, and that recordings are made at different days, in different realistic acoustic environments and with different microphones.

3.5.2 Speaker characteristics

How are the speakers for a speech corpus selected? Again, this strongly depends on the application one has in mind. For the development of a speech synthesis system experienced speakers, such as news readers or actors, are most appropriate. For the training and testing of recognition systems, on the other hand, the population of interest must be suitably sampled. There is no general agreement on the exact meaning of 'suitable' in this context. One definition would amount to random sampling of the population of interest. This operationalisation usually results in different numbers of samples from subpopulations in the population of interest. For example, when the total population of army personnel is sampled, the subpopulation of women is likely to be poorly represented. In the case of the training and testing of a recognition system for the army, this female under-representation might seem to be acceptable, because the recogniser would have to deal mainly with male speakers. However, it may appear that some of the influential heavy duty users are women and then the recogniser should better be designed to handle the few but important women with the same performance as for men. In general, random sampling has the potential drawback that extremely large numbers of samples are needed to ensure that rare, but nevertheless important phenomena are included. When, where, and why rare phenomena may still be important depends on the application for which the corpus is collected. In the case of fundamental research, on the other hand, the aim is often to compare subpopulations in some respect, and then it would be more appropriate to draw an equal number of samples from all subpopulations of interest. Uniform sampling of all subpopulations of interest guarantees that all relevant variation is included in the corpus with the smallest possible number of speakers. The application for which the speech corpus is collected not only determines the best sampling strategy, but it also influences the choice of speakers. For example, speech processing often involves spectral analysis of the recorded speech. Several analysis techniques, such as pitch extraction or formant extraction, are less accurate for high-pitched voices (women and children) than for low-pitched voices (men). If such analysis techniques are used and the sex of the speakers is of no concern for the research goal, it would thus be sensible to select only men for the speech corpus. In general, however, it is recommended to include all possible types of speakers in a speech corpus, unless there are imperative arguments to exclude specific speaker groups. Specifically, it is strongly recommended to include equal numbers of females and males in each corpus. Speaker characteristics, which are potentially important and should therefore be considered when selecting the speaker population are described and discussed below.

3.5.2.1 Stable $\leftrightarrow$ Transient speaker characteristics

The many speaker characteristics that may influence the speech signal can be divided in two main classes: relatively stable characteristics, and transient (temporary) characteristics. Stable speaker characteristics comprise on the one hand physiological and anatomical factors such as sex, age, weight, height, smoking/drinking habits, and possible pathologies, and on the other hand geographical and sociolinguistic factors. Transient (temporary) speaker characteristics cover factors such as a cold, or other mild afflictions of the speech organs, general physical condition (dependent on, for instance, the number of hours of sleep during the previous night), stress, and emotional state. Whereas transient speaker characteristics are very difficult to control, stable speaker characteristics are easier to take into account in the design of the speech corpus. For an overview of several important stable speaker characteristics, we refer to Scherer & Giles (1979). The most important stable speaker characteristics will be mentioned below.

3.5.2.2 Demographic coverage

Demographic factors form a very important set of relatively stable speaker characteristics which must be considered when designing sampling procedures for a corpus collection project. Each corpus should have sufficient demographic coverage. However, it is not always possible to determine all potentially relevant demographic factors *a priori*. Nor is the distribution of all factors in the total population always known. It is likely that the availability of detailed and reliable demographic data differs between the European countries. The availability of the data in lesser developed countries is even more questionable. In selecting speakers for inclusion in a corpus the possibility to assess certain characteristics is dependent on the recording protocol. If randomly selected speakers are recorded over the telephone, many personal characteristics cannot reliably be collected: self-report from the speaker is the only means of gathering the data.

3.5.2.3 Male/Female

Sex is known to have an enormous impact on speech quality. It is not well known at what age sex-related speech characteristics become prevalent. There is some evidence that sex-related speech characteristics are only partly due to physiological and anatomical differences between the sexes; cultural factors and sex role stereotypes also play an important role. Therefore, it is possible that the age at which sex-related differences become apparent differs between cultures and therefore between languages. See for general information on sex-related speech characteristics Smith (1979), Coates (1986), Philips, Stelle & Tanz (1987), and Brouwer & Haan (1987)). For the time being, no definitive recommendations can be given with respect to the age above which sexes should be distinguished and sampled individually. Unless the contrary can be motivated from the specific application the corpus is collected for, each corpus should comprise approximately equal numbers of speakers of both sexes. For some applications, recordings of young children may also be required. Children should be considered as a 'third sex', independent from adolescent or adult females and males. Speaker sex is known or suspected to affect at least four aspects of speech behaviour.

3.5.2.3.1 Pitch and intensity

Women are known to have higher average pitch than men. There are also indications that average intensity in female voices is somewhat lower than in male voices. Especially the higher pitch may affect spectral analysis techniques: pitch and formant extraction may be less accurate for high-pitched female voices than for low-pitched male voices. When a corpus is recorded to develop and test parameter extraction techniques, a realistic proportion of high-pitched female voices should be present. It should be realised that there is an interaction between sampling rate and the accuracy with which pitch frequency can be determined. In female and child speech even 20 kHz sampling frequency may not be high enough to obtain sufficient accuracy, as pitch frequencies may be as high as 500 to 750 Hz. Fortunately, sampling frequency can be increased using straightforward signal processing procedures whenever the need arises.

3.5.2.3.2 Overall spectral slope

Women are reported to tend more towards breathy a voice quality than males. It is not known whether this tendency is related to physiological and anatomical factors or that it is mainly due to culturally determined sex role stereotypes. Overall steeper spectral slope causes problems for some parametric signal processing techniques (e.g. formant extraction).

3.5.2.3.3 Accuracy of pronunciation

Women are reported to adhere more to standard pronunciation than men. It is not known whether this finding, made for (American) English and Dutch, generalises to other languages. It remains to be seen whether sexrelated pronunciation variation is best modeled and described on the level of phonemic representations of words or on the level of the phonetic implementation of what is essentially the same phonemic form. Awaiting results of further research in additional languages/cultures this factor is probably not sufficiently important to attribute great importance to it. Moreover, this aspect is very difficult, if not impossible, to separate from other sex-related factors, and will therefore be duly represented as long as the sexes are adequately represented in the corpus. Variation in pronunciation accuracy may also be caused by factors related to age and social status.

3.5.2.3.4 Vocabulary and syntax

Sex-related differences in vocabulary and syntax are certainly culturally determined. Here, the factor sex interacts with factors like age and social status. Differences on the level of vocabulary and syntax are only relevant when spontaneous speech is being recorded. If all speech material exists of read utterances, vocabulary and syntax are completely determined by the prompting material.

3.5.2.4 Age

Although the impact of speaker age on speech behaviour has not received much attention in previous research, there are indications that age influences at least two aspects of speech behaviour (Helfrich (1979)).

3.5.2.4.1 Voice quality

There has been some research on the relation between age and voice quality. Most studies were concerned with the question whether speaker age can reliably be estimated from the speech signal alone. It seems that people are moderately good at guessing age from speech signal characteristics, although reported correlation coefficients may be mainly determined by the ability to discriminate between very young, very old and adult but non-senior groups. The exact signal characteristics which enable people to guess the speaker's age are not well understood; neither is it possible to estimate their impact on the performance of automatic speech and speaker recognition. Until the questions about the importance and the exact nature of the impact of age on speech signals have been answered, it is recommended that attempts be made to sample the relevant age groups. In doing so, a distinction should be made between the group under 20, the group between 20 and 60 and the group over 60. If relevant, the group under 20 should, of course be sub-divided into toddlers, children, adolescents and young adults. However, the exact ages separating these subgroups is subject of discussion. Moreover, in many respects mental and physiological maturation may be more important than calendar age.

3.5.2.4.2 Vocabulary and syntax

Here the considerations described above in the paragraph on the impact of sex on speech behaviour apply in exactly the same way. There is some literature suggesting that vocabulary and syntax of the older generation are different from the younger speakers, but apart from obvious observations that the subject spontaneously discussed by senior citizens tend to differ there is little hard data to support the claim that age is more important a factor than, for instance, social group and education level.

3.5.2.5 Weight and height

As with speaker age, most research in the past has concentrated on the question whether people can estimate speaker weight or speaker height from speech recordings alone (Van Dommelen (1993)). It appears that people are moderately successful in this task. It will be clear that weight and height of speakers are highly correlated. The exact signal characteristics that enable people to guess the speaker's weight and height are not known. In a sufficiently large sample of speakers, most weight/height groups will probably be represented.

3.5.2.6 Smoking and drinking habits

Several investigations have shown that voice quality can change under the influence of smoking or the use of alcohol (Gilbert & Weismer (1974)). One of the most common consequences of smoking and drinking is premature ageing of the mucous membrane covering the vocalis muscle, resulting in a hoarse voice quality. Excessive drinking may eventually result in brain damage, which may in turn lead to severe speech disorders. The use of drugs can have a similar effect. In those cases it would be more appropriate to speak of pathological speech.

3.5.2.7 Pathological speech

The boundary that divides pathological speech from non-pathological speech is very difficult to draw. Hoarseness due to smoking can be regarded as a very mild speech disorder, whereas more severe speech disorders include, for instance, paralysis of the vocal cords and aphasia. Speech disorders can be divided into two main classes: those where there is a clear organic (anatomical, physiological, neurological) cause, and those where there is not. The latter category is usually referred to as functional disorder. However, in many cases there is no clear cut distinction between organic and functional speech disorders; often both types are involved, or it is unclear which of the two types is involved. Speech disorders can be described at five different levels:

- *Articulation disorders*
  This involves the distortion, deletion, or substitution of sounds or sound combinations. Usually such disorders are functional, but they may also result from lesions of the lips (e.g., a cleft lip), the palate (a cleft palate), the teeth, the tongue, the jaw, or the nose. Another possible cause of articulatory disorders is dysarthria, a damage to the central or peripheral nervous system, manifested by neuromuscular disability.

- *Resonance disorders*
  This involves lesions of the oral, nasal, or laryngeal cavities. Apart from functional causes, resonance disorders can result from, for instance, surgical removal of the tonsils, a cleft palate, or nose polyps.

- *Voice disorders*
  This involves lesions of the vocal cords, referred to as dysphonia. The voice may emerge as a whisper (no vocal-cord vibration), for instance due to paralysis; or vocal-cord vibration may be present to some degree, but accompanied by excessive air flow (a 'breathy' voice); or there may be irregular and therefore aperiodic vocal fold vibration, for instance due to the growth of abnormal tissue (nodules) on the vocal folds resulting in a 'hoarse' voice quality. Dysphonia may be caused by psychological and emotional factors, such as a severe shock, or by organic factors. A serious voice disorder is cancer of the vocal cords, which may lead to the surgical removal of the larynx (laryngectomy). Although the patients can learn alternative voicing mechanisms, their speech is usually severely degraded.

- *Language disorders*
  This involves disorders that do not affect the production of the speech message, but rather its content. These disorders are usually classified under the name *aphasia*. Patients suffering from aphasia may, for instance, use a reduced and incomplete sentence structure, have difficulty in wordfinding, use an inappropriate intonation, or make erratic pauses. The cause of aphasia is brain damage due to, for instance, a stroke, thrombosis, a tumour, or excessive drinking.

- *Rhythm disorders*
  The usual terms to describe the main rhythm disorders are stuttering (or stammering) and cluttering. Stuttering is a very complex phenomenon that is characterised by, for instance, a

repetition of speech segments, abnormal prolongations of sound segments, words being unfinished, or circumlocutions to avoid types of sound that cause problems. Stuttering varies enormously from person to person and from situation to situation. It is, for instance, well known that stutterers almost never stutter when they are singing. Both organic (genetic) causes and functional (environmental) causes are assumed to underlie the stuttering phenomenon. Another major category of nonfluency is cluttering. The primary characteristic here is that the patient tries to talk too quickly, and as a result introduces distortions into his rhythm and articulation. The description and theoretical study of cluttering is less advanced than that of stuttering. In addition, there is a considerable overlap between the categories of stuttering and cluttering.

For many purposes it is most appropriate to build speech corpora with a large variety of speakers. However, the speaker variability should be kept within reasonable bounds. Severely pathological speech will, in general, deviate substantially from 'normal' speech and thus it is usually not desirable to include this type of speech in a normal speech corpus. On the other hand, speakers with mild pathological disorders, such as hoarseness, can be included in for instance speech corpora designed for recognition.
Of course, research might focus specifically on pathological speech, for instance when a recogniser is developed for use as an environmental control device for handicapped persons. In that case pathological speech should of course be amply represented in the speech corpus. Pathological speech should also be present in a corpus designed to cover as much speaker variation as possible (a kind of 'all-purpose' speech corpus). A more elaborate discussion of pathological speech can be found in Perkins (1977) and Crystal (1980)).

3.5.2.8 Professional ↔ Untrained speakers

Professional speakers should be selected when recording very large corpora with very few speakers, for instance to develop text-to-speech systems. The major reason to prefer professional speakers for this purpose is their ability to keep pitch, intensity and speech rate constant, not only during one recording session, but also over several sessions, which may have to be scheduled on different days, perhaps spread over several weeks or even months. One possibly important drawback of using professional speakers must be emphasised: more often than not, professional speakers are not really representative of the 'normal' speech behaviour in the community. If the corpus is collected for the development of a text-to-speech system this may not be a problem. However, linguistic and phonetic findings based on a corpus comprising only speech of a small number of highly trained professional speakers should not be generalised without extreme caution.

3.5.2.9 Geographical and sociolinguistic factors

It is well known that both the regional and the sociolinguistic background of speakers can have a large effect on their speech. People speak differently dependent on the specific region(s) in which they were brought up, and dependent on factors such as the linguistic background of the parents, social status, and education level. It is widely assumed that the high-school period is most decisive for the regional or dialectal colouring in one's speech. Therefore it is strongly recommended to obtain information about the high-school period when collecting data about the speaker's background.
Dialectal speech or regional/dialectal colouring of the prestige variant of a language (like Received Pronunciation (RP) in British English or *Hochdeutsch* in Germany) are known to be perhaps the most important source of speaker-related variation. Not all languages have a widely accepted and well documented pronunciation standard, like RP in English. Given the enormous amount of literature on Dialectology one would assume that the impact of dialects on standard speech is well

understood. Unfortunately, this does not appear to be the case. Linguists and dialectologists appear to disagree about the number of major dialects in a language area, and about the boundaries between the areas where a specific dialect is spoken. Moreover, the majority of the dialect studies were based on written questionnaires. Although there are large amounts of recorded dialectal speech stored in the national Dialectology institutes, these recordings do not qualify as corpora, because they exist only on analogue tapes, with little or no detailed annotation. In collecting new corpora the factor *regional/dialectal colouring* should be properly accounted for. However, since the basic data to determine number of dialects and dialect boundaries are difficult to obtain and probably not always reliable, it is recommended that *dialect* is operationalised by *geographic region*. If necessary, processing of the corpus data can yield *post hoc* data on dialect differences. However, it has appeared that *post hoc* determination of the dialect background of a speaker as part of the transliteration/transcription process poses big difficulties. There is one additional factor which complicates the procedures for sampling dialectal influence, viz. the increasing mobility of the population. It is acknowledged that the impact of mobility is different between language areas and between countries. However, in sampling for a number of large telephone speech corpora in the U.S.A. (POLYPHONE; Voice Across America) a special 'dialect' called *Army Brat* was defined, for those speakers who had lived for short periods of time in many different parts of the country. It should be noted that the factor *dialect* not only affects pronunciation. More often than not, its impact on vocabulary and perhaps also on syntax is at least as important. Of course, the impact on vocabulary etc. can only come to light in corpus collection paradigms which allow the speaker to select his own words. In corpora comprising only read speech this factor should have no effect. *Sociolects* can be regarded as dialects spoken by a particular social class. A clear distinction between different social classes exists, for instance, in India, where each member of the society belongs to a specific caste. However, in most cultures it is very difficult to distinguish between social classes. The division into three categories *lower-class, middle-class, upper-class* seems to be most widely accepted for Western cultures. Elaborate schemes have been designed to determine a person's social class using factors such as education, profession, and income. In addition to social class membership a person's sociolect is, of course, also influenced by the linguistic background of the parents and the dialect regions in which he grew up. As is the case for dialects, sociolects may influence not only pronunciation, but also syntax and vocabulary. It is recommended that sociolects should be properly accounted for when collecting new corpora. It has been found that the impact of sociolects on speech behaviour strongly interacts with speaking style. Thus, the speech of a pipe-fitter who speaks in a formal way, may resemble the speech of a salesman who speaks in a casual way (cf. (Labov (1972)). This phenomenon probably also applies to regional dialects.

There is considerable uncertainty on how to treat dialects, regiolects and sociolects in corpora collected to develop speech technology, e.g. to develop connected speech recognition systems for use in telephone information systems. There may be large differences between countries and cultures in what is most appropriate in this respect. Of course, each operational recognition system should be able to handle the range of dialectal and regiolectal influences present in the speech of upper and middle class speakers produced in somewhat formal situations. The extent to which dialectal influences occurring in less formal speech, or in formal speech of lower class speakers must also be covered will depend very much on the application for which the recogniser is being developed. Another extremely important factor is the social acceptability to use strongly dialectal speech in a given situation. Acceptance is likely to differ strongly between regions in a given country.

If telephone applications are designed in such a way that all calls originating from a specific part of the country are handled in a local centre, one may envisage recognition systems which are adapted to the local regiolect/dialect, provided that suitable training corpora can be collected. When collecting speech corpora over the telephone by soliciting input from randomly selected subjects one should specify strict guidelines for deciding whether or not a specific speaker deviates too much from the 'standard' language for him to be included in the corpus. The speech of non-native speakers can be regarded as a special 'sociolect'. Some non-native speakers may speak the standard language of the country they reside in with only a slight accent, whereas others may speak the standard language with a very marked accent and/or a poor control over vocabulary and grammar. There seems to be no reason to exclude the former group of non-native speakers from a common speech corpus, whereas the latter group of non-native speakers would preferably be excluded, unless the research is specifically aimed at non-native speech or one wants to build an 'all-purpose' speech corpus.

Chapter 4

Corpus collection

4.1 Introduction

written/spoken corpus collection dimensions

The procedures section contains recommendations for each of the data collection dimensions. These recommendations should enable anyone interested in speech recordings to establish a suitable recording environment that will deliver data in a controlled procedure and in an acceptable technical quality.

4.2 Data collection dimensions

Data collection can be represented in the following dimensions (which are characterized by their extreme positions):

Visibility : open – secret

Environment : studio – on location

Communication situation : face to face – isolated

Communication mode : dialogue – prompted

Verification : online – offline

Data : single channel – multi-channel

These dimensions are not independent. For example, a secret (or *candid microphone*) recording will not be possible in a studio environment, nor will it be possible to do electro-magnetic articulography measurements outside a specially equipped studio.

4.2.1 Visibility: Open vs. Secret

It has been a concern that people who are aware that their speech is being recorded change their speech behaviour. For that reason some scientists have advocated recording paradigms using 'candid microphones'. Another concern, which is at least as important, is that speech behaviour is strongly influenced by the social situation in which the speaker performs. Speech produced in a court room, for instance, tends to be more formal in terms of vocabulary, syntax and phonetics than speech produced in one's living room. Also, read speech, especially when the linguistic material consists of isolated words, invokes a more formal and standard pronunciation than free conversation (cf. Labov (1972)).

To limit the influence of the fact that the speaker is aware of the fact that his or her speech is being recorded one can hide the microphone and the recorder. Provided that nothing else in the environment might alert the speaker of something unusual, and provided that nothing else in the situation is unusual, 'candid microphone' speech should be maximally natural. On the other hand, many scientists have had the experience that almost all speakers tend to forget that they are being recorded fairly quickly (say within less than five minutes), if they can be engaged in a task that requires cognitive activity different from reading aloud. Therefore, with the exception of very short recording sessions, using hidden microphones might not yield a big advantage. Thus the question arises whether the possible advantage of slightly better naturalness outweighs the disadvantages, the most important of which probably is the risk of loss of recording quality because on-line monitoring of recording level is extremely difficult. Another possible drawback of using hidden microphones is that much effort may be spent in recording speakers who, after debriefing, refuse to give consent for the use of the recordings, even for purely scientific research.

Nevertheless, a substantial part of the British National Corpus has been recorded in situations where at least some persons who participated in a conversation were not aware of the fact that they were being recorded (cf. Crowdy (1993)). As far as we know, only a very small number of persons demanded that the recordings in which they participated were erased. It is not known whether the material in the BNC has been analysed to investigate possible differences in recording quality and vocabulary between secret and open recordings.

One obvious way to make sure that the speaker cannot know that (s)he is being recorded is to tape telephone speech. But here again, publishing these recordings as part of a corpus requires that the speaker must be debriefed and asked for consent to keep the recordings.

4.2.2 Environment: Studio vs. On Location

4.2.2.1 Recording in a studio

Most of the older speech corpora have been recorded in sound studios. Studio recordings have the drawback that most subjects do not feel at home in that environment, with all possible impacts on their speech behaviour. However, as long as the speech to be elicited consists of lists of words, words embedded in carrier phrases, and the like, the 'abnormality' due to the unusual reading tasks may outweigh the contribution of the unusual situation.
Studio recordings have the advantage of superior signal-to-noise ratio, thanks to the excellent acoustic environment and — at least as importantly — the possibility to closely monitor recording levels, distance of the speaker to the microphone, use of superior but volatile condenser microphones, etc.

One must be aware, however, that 'studio' is not a well defined concept. Not all rooms called studio have good acoustic properties. It is not at all unusual to find rooms which have indeed relatively low ambient sound levels, but at the cost of very long or extremely short reverberation times. If studios are used to record large corpora, room acoustics calibration data should be provided with the speech recordings. To that end the procedures proposed in the guidelines for recording the EUROM-1 corpus can be used (see Appendix C): record standard pulsatile noises, e.g. caused by exploding balloons inflated to standard dimensions. From these calibration recordings the reverberation time and the overall frequency response of the room are easy to determine. It is important that the noise source is located at approximately the same point in the room where the speaker's head is during the recordings.

Many (small) corpora designed for basic speech research have been recorded in rooms which were not designed as audio studio. This is the case with most simultaneous recordings of speech and

EMG signals, which are typically made in research labs of hospitals. More often than not these rooms have had only limited treatment to provide acceptable room acoustics.

For many applications high quality speech recordings from one speaker at the time are required. These recordings should be free of background noise, including noises made by the speakers themselves. The following guidelines apply specifically to the common situation that speakers are recorded one at the time in a sound studio:

- Many speakers are somewhat nervous when they are recorded. It is the task of the experimenter to put speakers at ease with a friendly conversation before the recording session starts, preferably in the room where the recordings take place, so that the speakers can get accustomed to their surrounding. It is good practice to start recording sessions with a number of practice items, which enable speakers to get going.
- During the recording session speakers can produce various kinds of background noise, such as moving their chair, coughing, tapping with their fingers or their feet, turning text pages, etc. This can to some extent be solved by
  - making the recording conditions as comfortable as possible,
  - explaining to speakers in advance why background noise could interfere with the speech recordings,
  - presenting the text material as clearly as possible,
  - adding dummy items at the beginning and at the end of text pages,
  - present text material on a VDU instead of on paper,
  - adding one or more repetitions of the test items.
    It is recommended that a session manager is present during recording sessions who can immediately intervene whenever speakers do not properly perform their task.
    It is also recommended to use a VDU for presenting the prompting material whenever that is possible (i.e., when recording in a studio or on the experimenter's premises). When a VDU is used, care should be taken to avoid reflections from the screen, by positioning the screen at an angle from the axis connecting speaker and microphone. It should also be checked that the monitor does not produce audible noise of its own. One should also make sure that the type size is sufficiently large, even for subjects who wear (or should wear) glasses.
  - At the end of a long recording session, speakers can become hoarse. This can be prevented by taking a sufficient number of breaks in which the speaker can drink some water, or by splitting the long session up in smaller sessions.
  - When speakers have a reading task, the pronunciation of a specific test item can be influenced by its position in the total list of test items. To compensate as much as possible for such order effects, each speaker should have a different random ordering of test items, unless there are compelling arguments not to follow this procedure.
  - Speakers are inclined to 'fade out' their speech (especially their intonation) at the end of a list of items. It is therefore recommended to add some dummy items at the end of (partial) lists of items. This concern may not apply if individual items are prompted by displaying them on a VDU.

4.2.2.2 Recording on location

Corpora recorded in the field have the advantage that the speaker is acting in an ecologically realistic environment. In most cases the price to be paid for this advantage is a substantial loss in signal-to-noise ratio, either because of high ambient noise levels, or limited possibility to monitor recording levels, distance to microphone, etc. or both. If ecological reality dictates recordings in the field one should nevertheless plan for conditions which allow an audio engineer to monitor the procedure.

Complete ecological validity may not be feasible. For instance, recording a speech corpus in a running car cannot safely be accomplished if the speaker is in the driver's seat.

Two important classes of recordings on location are recordings

- in an actual application, and
- on the telephone.

Recording speech in actual applications (which are based on speech input) is one obvious way to obtain 'realistic' speech data. At least in some countries it is not legally required to advise the user of the service that her or his speech is being recorded, as long as the recordings are only used for research purposes internal to the company which runs the service. This procedure is probably most often used in pilot versions of an application, where the number of users is limited, so that one may realistically hope to be able to process all recorded speech. Such recordings are necessary to systematically evaluate the success or failure of the speech input parts of an application by relating the speech recordings to the log of the use of the application.

Recording speech on the telephone (preferably digital, i.e. ISDN) is suitable for the gathering of limited amounts of speech material from a large number of speakers (POLYPHONE; Damhuis et al. (1994)).

A possible drawback of telephone recordings is the limited bandwidth of the speech signal, typically between 300 Hz and 3000 Hz, which may pose problems for some kinds of basic speech research. For example, the absence of low-frequency components prevents a proper pitch analysis of the recorded speech using methods which rely on the presence of the fundamental frequency (frequency-domain methods, such as the 'harmonic sieve', may yield satisfactory results). The absence of high-frequency components prevents for instance the proper spectral analysis of consonants, especially fricatives. Apart from the limited bandwidth of the speech signal, telephone channels can also give a substantial loss in signal-to-noise ratio, especially in the non-western countries where digital telephone systems are not yet commonplace. Even in modern digital telephone networks signal-to-noise ratio suffers from the limited dynamic range that can be accommodated with 8 bit A-law (or U-law) coded samples.

If these drawbacks are not important for the research goal, recording of telephone speech appears to be a simple way to collect a large amount of speech data in a very short time. For some applications, such as the training and testing of a telephone speech recogniser, a speech corpus with telephone recordings is of course indispensable. It should be emphasised that telephone speech is suitable for many linguistic research projects, including research into most aspects of dialects and regional language variants as well as all aspects related to spoken language syntax and vocabulary.

4.2.3 Communication Situation

In this section I want to compare

4.2.3.1 Face to face

and

4.2.3.2 Isolated

communication situations with each other. Is this a sensible thing to do?

Input requested here!

4.2.4 Communication Mode

In this section I want to compare

4.2.4.1 Dialogue

and

4.2.4.2 Prompted Interview

with each other. The following text is the one by Els den Os from the chapter previously known as 2.5. I am currently modifying it.

4.2.4.2.1 Text den Os

When making studio (or office) recordings, the text to be recorded can in principle be presented both in written and in auditory form. Most researchers seem to have preferred written prompts, mainly because it was thought that spoken prompts would lead the speakers to adopt the same pronunciation, loudness, speech rate, etc. as the speaker who prepared the prompting material. There is substantial literature showing that speakers adapt their speech behaviour to the person(s) they are speaking to. However, the impact of this adaptation effect is not well known and only partly understood. It must be expected that adaptation depends on a lot of factors that are difficult to control and that all have to do with the social relation between the partners in the conversation. For this reason the Edinburgh MAP corpus (H. Thompson), collected by the HCRC of Edinburgh University, included dialogues with two familiar and two unfamiliar speakers.

The simplest written prompts are printed texts containing instructions and/or stimulus material for the speakers. When the recordings are made in a studio, written prompts can also be controlled by a computer. A program furnishes all the necessary indications to the speaker, such as *Pronounce the following sentence...* on a terminal screen. The advantage of displaying prompts on a terminal screen is that the entire recording procedure is organised in a neat way. Some kind of verification procedure can be included to limit the number of errors. Timing of utterance prompting can also be conducted by the computer. This may to some extent enhance the uniformity in speech rate between different speakers. On the other hand, more natural speech may be obtained, if speakers are allowed to establish their own pace.

When recording telephone speech written prompts cannot always be used because the address of speakers (to send them a prompt sheet) is unknown.

We know of no reports showing that isolated word speech recognition systems trained with speech elicited by asking callers to repeat the prompts perform worse than devices trained with words prompted in other ways.

It has been the experience of many corpus projects that it is very difficult to design unobtrusive questions which have a very high likelihood of prompting specific answers; in a way, the two aspects are contradictory: the higher the likelihood that there is no other answer than the intended one, the more 'stupid' the question seems to be. On the other hand, asking questions — even stupid ones — comes much closer to the situation a caller will find himself in when using a Voice Response System than asking a caller to repeat words or phrases verbatim.

4.2.5 Verification

4.2.5.1 On-line verification

The following text is that of Els den Os. I am currently rewriting it, but I will need further input. **Input requested here!**

4.2.5.1.1 Text den Os

Many of the older speech corpora recorded in studio environment and consisting of read speech only have employed a recording paradigm where a trained experimenter monitored all productions for exact conformance with the prompting text. If a deviation of that text was noted, the speaker was requested to repeat the item until it was correctly produced. Within this on-line monitoring paradigm two sub-paradigms can be distinguished: one in which each error is signalled to the computer by the experimenter without interrupting the speaker (so that the erroneous items could be presented a second time after the completion of the item list); and one in which the speaker is immediately alerted that an error has occurred so that the last item can be repeated. With direct recording to disk file (especially when individual files are planned for each item) both techniques can save an enormous amount of post-processing time, compared with a protocol in which the speech is recorded on a sequential medium (irrespective of the analogue or digital recording technique).

Therefore, if on-line verification is planned, use of a computer-controlled workstation that records directly to computer files is strongly recommended. One such workstation has been built in the ESPRIT SAM Project.

On-line verification is only possible with read speech. If the text to be recorded includes contemporaneous, free speech it cannot be predicted beforehand what the speaker will say. In these cases there is no way around the time consuming *post hoc* transliteration.

4.2.5.2 Off-line (or *Post-hoc*) verification

Post hoc transliteration was employed in collecting some (very) large corpora like Voice Across America and POLYPHONE. Characteristic for the recording protocol used in these corpus collection projects was that the recording workstation performed completely automatically and unattended. It has been suggested that the read items in these corpora might have been verified on-line by using an automatic speech recogniser running in verification mode, but except perhaps for an Italian corpus consisting of isolated utterances of a small vocabulary we know of no corpus collection projects that have indeed employed automatic verification and re-prompting in a fully automatic recording workstation.

On-line verification is the only practical paradigm that guarantees that the corpus will indeed contain exactly the items and the number of repetitions planned for during corpus design. The procedure has one very important limitation, that should not be underestimated: it will yield a corpus which is (virtually) completely devoid of disfluencies, out-of-vocabulary words, coughs, sneezes, etc. Cleaned-up corpora of the type implied here have misled engineers to think that speech recognisers had reached performance levels sufficient for actual applications. What they failed to realise — due to the absence of these phenomena from the training materials — was that in real life disfluencies etc. abound, and that these phenomena may be more important in determining the real life performance of speech recognisers than recognition error rate on a clean corpus. For this reason it is strongly recommended to use post hoc transliteration whenever that is possible. In making this recommendation it is acknowledged that recording disfluencies etc. makes no sense in recording speech material for carefully designed perception experiments.

4.3 Monitoring

physical validation phonetical validation

Hans Tillmann and I will write this section.

4.4 Procedures

The following sections describe procedures which have been used in practical speech recordings.

4.4.1 Equipment

The basic equipment needed for speech recordings consists of

- microphone
- amplifier and/or processor
- recording device
- storage devices

The choice of equipment depends on the choices on the dimensions visibility, environment, communication mode, and on the data to be recorded.

The quality of the recording channel itself (microphone and recording medium) is determined by three characteristics: signal-to-noise ratio, bandwidth, and dynamic range.

For every speech recording a *log* or *journal* should be kept. It contains the essential administrative information about recording setup, personnel involved, speaker data, and recording time and date.

It is necessary to record at least the following data for a recording session:

- recording date and time
- recording engineer
- equipment used
- speakers
- recording environment

Recording time and date and the recording engineer are independent of the number of speakers or channels recorded. The environment, speakers, and equipment may differ for each channel and thus should be written down separately for each channel. A separation of recording dependent and channel dependent data is thus advisable, and this separation should be made explicit in the layout of a form (or a database structure):

rec_id	date	time	engineer	remarks
M0127D	22.01.95	17:10	CSC	VM scenario "Time Table" April-June calendar

id	channel	type	recorder	microphone	environment	speaker
342	left	audio	sony DAT	HMD416	studio	UAA
343	right	audio	sony DAT	HMD416	office	AJB
344	A	electro-palatography	c:/epg/300	palate UAA	studio	UAA
...	...	...	...	...	...	...

4.4.1.1 Microphone

Microphones can be classified by their

- directional properties,
- transducer principle, and
- position relative to the speaker

A well-documented speech corpus should contain data about the microphone, such as make and type (condenser, dynamic, etc.), position of the microphone relative to the speaker or speaker's mouth, possible calibration procedures, etc.

Unidirectional microphones are sensitive to the direction where the sound comes from. Unidirectional microphones are preferred when a single speaker is recorded in a laboratory environment.

Omnidirectional microphones are not sensitive to the direction where the sound comes from. They are suitable for recordings on location, or when the speaker is moving, e.g. working, walking, or driving a car. Omnidirectional microphones may be used to record several speakers if it is guaranteed that their turns will not overlap; otherwise multiple unidirectional microphones should be used (see ...).

Microphone arrays are a possible means of sound source location. Computer controlled microphone arrays can focus on a single sound source, therewith improving the effective signal-to-noise ratio dramatically (cf. Flanagan et al. (1991)).

The *electrical transducer principle* is a second dimension along which microphones can be distinguished. For most purposes in speech research the differences between the tranducer principles is not very important, with the exception of carbon button microphones. *Carbon button microphones* are used in older telephone handsets. They may distort the frequency response of the signal quite considerably. Moreover, their transmission properties may change significantly over time. *Electret microphones* are more stable than carbon button microphones. They can have almost flat frequency response in the telephone bandwidth, at least in principle. The actual frequency response of a microphone depends primarily on the acoustic properties of the case in which it is encapsulated. Badly designed handsets therefore can have bad frequency response characteristics, regardless of the use of electret transducers. For basic research into the characteristics of the glottal sound source the phase distortion of the microphone is as important as its amplitude response. For virtually all other research and development purposes in speech phase response is immaterial.

Finally, the *microphone position* relative to the speaker's mouth can be used to distinguish types of microphones.

Headset microphones usually are attached to headphones via an arm. The position of the microphone relative to the articulatory tract is fixed, and the speaker is free to move the head. However, the microphone has to be positioned very carefully to avoid noise through breathing, and speakers often feel uncomfortable with a headset.

Close-up microphones are attached to the speaker's clothes, usually on the chest. The microphone does not disturb the speaker and it is quite close to the articulatory tract. However, the distance of the microphone varies greatly with body movements, and new noise sources, e.g. frothing of clothes, are introduced.

Table-top microphones usually are uni-directional microphones placed approx. half a meter away from a speaker. The microphone does not disturb the speaker, and the distance of the microphone varies only very little with body movements. However, with more than one speaker in a room there is little channel separation, and new noise sources, e.g. interference from room echo, tapping on table, movements of prompt sheets, are introduced.

Room microphones are omni-directional microphones that are placed in specified positions in a room, independent of the position of a speaker. They are independent of speaker position and can be hidden completely. However, there is little (if any) channel separation, and surrounding noised interfere with the speech signal.

4.4.1.1.1 Recommendations

- If acceptable in the recording environment, and for optimal acoustical quality, use headset microphones.
- Place the microphone slightly to the left or the right of the mouth and a bit below the lower lip to avoid breathing noises. Take care that no cables touch the microphone arm, and that the speaker is comfortable with the headset.

4.4.1.2 Amplifier/Processor

The signal coming from a microphone must be amplified to be recorded. In many cases, some processing is also needed, e.g. analog to digital conversion, transformations for different encoding schemes, filtering to reduce noise, etc.

Some processing steps have to be performed only once, e.g. analog to digital conversion. Others will be performed repeatedly, e.g. the transformations for different encoding schemes.

4.4.1.2.1 Recommendations

- Define a standard setup and procedure for all the steps from recording the signal to storing it.
- Choose *de-facto* accepted standards for the storage formats, and use standardized conversion tools. For normal speech recordings the standard quantisation and sample rates should be 16 bit (linear encoding) and 16 KHz, and 16 bit 8 KHz for analog telephone speech. For ISDN telephone recordings, one should use the ISDN standard 8 bit a-law encoding at 8 KHz sample rate.
- Use the same equipment wherever possible (and appropriate).

4.4.1.3 Recording Device

Basically, there exist two types of recording devices: tape drives, and computers with hard disks. Recordings to tape are either analog (audio tapes, compact cassette, video tapes) or digital (DAT), whereas recordings to hard disk are always digital.

Ongoing development in the field of audio technology has shifted the emphasis away from analogue recording media to digital recording media. The traditional recording medium has been the reel-to-reel magnetic tape. Apart from a relatively poor signal-to-noise ratio of typically 60–70 dB, this medium suffers from mechanical problems such as flutter and wow. Moreover, the quality of an analogue speech recording severely degrades after it has been copied repeatedly. Because of these drawbacks, it is strongly recommended to use digital media for the recording of speech. The most widespread digital medium for recording of speech signals is the DAT (Digital Audio Tape). This medium is strongly recommended. Recordings are made on two channels with standard sampling frequency of 48 kHz, and 16-bit resolution. Another option, that can only be used in a laboratory environment, is to record the speech directly on a high capacity computer disk. Two other digital audio media, the CD-ROM and WORM (Write Once Read Many), are less suitable for speech recording, because they cannot be erased. That is, data (for instance, speech recordings) can only *once* be written to a CD-ROM or WORM; afterwards, the stored data can be read as many times as one likes (compare a gramophone disc). The CD-ROM and WORM are especially useful for the permanent storage of selected recordings in a database.

The recording devices can be characterized according to the following criteria:

- portability,
- capacity, and
- ease of use.

The *portability* of a recording device is determined primarily by its size and weight, and secondarily by its operating requirements, e.g. power supply, environmental conditions, etc.

Tape drives, analog or digital, come in all sizes, including Walkman-sized DAT recorders. Usually, tape drives are optimized to record or playback signals, i.e. they do not produce very much noise themselves. Portable tape drives usually have only a reduced set of features, they can operate on batteries and are quite immune to adverse environments (some are even water-resistant). Non-portable tape drives offer more features (e.g. remote control, manual setting of recording parameters, computer interfaces), require a permanent power supply and operate in the usual office environments.

Computers too can be divided into portable and desktop computers. In general, they produce significant noise during operation (hard disk spinning, keyboard clicks, system alerts etc.) and must thus be shielded from the signal to be recorded. Furthermore, sound cards in computers are subject to interferences from other devices inside the computer, e.g. noise from the bus, the processor etc. Portable computers are about the size of an A4 book and weigh approx. 2 kg. At present, only high-end portable computers are equipped with signal processing facilities (e.g. signal processor, 16 bit quantization, sample rate ¿ 8 KHz) required for speech recordings.

The *capacity* of tape drives is almost unlimited because full tapes can be replaced by empty ones quickly and at low cost. Typically, an analog compact cassette holds about 90 minutes of stereo signals, a video cassette up to four hours, a digital DAT tape up to two hours.

The capacity of computers for speech recording is mainly limited by the capacity of the hard disk. A 1 Gigabyte disk can store approx. 8 hours of mono signals (16 bit quantization, 16 KHz sample rate). Such disks are becoming common on many desktop computers and even in portable computers, so that hard disks are suitable recording devices for very many speech recordings already. The major limitation of recording to hard disks is that the hard disk cannot simply be exchanged against another one. This means that the data on a hard disk has to be saved to some backup medium, e.g. magnetic tape or CD-ROM.

Ease of use must be seen under two aspects: first, the ease with which the device can be used to perform the recording; second, the ease with which the recorded data can be accessed for further processing.

Tape drives are easy to set up and speakers are used to them. Howver, especially for analog recordings, it is quite cumbersome to access recordings for further processing. The appropriate tapes have to be located and the tape drive has to be attached to a computer.

Computers as recording devices are still uncommon. They require the expertise of an engineer to be set up correctly, and speakers are easily distracted by the presence of a computer. However, computers offer significant advantages over tapes: recordings can be fully automated, administrative data is collected together with the recordings, and data is available immediately, either for control purposes or further processing.

4.4.1.3.1 Recommendations

- Use digital recording devices.
- Use a computer for the recording to automate recording procedures and for easy access to data for further processing.

4.4.2 Speaker Recruitment

Speaker recruitment can be characterized along the following dimensions:

- number of speakers
- speaker qualifications
- demographic distribution
- speaker availability (once/repeatedly; within short/long period)

Speakers should be thought of as a valuable resource in speech recordings. It is therefore advisable to build a *speaker database* which contains for each speaker

- a unique speaker id
- administration data (name, address, telephone)
- personal information (place and date of birth, languages, education, etc.)
- physiological data (sex, size, weight, etc.)
- speaker history (list of recordings, etc.)
- remarks

Preferably, such a database is implemented using a database management system on a computer. This way, data can be entered easily during the preparation of a speech recording. Such a database can also be held on forms in a folder, but then the extraction of speakers according to specific criteria other than the primary ordering criterium is difficult and error-prone.

Recruiting a small (i.e. 1 to 5) or medium (5 to 50) number of speakers is no problem. Depending on the requirements, colleagues, friends, and relatives can be asked to participate. However, one cannot expect any demographic balance in small sets of five or less speakers. The advantage of using friends and relatives is that they may be available for a long period of time, and that they in general can be used for more than one recording.

The recruitment of a large to very large number of speakers is completely different from that of a small number of speakers. Accessing the speakers, scheduling their recordings, evaluating the recordings, and storing the data become so large tasks, that they cannot easily be performed by a single person. Accessing a large number of speakers requires either

- a contact address,
- a public call, or
- a hierarchical recruitment scheme.

Contact addresses, i.e. telephone number or postal address, are expensive: address brokers charge for each address bought, with the risk of the address being useless (the address is wrong, the person is not willing to co-operate, etc.) entirely upon the buyer. Market research institutes have large address databases from which they can select subsets according to specific criteria, but they in general do not give away these addresses. Although addresses allow persons to be contacted directly, e.g. through mail, telephone, or interviewer visits, the rate of return is rather low, typically in the range of less than 5for mail, 25 % for telephone, and 50 % for interviewer visits.

Public calls for participation, e.g. newspaper advertisement or article, Internet posting, radio or TV announcement, may reach a very large audience. In many cases, a public call can be arranged at little expense — newspapers, especially the science editors, are willing to co-operate, Internet postings are virtually free, and radio or TV announcements are affordable). The rate of return is usually very low (less than 1 %) but this is compensated for by the sheer size of the audience reached. However, the means to determine the response to a call for participation are limited. Also, the number of callers will not be evenly distributed over time (most people will call immediately after having received the call), which may cause capacity problems. People responding to a public call for participation are highly motivated; however, this does not hold for the population as a whole and thus introduces a bias.

In *hierarchical recruitment* the task of recruiting m speakers is divided into n tasks of recruiting m/n speakers. Hierarchical recruitment works well if the burden of recruiting speakers can be mapped to some real-world hierarchy, e.g. the employee hierarchiy in a company. The rate of return strongly depends on the success of a person persuading others to participate.

In all three recruitment strategies, incentives may help to increase the motivation to participate and thus the rate of return. Incentives can either be gifts (e.g. telephone cards) or the participation in a lottery with a grand prize. However, such incentives clearly make the recruitment of speakers even more expensive.

4.4.2.0.2 Recommendations

- Speaker addresses are a valuable resource. One goal of the speaker recruitment should thus be to establish or extend the database of speakers.
- In a project with a large project partner, e.g. a large company, the hierarchical recruitment scheme should be tried.
- If sufficient time is available, then a mix of a public call for participation and contact addresses is promising: a public announcement provides an address from which further information, prompt sheets, etc. can be obtained. Any such request is then registered and the person can be addressed directly.
- If time is short and manpower is limited, then speaker recruitment via a market research institute is advisable. This is especially true if there is a high degree of interaction between the speaker and the recording institution, e.g. agreements to be signed, posting of incentives, etc.

4.4.3 Scheduling speakers

Scheduling speakers is important to make optimal use of the recording capacities within a given period of time. Proper scheduling avoids speaker frustration (caused e.g. by having to wait, ever-busy telephone lines, etc.) and allows a maximum number of recordings within the given recording capacity.

If speakers are recorded in a studio, a time slot is reserved for each speaker. This time slot must be sufficient for

- the technical setup of the recording
- the registration of the speaker
- the instruction and accustomization of the speaker
- the speech recording session itself
- a cleanup procedure (backup of data, reset of technical equipment, etc.)

In general, five minutes for each of the side-tasks should be sufficient.

If speakers have to travel far then it is almost inevitable that some of them come late or do not appear at all. In such cases it is advisable have some speakers available upon short notice. In any case there must be a person responsible for the scheduling, and this person must be reachable directly by telephone.

For telephone recordings, the number of speakers calling at any one time must be matched to the capacity of the telephone equipment. If potential speakers do not get through because of busy lines, they are likely not to retry. Furthermore, telephone recordings should be possible 24 hours a day, or it must be clear to callers that the service is operational only for a specific period during the day. Note that recording 24 hours a day requires that the recordings be performed automatically because only in rare cases will human operators be available for 24 hours. Again, speakers must be able to reach an operator via telephone, e.g. to report problems or make suggestions.

4.4.3.0.3 Recommendations

- Spread the number of speakers evenly over the recording period, but leave the last 10 - 15% of the period as a reserve to cope with delays and interruptions
- In studio recordings, assign a time slot to every speaker.
- In telephone recordings, ensure that the recording system is operational without supervision, and that the number of speakers is within the capacity of the telephone system.
- One person should be responsible for the scheduling of recordings. This person must be reachable directly via telephone.

4.4.4 Speaker Prompting

Speaker prompting is used to elicit directly from a speaker a certain type of speech data, e.g. numbers, dates, times, etc. Such data is much more difficult to obtain in dialogues or role-play. The major problems with speaker prompting are a decrease in the spontaneous quality of the speech, ambiguous prompts that lead to unexpected responses, and the rigid structure of a prompting script, i.e. a sequence of prompts. However, in many applications prompting a user for input is a natural situation, and thus the decrease in spontaneity in the speech is highly welcome.

Four types of prompts can be distinguished:

information or feedback prompts , e.g.
"You have now completed half the questionnaire."

instruction prompts , e.g.
"Please read the number under topic 1)"

questions , e.g.
"what day of week is today?"

repetition prompts
in which a speaker is asked to repeat what has been prompted, e.g. "three four five".

The possible responses to prompts may vary greatly. It is thus advisable to instruct the speaker which responses are expected, e.g. "Please answer the following questions with yes or no". However, restricting the set of allowed responses too strongly will lead to unnatural speech.

In face to face communication situations, there is an influence of the interviewer onto the speaker even when the catalog of prompts is fixed, e.g. in a prompt sheet. Visual communication, deictic references (like pointing with a finger to an item to be read), play a significant role. The interviewer guides the speaker through the script and may immediately correct any errors.

In telephone recordings prompts may be output by a computer or a human interviewer. The advantage of computer prompting is that all speakers hear identical prompts. One disadvantage is that a computer based prompting system strictly follows a predetermined prompting script and may not notice that the speaker is not responding correctly. Computer prompting scripts should thus not take longer than 15 minutes, and the script should be divided into several small units. Between the units, feedback should be given to the user to inform him or her of the status of the recording.

Human interviewers immediately realize whether a response from a speaker is correct, and they are able to correct wrong responses immediately. However, each prompt is an individual utterance so that variations among responses may also result from the prompts.

Each prompting script, should be thoroughly tested before the actual recording of data. The test participants should be candidate speakers, and test conditions must be as similar as possible to those of the actual recording. In the case of computer prompting, it is useful to have a prompting simulator which can be adapted to new prompting scripts easily.

4.4.4.0.4 Recommendations

- Formulate prompts so as to avoid ambiguities
- Restrict the number of allowed responses
- Test prompting script on 1% (minimum 10 speakers) of the candidate speakers under realistic conditions
- Give sufficient feedback to the speaker (after one third, half, and three quarters of the script)

4.4.5 Cost

The cost of a speech recording is determined by the cost for personnel and equipment and by the period of time. The total cost estimate is defined in a budget, and at given times the actual expenditures are compared to the budget plan.

A speech recording project is usually defined by scientifically trained experts, e.g. speech engineers, phoneticians, etc. Once the project has been defined, the data recording can begin.

The minimum personnel requirements for speech recordings (of a large number of speakers) are a project administrator and supervisor, and a system operator; both should be available for the whole recording period. Depending on the speech recording setup and the processing of the signal data, interviewers, scientific personnel, and "Hilfskräfte" are necessary.

The administrator is responsible for the budget and the supervision of the project as a whole, the recruitment of speakers, scheduling of recordings, and the organisation of the data evaluation. The system operator is responsible for the technical and data processing aspects of the speech recording, i.e. the setup of equipment, storage and backup of data, etc. Interviewers are needed for speech recordings in face to face communication situations. A first evaluation of the technical quality of recordings can be performed by "Aushilfskräfte", whereas further processing, e.g. the transliteration or a phonetic segmentation and labelling of utterances require trained experts or scientific personnel.

The cost of personnel is the sum of salary and related infrastructure (room, desk, computer, telephone) and working materials costs. In many cases existing resources can be reused, but it should be clear that they have to be accounted for in the budget.

The cost for equipment consists of the acquisition and maintenance costs. Again, in many cases existing equipment can be reused and it must be accounted for in the budget. Maintenance costs are significant cost factors which often exceed the original acquisition costs. For time-critical projects, maintenance contracts with a guaranteed repair time should be considered.

A speech recording can be divided into the following phases:

- initialization and test phase
- preparation
- recording
- evaluation
- clean up

All phases are strictly sequential except for recording and evaluation which can be executed in parallel.

Initialization and test, preparation and cleanup take roughly constant time. The initialization and test phase must be considered very important because wrong decisions here will affect the rest of the project. Preparation can be short if the initialization and test phase results in a good procedural setup.

The duration of the recording and evaluation phases depend directly on the number of recordings. As a rough estimate, double the speaking time (prompts and responses) to get an estimate of the time needed to perform an individual recording (speaker instruction, cleanup, etc.). Depending on the quality of the evalaution, the time needed for evaluation may be double (technical evaluation) to ten times (phonetic evaluation and transliteration) the speaking time.

4.4.5.0.5 Recommendations

- Define a detailed budget for the project as a whole and the speech recording tasks
- Assign a project supervisor and a system administrator; both persons should be available throughout the project
- Exploit parallelization of tasks
- Include overhead in your time calculations: for recordings, 100% overhead on speaking time, for evaluation 100 - 1000%.

4.4.6 Multi-Channel Recording

For some applications one must record a number of physiological signals besides the acoustic signal, such as a laryngograph signal, an electromyograph signal (EMG), air pressure or flow in the vocal tract, articulatory parameters, X-ray data, etc. The major drawback of recording such additional signals is that speakers have to be bothered with measuring equipment, such as a strap with electrodes round the neck in the case of laryngographic recordings. One should be aware that the measuring equipment may interfere a natural speech production. Therefore, it is recommended to use additional signal recordings only for basic speech research and for specialised purposes, such as examination of voice pathology, and else confine oneself to the basic acoustic signal. For some applications (audio-visual analyses) it may also be useful to make video recordings.

- Electroglottography (Laryngography)
- Electropalatography
- Electromagnetic Articulography (EMA)
- Cineradiography
- Air-flow measurements
- ...

to name a few.

Most of these multi-channel recordings require a high technical effort. The placement of sensors may disturb the speaker (X-rays are even dangerous), and due to the considerable effort, only few corpora of these kinds of measurements exist.

4.4.6.0.6 Laryngography

In a laryngograph recording a pair of electrodes is attached to the throat of the speaker on each side of the thyroid cartilage (Adam's apple). This sensor produces a signal proportional to the amount of contact between each vocal fold, e.g during phonation.

Laryngography recordings were taken at the Eurospeech 93 conference in Berlin, where speakers were recorded during their presentations. The data is available in the TED corpus from the BAS (bas@sun1.phonetik.uni-muenchen.de).

Requirements: Laryngography sensors, DAT tape recorder or computer interface; 8 bit quantization approx. 10 kHz sample rate.

4.4.6.0.7 Electropalatography

Electropalatography registers the contact of the tongue with the hard palate during articulation. The speaker places a customized thin artificial palate in his or her mouth. This artificial palate contains an array of electrodes which record contact with the tongue.

The data recorded by each electrode is combined to a two-dimensional representation of the palate at any given point in time.

¡picture here¿

Requirements: Artificial palate individually tailored to a speaker, multi-channel recording device, e.g. computer with a suitable interface; 64 bit quantization (i.e. typically an 8 * 8 array), sample rate 200 Hz.

4.4.6.0.8 Electromagnetic Articulography

Electromagnetic articulography (EMA) measures the movement of the tongue and other articulators through tiny induction coils attached to the tongue. The head of the speaker is enclosed by a helmet which holds two (or more) coils that create an electromagnetic field; The signal induced in the coils on the tongue is proportional to the distance from the transmitter coils on the helmet.

¡picture¿

The EMA provides essentially the same kind of data as the microbeam X-ray (see section ...) but uses a different technology.

Requirements: Articulograph, multi-channel recording device, e.g. a computer with a suitable interface; data rate depends on the quantization, the number of sensors, the number of transmitters (typically 10 sensors and 3 transmitters), and the sample rate (typically 250 Hz).

4.4.6.0.9 Cineradiography

X-ray measurements are rarely performed today because of the health hasards they impose on the speaker. However, early recordings are still available on film or, in digital format, on laser disk (Bateson at ATR, Japan).

X-ray measurements show the modification of the articulatory tract during articulation. The movement of the jaw can be seen clearly; tongue and lip movement are often less clear due to the fact that they do not show up very clearly on X-ray. The movement of the vocal folds is too fast to be recorded at the slow frame rate of film recordings.

Requirements: Seldom performed.

4.4.6.0.10 Air-flow measurements

In air-flow measurements the speaker wears a mask (usually designed to separate oral and nasal airflow). Flow is usually derived from the pressure drop across a wire-mesh located in a flow head mounted in the mask ¡picture¿

The measurements yield data on the speed, direction, and volume of air flow. Depending on the type of sensor and attachment, the measurement requires that the speaker does not move during articulation.

Requirements: Air flow sensors, data acquisition hardware. The data rate depends on whether phonatory components of airflow need to be captured.

4.4.6.0.11 X-ray microbeam

X-ray microbeam provides two-dimensional movement data (usually in the mid-sagittal plane) of selected fleshpoints on the tongue and other articulators. It uses a point-tracking technique to reduce the radiation exposure to the subject to acceptable levels.

Requirements: The equipment is only available at the dedicated microbeam facility in Madison, Wisconsin. Data rate: each fleshpoint is tracked at about 100-200Hz. Typically about 10 fleshpoints are tracked simultaneously.

4.4.6.0.12 Nuclear magnetic resonance imaging

Nuclear Magnetic Resonance Imaging is a static (up to now) imaging technique with very good resolution of the soft tissues in the vocal tract. Slices can be freely chosen i.e sagittal, coronal etc.

Requirements: a friendly hospital; sample rate ¡ 1Hz, image resolution 256*256 pixels (typical, with 8 bits pixel depth)

4.4.6.0.13 Ultrasound imaging

Ultrasound imaging can be used for obtaining sagittal and coronal images of the tongue(for those locations on the tongue where no air intervenes between transducer and tongue; the transducer is usually held under, and moves with the jaw).

Requirements: Ultrasound machine. The data is usually stored as standard video data. A frame-grabber is needed if data is to be digitized.

4.4.7 Wizard of Oz

section still missing — can anyone serve with a practical guide on how to set up a WOz experiment?

4.4.8 Legal Aspects

There are differences between countries in legal objections against collecting, storing and disseminating demographic and personal characteristics of speakers in a corpus. In all countries name and address data may freely be stored and published, as long as it is guaranteed that no other data about the person can be linked to name/address. Limitations in this respect must be carefully checked with legal advisers in the country or countries of interest. Another legal issue that must be considered is the consent to record speech and subsequently to make the recordings available to other parties. In legal terms, making available recordings is probably equivalent to publishing them. There are likely to be differences between countries in legislation about recording speech. In some countries recording is legal as long as one of the parties involved in a dialogue gives explicit consent. Probably, the law only states that recording for own use is legal; publication or any other way of dissemination may be illegal, or bound by much stricter regulation. To avoid problems in this area, it is recommended to have speakers sign a statement to the effect that they know that the recordings are made for later dissemination and publication. If that is not possible, e.g. when collecting large corpora of telephone speech, it is recommended that each speaker is explicitly advised to abort the call if (s)he has any objections against recording and subsequent publishing of the speech to be produced. It is very important to consult with legal advisers about the correct formulation of these statements and advises. Special care must be taken when recording and publishing corpora with pathological speech or with speech of very young children who are not yet able to give conscious consent for publishing. Especially with rare pathologies it may be very easy to trace the speech back to the patient, even if name and address data is not coupled to the data about the pathology. With pathological speech of mentally healthy adult patients a carefully formulated written and signed consent form may be sufficient. With speech of very young children carefully formulated consent forms signed by the parents are necessary. Legal issues (and ethical ones as well) are especially relevant in the case of surreptitious recording of speech. In order to circumvent the observer's paradox (see section 3.4.1.7), researchers might want to resort to surreptitious recording of speech. This could, for instance, be done with a concealed microphone 'in the field' or by tapping telephone conversations. In a strict sense invading the privacy of someone's personal speech might be regarded as illegal by the authorities. And even if it would not be illegal, it might still be regarded as unethical by some people. As far as we know, no linguist has ever been tried for recording speech data surreptitiously. However, to minimize the risk of breaking the law in any way, and to conform to ethical norms as much as possible, the following guidelines can be taken into consideration:

- Be sure that recordings are made only to gather speech data that could not be otherwise obtained. Also remember that any possible tortious or criminal motive would certainly render a surreptitious recording illegal.
- To protect the privacy of persons who were surreptitiously recorded as much as possible, do not disclose their identities to anyone, do not use real names in published examples, and do not publish long transcripts in which the speakers could be identified.
- The invasion on privacy strongly depends on the context in which the recordings were made. For example, surreptitious recording of conversations in a public place, such as a bar, a supermarket, or department store, would be less invasive on the privacy of people than surreptitious recording of conversations in a bedroom. In public places speakers probably produce speech with the full awareness that some passerby may hear what is being said. As another example, surreptitious recording of telephone conversations will be more invasive on the privacy of people when they talk about confidential matters concerning various intimate details about their lives than when information about plane schedules is asked (as, for instance, in Wizard of Oz indexWizard of Oz technique experiments). Also the age of the speakers who are being surreptitiously recorded may be important with respect to confidentiality; in general, the speech of (young) children will be regarded as less confidential than the speech of adults.
- And last but not least, it is strongly advised to inform people afterwards that their speech was surreptitiously recorded and offer the opportunity to review and/or destroy all or part of the recordings. The argument for debriefing is perhaps somewhat less compelling when the researcher stays away from private conversations, but both from a legal and an ethical point of view debriefing would always be preferable.

A more elaborate discussion of legal and ethical aspects of surreptitious recording of speech can be found in Larmouth (1985) and Murray & Murray (1985).

Chapter 5

Corpus representation

5.1 Introduction

In this chapter two important stages in the representation of spoken language corpora are discussed:

- Transcription of the speech.
- Design of the data base.

Transcription means that a symbolic representation of the speech is made. The transcription makes it possible to easily find your way in a speech corpus. Design of the data base means the way in which speech, speaker information, information on data collection, and transcriptions are stored in a structured way. For both transcription and data base structure no firm standards exist at this moment. Partly this is due to the fact that the representation of corpora very much depends on their future use. However, general remarks and recommendations can be made. It is expected that the need for standards will grow in the near future as the number of corpora which are collected increases and the exchange of corpora will be greater. Furthermore, the foundation of the European Language Resources Association (ELRA) in 1995 will stimulate the standardisation of spoken language corpora.
In section 5.2 transcription of spoken language corpora will be discussed and in section 5.3 we will deal with the storage and the data base structure of corpora.

5.2 Transcription of spoken language corpora

Transcriptions are used in many fields of linguistics, including phonetics, phonology, dialectology, sociolinguistics, psycholinguistics, second language teaching, and speech pathology. Transcriptions are also used in disciplines like psychology, antropology, and sociology. The type of transcription very much depends on the aim for which the transcription is made. This aim especially determines the degree of detail that is required. For example, if a corpus has been designed to investigate the amount of time several speakers are speaking at the same time in a dialogue, a very global transcription will be sufficient. If a corpus has been collected to initialise the training of a phoneme-based speech recognition system, one might want to have a very precise, segmental transcription. The precision of transcription can vary between these two extremes. Detailed, segmental transcriptions of large scale spoken language corpora with many speakers and much (spontaneous) speech cannot easily be achieved. This would be too time-consuming and too expensive. Therefore, most large speech corpora are provided with verbatim transcriptions, i.e. word level, orthographic representations of what has been said, see e.g. the ATIS and Switchboard corpora. However, a medium sized corpus of read speech can be provided with a segmental transcription and even with labelling at the segmental level, see e.g. the TIMIT corpus, which consists of 630 speakers each reading 10 sentences, and also the Phondat (1990 and 1992, both read speech) and Verbmobil (1993, spontaneous speech) corpora. An important point that should be made here is that all types of transcriptions are the result of an analysis or classification of the speech. Transcriptions are *not* the speech itself, but an abstraction from it. Yet often, transcriptions are used as if they were the speech itself. It is evident that this can only be done if transcriptions are made very consistently and precisely. In section 5.2.2, levels of transcription are presented, starting from global orthographic transcriptions and ending with detailed phonetic transcriptions. For all types of transcriptions, we will emphasise their use in spoken language corpora meant for linguistic research as well as speech technology. In the next section some remarks will be made about transcriptions of read versus transcriptions of spontaneous speech.

5.2.1 The transcription of read vs. the transcription of spontaneous speech

The transcription of read speech starts from the written texts, which makes them somewhat easier to perform than transcriptions of spontaneous speech where everything needs to be written down from scratch. In the case of read speech, planning and word seeking processes are not involved. These processes of spontaneous speech have great impact on the speech that is produced. We all know that spontaneous speech is not perfectly fluent: speakers produce many filled pauses, mispronunciations, false starts, and repetitions. In addition, depending on the formal setting, speakers will use colloquial speech and sloppy pronunciations. These properties of spontaneous speech make that all types of transcriptions, global as well as precise ones, are more difficult to perform for spontaneous speech than for read speech. In case of read speech, the written texts will induce less disfluencies and less sloppy pronunciations.

Another important distinction between read and spontaneous speech in relation to transcriptions is that for read speech it is clear what an utterance is: the written sentence, mostly starting with a capital letter and ending with a period, question or exclamation mark. For spontaneous speech this is not immediately clear. Depending on the type of spontaneous speech, one has to define the criteria for utterances. For dialogues and other forms of conversation in which more than one speaker is involved, it is usual to let utterances more or less correspond with speaker turns (see the Guidelines issued by the Text Encoding Initiative (TEI) in Sperberg-McQueen & Burnard (1994) and Switchboard). For monologues, utterances can be defined as stretches of speech mostly preceded and followed by a pause and having a more or less consistent syntactic, semantic, pragmatic, and prosodic structure (see the criteria developed by the Network of European Reference Corpora (NERC) in French (1991) and French (1992) and the Dutch Speech Styles Corpus (Den Os (1994))).

5.2.1.1 Transcription of dialogues

When two (or more) persons are conversing together interruptions frequently occur. This is true for informal conversations between friends, for formal information requests, for face-to-face situations, and for telephone conversations. These interruptions may be complete utterances, but also e.g. affirmative "yes", or "mm". These interruptions resulting in simultaneous speech, must be annotated in the transcriptions. If a dialogue between two persons is concerned, it is possible to indicate simultaneous speech in a clear way. For example Switchboard uses "#", one before and one after each of the segments spoken at the same time, to indicate that the two speakers of the telephone conversation speak at the same time, e.g.
A: # Right, bye #
B: # Bye bye #
However, when more than two speakers are conversing, it is not possible to indicate the interruptions and simultaneous speech in a clear and simple way. For these cases a so-called "score notation" can

be used. As for music score notation, the different speakers are given a separate track, one above the other. The tracks must be synchronised. A computer programme "syncwriter" that handles these types of conversation has been developed for the Apple Macintosh. Information about this programme can be obtained at ger.xse0014@applelink.apple.com.
It is also possible to collect dialogues in which simultaneous speech is avoided. In part of the Verbmobil corpus the dialogue partners are recorded separately; the partners press a button when they start to speak and this directs the recordings. The recordings are made in two rooms, separated by a glass screen, so that the speakers can see each other. The speakers can hear each other by means of head phones. Of course, this situation is not as natural as when both speakers are allowed to speak at the same time.

5.2.2 Levels of transcription

All types of speech, whether read or spontaneous, whether monologues or dialogues, can be transcribed at different levels. The most global one, the orthographic transcription, will be used in the case of large corpora, or in cases in which the research questions can be answered on the basis of a global indication of what has been said. More precise transcriptions will be used for smaller corpora. A distinction can be made in the following types of transcriptions:

- Orthographic transcription,
- Phonemic transcription,
- Allophonic transcription,
- Phonetic transcription,
- Prosodic transcription.

An orthographic transcription gives a verbatim description of what has been said using the normal spelling of a language. In a phonemic transcription each symbol refers to a single phoneme, in whatever context this phoneme occurs. In an allophonic transcription a single phoneme may be represented by different symbols, depending on the structural or environmental context in which the phoneme occurs. Sometimes a phonemic transcription is called a *broad transcription*. Phonemic and allophonic transcriptions provide generalised, abstract statements about the pronunciation of a language or dialect. The transcriptions are also called *systematic transcriptions*, since they are used to make explicit observations on rule-based generalisations about regular, patterned activities in the accent concerned (Laver (1994: 550)). A phonetic transcription makes more or less specific comment on phonetic details of pronunciation. Therefore, it is sometimes called a *narrow transcription*. A combination of phonemic and allophonic transcription is possible, e.g. when only the allophones of certain phonemes are given. In a prosodic transcription prosodic features like intonation, rhythm, and accents are presented. In the following we will deal with the different levels of transcription in relation to spoken language corpora.

5.2.2.1 Orthographic transcription

By orthographic transcription (or transliteration) we mean that the standard spelling of a given language is used to indicate the spoken words. Orthographic transcriptions are used in large scale speech corpora and in corpora used for research in which details about pronunciations of words are not important, e.g. in lexical research. In the first case a detailed transcription of at least part of the corpus is perhaps desirable (especially if such a corpus is used for the training of speech recognition systems), but because of the huge amount of work this cannot always be done. In the latter case precise transcriptions are not even necessary.

The use of orthographic spelling necessarily implies a compromise between the sounds heard and what has to be written down. Especially for spontaneous speech the difference between what is heard and the symbolic representation may be large.

5.2.2.1.1 Reduced word forms

Because of the discrepancy between what is heard and what has to be written down, it is has been decided in many spontaneous speech corpora to indicate reduced word forms. One should use the reduced word forms which are given in the standard dictionary of a given language. However, for the sake of consistency one is sometimes forced to use forms which are not present in the dictionary. For example, in German prepositions and articles are written as a single word in case of a reduction of the number of syllables ("zu der" is written as "zur"). These are existing forms in the Duden dictionary. However, in the Verbmobil corpus one is allowed to write "fürn" for "für den", although this form does not exist in the dictionary. Also for the Dutch Speech Styles corpus it was decided to indicate reduced word forms. Criteria for indicating reduced forms in a transliteration may be a) frequency of occurrence of these forms and b) reduction in the number of syllables.

5.2.2.1.2 Dialect forms

Even in speech corpora which are meant to collect the standard variant of a given language, speakers may have their own idiolect or may use words that have a dialect basis. These words have to be marked in the transcription. Verbmobil chose to indicate dialect words which are not in the Duden dictionary in an orthographic way. It is possible to give information about the meaning behind these words, for example:
moin, moin <; norddeutsche Grušformel>, that is the dialect form of "good morning" spoken in the North of Germany.

5.2.2.1.3 Numbers

In orthographic transliterations numbers are usually not represented as digits, but they are written down as words. In some cases it is decided to deviate from the standard spelling in order to avoid very long words. For example, in the Verbmobil corpus the numbers 13 till 99 as well as the hundreds from 1 till 19 (the years) are written as a single word. However, all the other numbers are written separately, not conforming to the rules of the dictionary. Thus:
1993: neunzehnhundert dreiundneunzig
349614: drei Millionen neunundvierzig tausend sechshundert vierzehn
349: dreihundert und neunundvierzig

5.2.2.1.4 Abbreviations and spelled words

In orthographic transcriptions one usually writes the full forms of abbreviations: "a.o." is written as "among others", German "usw." is written as "und so weiter". Abbreviations which are spoken as words are spelled as words, e.g. Benelux, OPEC, NATO.
In spoken language corpora like Polyphone, one type of items speakers had to read was spelled words. Words can be spelled in different ways: the names of the letters can be pronounced, like A, B, C, but one can also use words beginning with the letter concerned, like Adrian, Bernard, Christian. In transcriptions spellings have to be indicated, also when they do not concern whole words, but only part of words: USA-trip, Vitamin-C. In the Vermobil corpus, spellings are indicated

by capitals preceded by a space (except in case there is a hyphen) followed by $: $U $S $A-trip, Vitamin-$C.

5.2.2.1.5 Interjections

Interjectives like "au", "ah", "oh", and "mm" , or the French "hein" must be indicated according to the standard spelling of these forms in the given language. If there is no standard spelling for a certain interjective, one has to decide for a certain spelling.

5.2.2.2 Orthographic transcription of read speech

When orthographic transcriptions are used for corpora containing read speech, the original written text may function as the default transcription. The transliteration indicates how well the written text was read by the speaker. For single words or short sentences, speakers will make relatively few mistakes, as has been experienced with e.g. the Dutch POLYPHONE Corpus and the Groningen Corpus. For read texts (even short ones) it turns out that speakers often make mistakes. These mistakes are mostly related to deletions of words, false starts, or hesitations. They may also use words that were not in the texts or they may use a different word order. Furthermore, speakers may mispronounce words. For example, they may add syllables, leave them out, or they may scramble syllables.

Depending on the use of a corpus of read as well as spontaneous speech, these disfluencies must be indicated in the transcriptions. If the corpus will be used for initialising speech recognition systems for instance, all sounds must be indicated, including hesitations, filled pauses, etc. One can also think of research on reading errors which will need these disfluencies to be written down. On the other hand, if a corpus will be used to find the type of syntactic structures that people use in a certain dialogue system, it is not necessary to indicate all events that occur in the signal. In section 5.2.5 we will give an overview of annotations which can be used in a transcription. These annotations will concern types of (non-) verbal sounds by the speakers (hesitations, coughing, laughing), but also types of background noises (slamming of doors, ringing of telephones).

5.2.2.3 Phonemic transcription

By phonemic/phonotypical transcription we mean transcription based on the phonemes of a given language. A phonemic transcription is sometimes called a broad transcription. Of course, phonemic transcriptions cost more time than orthographic transcriptions. One has to listen more precisely to what has been said. For phonemic transcriptions, the IPA (International Phonetic Association) system is mostly used. SAMPA is considered to be a computer version of part of the IPA system (see Appendix A). SAMPA is based on ASCII characters. There are also other computer versions, as the one indicated in the Table.

Table 5.1: *Table of number-to-symbol correspondence for the LaTeXfont wsuipa11.*

0	ɐ	32	ɣ	64	∞	96	ʒ
1	ɑ	33	ɤ	65	ᵽ	97	ʓ
2	α	34	ɧ	66	þ	98	ʔ
3	ɒ	35	ħ	67	ɸ	99	ʕ
4	ʌ	36	ɦ	68	ɾ	100	ʖ
5	ƀ	37	ɧ	69	ɼ	101	ˀ
6	ƀ	38	ɥ	70	ɽ	102	ˁ
7	b̸	39	ɨ	71	ɹ	103	⌐
8	ɓ	40	ᵻ	72	ɻ	104	
9	β	41	ɩ	73	ɺ	105	ˈ
10	ȼ	42	ɪ	74	ʀ	106	˥
11	ç	43	ᵻ	75	ʁ	107	
12	ƈ	44	ɟ	76	ʂ	108	⊥
13	đ	45	ƚ	77	ʃ	109	⊤
14	đ	46	ɫ	78	ʆ	110	⊣
15	d̸	47	ɬ	79	σ	111	⊢
16	ɗ	48	ɭ	80	ʈ	112	ˑ
17	ɖ	49	ɮ	81	ʧ	113	ː
18	ʤ	50	λ	82	ɟ	114	¨
19	ð	51	ƛ	83	θ	115	˓
20	ᴅ	52	ɱ	84	ʉ	116	˄
21	ə	53	ɯ	85	ɏ	117	˅
22	ɚ	54	ɰ	86	ʊ	118	˂
23	ɘ	55	ɲ	87	ᴜ	119	˃
24	ε	56	ŋ	88	ᵾ	120	˚
25	ɜ	57	ɳ	89	ʋ	121	˳
26	ɝ	58	ɴ	90	ʍ	122	˓
27	ʚ	59	ʘ	91	χ	123	˜
28	ɡ	60	ɵ	92	ʎ	124	˷
29	ɠ	61	ɔ	93	ʏ	125	ˇ
30	ɢ	62	ω	94	ʑ	126	˓
31	γ	63	ɷ	95	ʐ	127	⁀

It is possible to convert an orthographic transcription into a phonemic representation by means of automatic grapheme-to-phoneme conversion. Of course, one cannot be sure that the phonemic representation that has been obtained in this way, is the one that was actually pronounced by the speaker. Possible variations in the pronunciation of words are not incorporated in the present text-to-speech systems. For many purposes, the automatic grapheme-to-phoneme conversion will be sufficient. However, for initialising training of speech recognition systems this level of transcription is perhaps too global.
It is possible to use the automatically obtained phonemic representation as the starting point of the phonetic transcription of what has actually been pronounced.

5.2.2.4 Allophonic transcription

Phonemic transcriptions focus on the phonemic contrasts in the language, whereas allophonic transcriptions focus on the ways in which spoken material reveals differences associated with structural and environmental context. If different symbols are used for a single phoneme when this phoneme occurs in different contexts, one speaks of an allophonic transcription. It is possible to combine phonemic and allophonic transcriptions, e.g. by indicating only the allophonic variants of a subset of phonemes.

5.2.2.5 Phonetic transcription

A phonetic transcription is even more precise than a phonemic one. A phonetic transcription represents the pronunciation of words of individual speakers. No attention is paid to phonological theory. Allophonic and phonetic transcriptions are sometimes called narrow transcription.
Diacritical marks like the ones for labialisation, nasalisation, synchronic articulation are used to indicate the exact pronunciation of sounds. Furthermore, special signs are used to indicate length, stress, and tones. This narrow phonetic transcription is very time consuming to make and cannot be done for large speech corpora. It has been found that providing reliable phonetic transcriptions for large corpora is hardly feasible (cf. Cucchiarini (1993)). However, detailed transcriptions of a small number of specific phenomena (e.g. presence/absence of diphthongation, voiced/voiceless character of fricatives) can be made relatively fast and reliably if the occurrences of these phenomena can be retrieved quickly with the aid of annotation and direct access to files offered in a computerised speech corpus (cf. Van Hout (1989) and Van Bezooijen & Van Hout (1985)).

5.2.2.5.1 The CRIL convention

CRIL (Computer representation of individual languages) conventions have been defined and proposed by a working group at the 1991 Kiel convention of the International Phonetic Association. The conventions consist of two parts:

- A mapping of IPA categories to standard IPA numbers and IPA symbol names.
- Three distinct levels of representation for natural speech production.

The first component has been introduced to enable broader use and dissemination of the descriptional IPA categories; i.e. the IPA symbols as well as the IPA diacritics needed for the narrow transcription of regular and defective speech utterances.

The second component of CRIL is devoted to a standardised representation of natural speech productions and introduces three systematically distinct levels for specifying what could be called the text of a spoken utterance. These levels are:

1. Orthographic level
   This level contains the orthographic representation of the spoken text.

2. Phonetic level
   This level specifies the phonetic form of a given word in its full (unreduced) segmental structure which only appears when a word is spoken in isolation.

3. Narrow Phonetic level
   This level gives a narrow phonetic transcription of the actually spoken words. It is only on this level that phonetic categories can be directly related to the speech signal itself.

The CRIL conventions have been shown to work very well for the characterisation of speech data in the German PHONDAT and Verbmobil-PHONDAT corpora.

During the International Conference on Spoken Language Processing (ICSLP) in Banff in 1992 a workshop was held on Orthographic and Phonetic Transcription. The workshop goals were to agree on areas where community-wide conventions are needed, to identify and document current work, and to establish a means of future communication and continued cooperation. The notes from these workshop are available through anonymous ftp (cse.ogi.edu).

5.2.2.6 Prosodic transcription

Usually, some kind of prosodic information is indicated in the levels of transcription mentioned above. This is especially true for narrow phonetic transcriptions, in which attention is paid to stress and tones. But even in orthographic transcriptions (ATIS, Switchboard, Dutch Speech Styles Corpus) some prosodic variables are marked, such as lengthening of sounds, pauses in words and utterances, emphatic stress, and intonational boundaries. A systematic and theory-based system for transcribing the intonation patterns and other prosodic aspects of English utterances, the ToBI system, has only recently been developed.

5.2.2.6.1 ToBI transcription system

For an overview of existing prosodic transcription approaches, see Llisterri (1994). Here, we shortly present the ToBI transcription system.
The ToBI transcription system (Tones and Break Indices) (cf. Silverman et al. (1992)) is meant to be a standard for prosodic transcriptions. The system makes use of speech files, fundamental frequency contours, and label files. A ToBI transcription of an utterance (next to the speech recording and a fundamental frequency contour) consists minimally of symbolic labels for the following four tiers: orthographic tier, tone tier, break-index tier, and miscellaneous tier. The orthographic tier specifies the words in the utterance using ordinary orthography. The tone tier specifies the tonal properties of the F0 contour of the utterance. The break-index tier specifies the degree of disjuncture between words in the orthographic transcription, and the miscellaneous tier is used for additional ToBI notations and for individual or local additions. The tone and break-index tiers represent the core prosodic analysis. The system does not handle all types of prosodic factors, but it is flexible enough to be adjusted to one's own needs. It is recommended to use this system if prosodic information is necessary for a corpus. Information about the ToBI system, as well as guidelines, can be obtained by anonymous ftp (kiwi.nmt.edu).

5.2.3 Segmentation and Labelling

5.2.3.1 Definition and motivation

5.2.3.1.1 Definition

Segmentation is the division of a speech file into non-overlapping sections corresponding to physical or linguistic units. Labelling is the assignment of physical or linguistic labels to these units. Both segmentation and labelling form a major part of current work in linguistic databases.

The term 'transcription' may be used to refer to the representation of a text or an utterance as a string of symbols, without any linkage to the acoustic representation of the utterance. This was the pattern followed by speech and text corpus work during the 1980s, such as the prosodically-transcribed Spoken English Corpus (Knowles, Taylor & Williams (1995)). These corpora did not link the symbolic representation with the physical acoustic waveform, and hence were not fully machine-readable. A recent project, MARSEC (Roach et al. (1993)), has generated these links for the Spoken English Corpus such that it is now a segmented and labelled database. This is the form that is most useful to researchers in speech and language technology.

The types of segments that may be delimited are of various kinds, depending on the purpose for which the database is collected. The German PHONDAT and Verbmobil-PHONDAT corpora use the CRIL (Computer representation of individual languages) conventions formulated by a working group at the 1991 Kiel convention of the International Phonetic Association. These conventions propose three levels of representation: orthographic, phonetic and narrow phonetic. The orthographic level contains the orthographic representation of the spoken text. The phonetic level specifies the phonetic form of a word in citation form. The narrow phonetic level gives the phonetic labelling of the particular token of the word that was recorded.

A more detailed system of levels of labelling has been proposed by Barry & Fourcin (1992), which includes the above three levels. Each given speech corpus will choose one or more of these levels, which are described in detail under 5.2.3.2 below, and which grew out of the SAM project for the major European languages.

The format of label (transcription) files varies widely across research institutions. The WAVES format is becoming popular, and has the advantage of being human-readable. It is advisable to use a label file format that can easily be converted to a WAVES label file, for the sake of portability across different systems.

During the International Conference on Spoken Language Processing (ICSLP) in Banff in 1992, a workshop was held on Orthographic and Phonetic Transcription. The workshop goals were to agree on areas where community-wide conventions are needed, to identify and document current work, and to establish a means of future communication and continued co-operation. The notes from this workshop are available by anonymous ftp from cse.ogi.edu.

5.2.3.1.2 A caveat

The treatment of speech as a sequence of segments to be delimited is to some extent a convenient fiction, made necessary by the requirements of speech technology. For example, it is notoriously difficult to define the boundaries between vowels and glides, or between a vowel and a following vowel. in addition, information about the place of articulation of a consonant is usually contained in its neighbouring vowels rather than the consonant itself. In the case of place assimilation, electropalatographic studies have shown that there is often a residual gesture towards the underlying segment (Nolan (1987)). Hence one cannot describe the speech signal as a simple string of discrete phones in absolute terms.

Notwithstanding the above, Roach et al. (1990) argue that the attempt to segment speech is valid, as many segments (especially some consonants) have very clear acoustic boundaries. Where clear acoustic boundaries do not exist in the speech signal, selecting a fairly arbitrary point is better than doing no segmenting at all, from the viewpoint of speech technology research. Since segmented corpora are required for training HMM-based recognisers, problems of this kind could be cancelled out by including a great deal of data of the problematic kind, so as to avoid skewing the statistical models with only one view of the boundary location.

5.2.3.1.3 Use in speech and language technology research

A segmented and labelled speech database is needed for training the HMM models used in many recognisers, as well as for testing them. In addition, such a database provides the raw data needed for deriving rules for text-to-speech synthesis (rules of duration, intonation, formant frequencies, etc.). In HMM-based speech recognition, vast amounts of training data are needed, especially for speaker-independent systems. Thus it is important for the level of segmentation and labelling to be accurate and reliable, as well as consistent across the entire speech database.

5.2.3.1.4 Use in linguistic research

A segmented and labelled speech database is also a primary resource in basic linguistic research, particularly in the case of little-researched languages. Such a database can yield fundamental information on the acoustic parameters of speech sounds of the language, as well as more detailed information on such things as patterns of duration variation according to linguistic context.

5.2.3.2 Levels of segmentation and labelling

The various levels of segmentation described below are based on those outlined in Barry & Fourcin (1992) and Autesserre, Pérennou & Rossi (1989), with some slight differences.

5.2.3.2.1 Recording script

In the case of speech read from a script, the simplest and quickest level of annotation is the orthographic form of the words in the recording script, as these are readily available. However, most researchers will need more detail than this, and so other levels must also be considered.

5.2.3.2.2 Orthographic

Even in the case of read speech, the orthographic level will be needed, since the speaker will not always follow the recording script exactly. The speaker may insert, delete or transpose material, and there may be false starts and self-corrections. The orthographic level will include the orthographic form of what the speaker actually said (this will be the initial level in the case of spontaneous speech). Decisions will have to be made on how to spell non-standard vocalisations, abbreviations, dialect words etc., and also on where to divide words in the case of languages with many compound word forms. For example, contractions (such as English “I'm”, “won't”) may need to be included in a small lexicon of anomalous forms, used together with the main lexicon when transcribing the speech orthographically. During the transcription of spontaneous monologues in the Dutch Speech Styles Corpus, this small list was found to be of a manageable size (about 30 items). The comparison of this level with the level of the recording script may be of interest to linguists studying speakers' language performance, but otherwise the level of the recording script is not of immediate interest.

Recommendation 1

When transcribing a corpus orthographically, it is advisable to generate a list of all unique word forms found in the transcription. This list will then form the input to a grapheme-to-phoneme conversion module (which may involve accessing a phonemic dictionary and/or running letter-to-sound rules). The output of this module will be the (citation) phonemic form of the speech, which can form a basis for later adaptation to various accents of the same language.

5.2.3.2.3 Morpho-syntactic

The database may also be labelled at the morpho-syntactic level. This is of particular interest to those wishing to study the relationship between prosody and syntax. The researcher will label such items as clause and phrase boundaries, compound word internal boundaries, and possibly affix boundaries also.

5.2.3.2.4 Citation-phonemic

The citation-phonemic level (also referred to as the 'phonemic' level by Barry & Fourcin (1992) may contain the output phoneme string derived from the orthographic form (by lexical access, by letter-to-sound rules, or both). There are various possibilities for representing the phoneme symbols. Some platforms have the facility to display the full range of IPA symbols, such as the symbols used by the latex font wsuipa11 (see table below).

[insert latex source for original Table 2.1 here, and re-number it]

However, many researchers will need to use some other means, such as an alphabetic or numeric representation of IPA symbols. The numeric equivalents of all IPA symbols are displayed in Esling & Gaylord (1993). An alphabetic equivalent system is used on the newsgroup sci.lang (see Appendix A), comprising all IPA symbols.

If the requirement is narrowed to symbols only for the main European languages, then the SAMPA system (see Appendix A) will be sufficient. This system has the advantage that, for any given language, only one grapheme is used per phoneme, with no spaces in between. Other systems that have been proposed (principally for English) allow a string of two or more graphemes to represent a phoneme, but a space between each phoneme representation is then necessary.

In the case of English, there are still more alphabetic systems that have been used in the past, such as (for American English) ARPABET and KLATTBET, and (for British English RP) Edinburgh's Machine-Readable Phonetic Alphabet, all of which use short grapheme strings separated by a blank space. However, a language-specific set of alphabetic phoneme symbols has not yet been devised for all possible languages.

Recommendation 2

If the corpus is confined to one language, and if the labelling is to be alphabetic rather than true IPA symbols, then it is advisable to use a language-specific set of characters. This avoids the notational complexity necessary when all symbols must be kept distinct across all languages, as is needed in the study of general phonetics.

5.2.3.2.5 Broad phonetic (or phonotypical)

The citation-phonemic labelling will not show such phenomena of running speech as place assimilation, consonant deletion, French liaison, RP linking-r and vowel reduction, etc. Hence there is a need for this more detailed level, which may (at least initially) be generated entirely by phonological rules from the citation-phonemic representation. It uses only symbols that have the status of phonemes, marking the output of connected speech processes that either insert or delete phonemes, or transform one phoneme into another.

It would be possible to segment and label a small sample of speech manually at this level of labelling, and then use this speech to train HMM models. The models would then be used to segment the main part of the database automatically. The resulting segmentations could be used as input to a process of manual post-editing, which would also be at the broad phonetic level, and ought not to require too much manual intervention. Alternatively, if the database is very large, there would be no manual intervention (in which case a certain loss of accuracy would have to be accepted).

This is the level referred to as 'phonotypical' for French labelling by Autesserre, Pérennou & Rossi (1989). It may be derived initially from the citation-phonemic labelling by phonological rules which, for example, reduce or delete unstressed vowels in English, or which delete some cases of word-final schwa in French. At this stage, it does not require reference to the sounds actually produced by the speaker (though this may come later, in manual post-editing). The 'broad phonetic' level of Barry & Fourcin (1992) is intended as a phonemic-level representation of the speaker's tokens, ie. the labeller makes reference to the speech signal. The 'phonotypical' level of Autesserre, Pérennou & Rossi (1989) is intended to be derived purely by rule. Both approaches are possible: however, in the context of speech recogniser training, there are certain advantages in deriving this level of labelling purely by rule, as this is relatively quick and easy. This level of labelling uses a limited inventory of symbols (and so it is a practical proposition for very large databases), while also offering more phonetic detail than the citation forms. So there is a balance between accuracy of representation and ease of derivation of the representation. However, there will still be some discrepancies between completely rule-derived labelling (where the corpus is automatically segmented) and the speaker's utterance. Hence there is some loss of accuracy, which may possibly be offset by a large volume of data.

5.2.3.2.6 Narrow phonetic

The narrow phonetic level of labelling is the first one where the labeller cannot avoid listening to the recording and/or inspecting the waveform and spectrogram. This is because it attempts to represent what the speaker actually said at the time of recording. This consideration immediately increases the time and effort necessary, as every part of the speech must be inspected manually.

The inventory of symbols is increased to include sounds that do not have phonemic status in the language (such as a glottal stop for English, or a symbol for aspiration after a voiceless plosive). It is at this level that different allophones may be represented, as well as devoicing or voicing, and secondary articulations such as nasalisation or labialisation. One segment at this level (eg. a voiceless plosive) may correspond to more than one segment on the acoustic-phonetic level (eg. the closure phase and burst of a plosive). There are potential problems in determining the boundaries of segments at this level, especially in the case of 'desequencing' or transposition of two sounds (Barry and Fourcin, op. cit.). However, this level of representation will be far more accurate as a record of what was said, and hence will yield more reliable HMM's when training a speech recogniser. This implies that not so much data will be needed, provided it is accurately transcribed at this level.

Recommendation 3

It is better not to embark without good reason on a level of labelling that requires the researcher to inspect the speech itself, as this greatly increases the resources needed (in terms of time and effort). If the broad phonetic (ie. phonotypical) level is considered sufficient, then labelling at the narrow phonetic level should not be undertaken.

5.2.3.2.7 Acoustic-phonetic

This level of labelling distinguishes every portion of speech that is recognisably a separate segment of the acoustic waveform or spectrogram. Hence stop closures, release bursts, and aspiration phases will all be segmented separately, as will glottal onset of a vowel, or separate voiced and voiceless portions of a fricative, nasal or stop. The labelling will be done in terms of these recognised articulatory categories, which can be easily related to the labels at higher linguistic levels. The boundaries of some segment types (glides and vowels) will to some extent be arbitrary, and a criterion for homogeneity such as "sameness of change" will have to be employed (eg. in the falling second formant for a palatal glide).

5.2.3.2.8 Physical

The most detailed level of labelling is the physical level. This does not need to relate only to an acoustic record, but could have separate tiers related to different types of input (eg. nasal transmission detectors, palatography). However, acoustic parameters are likely to be the most frequent type of representation, and different types may be needed for particular application areas (eg. filter bank output energies, formant frequencies, LPC and cepstral coefficients, fundamental frequency or electroglottograph output waveform). The physical events may be overlapping or discrete in time, with each parameter allotted a separate annotation row (eg. nasal resonance, periodicity, high-frequency noise). This level of labelling has not been generally used in speech technology research to date, but it has the potential to serve as a resource for developing speech synthesisers with greater 'naturalness', and speech recognisers which include more speech knowledge in their algorithms.

5.2.3.2.9 Non-linguistic phenomena

Another level of labelling may be used for non-speech phenomena. This includes speaker noises such as coughing, laughter, and lip smacking, as well as extraneous noises such as the barking of dogs and the slamming of doors. In addition, this level could also be used to label paralinguistic information such as disfluencies and filled pauses. The type of representation used for such annotations will depend on the purpose of the database. An annotation system such as that proposed by the Text Encoding Initiative is very elaborate and makes heavy demands on a transcriber, but also makes it possible to derive all relevant information from a transcription. While the TEI system makes use of SGML, which guarantees that existing software can be used, there is a large initial learning curve for the transcriber, which multiplies the possibility of human error in the transcription. Other annotation systems (such as those used in ATIS and Switchboard) are less elaborate, but also easier for transcribers to learn. The conventions used in ATIS, Switchboard, Polyphone and the Groningen corpus consist of different types of brackets with possible additional glosses. Retrieval software referring to these particular annotations must be designed in a more or less ad hoc way, which is less convenient than the TEI system. However, it is possible to provide standard scripts for a speech corpus. It is important to find the correct balance between the sophistication of the annotation system and the practicality of the system from the transcriber's point of view.

The following types of phenomena could conceivably be annotated on this level of representation:

Omissions (in read text) : Words from the reocording script which were omitted by the speaker may be indicated. In spontaneous speech, it is very difficult to know whether a speaker has omitted words which he or she actually intended to say, and so omission is only relevant in the case of read speech.

Verbal deletions and corrections (implicit or explicit) : Words that are verbally deleted by the speaker may be indicated. These are words that are first uttered, then superseded by subsequent speech. This can be done explicitly, as in "Can you give me some information about the price, I mean the place where I can find it". Alternatively, it can be done implicitly, as in "Can you give me some information about the price, place where I can find it". Verbal deletions may be indicated in both read and spontaneous speech.

Word fragments : Word fragments comprise one or more sounds belonging to one word. For example, in ATIS word fragments are indicated by a hyphen, as in "Please show fli- flights from Dallas".

Unintelligible words : Sometimes only part of a word is unintelligible, in which case only the intelligible part is transcribed orthographically. If a word is completely unintelligible, this fact will be annotated on the level of 'non-linguistic phenomena'.

Hesitations and filled pauses : Filled pauses (such as 'uh' and 'mm') may be indicated. Some annotation conventions (eg. Polyphone and Switchboard) annotate only one or two types of filled pause ('uh' and 'mm', or purely 'uh'). Other systems (eg. ATIS and Speech Styles) annotate more than two types (eg. 'uh', 'mm', 'um', 'er', 'ah'). The types of filled pause vary across languages (for example, the English 'er' is not used in Dutch). The recommendation is to use at least two types: one vowel-like type 'uh', and one nasal type 'mm'.

Non-speech acoustic events : These can be made either by the speaker or by outside sources. The first category includes lip smacks, grunts, laughter, heavy breathing and coughing. The second category includes doors slamming, phones ringing, dogs barking, and all kinds of noises from other speakers. The Switchboard corpus uses a very extensive list of non-speech acoustic events, ranging from bird squawk to groaning and yawning. The recommendation is that these events are annotated at the correct location in the utterance, by first transcribing the words and then indicating which words are simultaneous with the acoustic events.

Simultaneous speech : For dialogues and interviews, words spoken simultaneously by two or more speakers may be indicated.

Speaking turns : Discourse analysis makes use of indications of different speaking turns and initiatives. While these are not generally used in speech technology, it would always be possible to transcribe them.

5.2.3.2.10 General recommendations for transcribing

The following are some general recommendations of a fairly practical kind.

Utterance information : The recommendation is that transcribers annotate information about the process of transcribing, as some indication of the difficulty of transcribing may be useful in later analysis. The transcribers of the Switchboard telephone corpus annotated (on a scale of one to five) the following characteristics of a conversation: difficulty, topicality, naturalness, amount of echo from speaker B/speaker A, static on A/B, background noise on A/B.

Labelling procedure : Where labels are annotated at more than one linguistic level, the recommendation is to listen to only one level at a time. This is because, in ordinary conversational listening, hearers mentally delete all hesitations, false starts etc. and do not analyse intonation explicitly. However, transcribers must learn to hear all these events in an analytical way.

The easiest method is to listen to and label the words first, and then to label the prosodic phenomena and other annotations.

Transcriber features : In the case of orthographic transcriptions, it is not necessary to use experienced transcribers. However, in the case of phonemic and phonetic transcriptions, the recommendation is that the transcribers should be trained phoneticians, or otherwise accustomed to listening to speech in a very analytical way.

Transcribing time : An orthographic transcription of spontaneous speech will require about ten times real time (ie. ten times the duration of the speech in question). An orthographic transcription of read sentences will require about three times real time, while an orthographic transcription of read continuos texts will require about five times real time. A phonemic or phonetic transcription will require very many more times real time: the precise factor cannot be determined as so many variables are involved.

Checking of transcriptions : It is always necessary to check transcriptions, and this can be done in different ways. An independent transcriber can transcribe the whole or a sample portion of the corpus. Alternatively, a non-transcriber can scan the transcription while listening to the speech, which is less time-consuming. In the latter case, the recommendation is to perform the checking procedure over the database in the reverse order to that in which it was transcribed. This is because the transcriber will have become more consistent in his/her use of the transcription system by the end of the database, and so the final part of the database will train the checker as he/she begins to work with the particular system used.

5.2.3.3 Manual segmentation

Manual segmentation refers to the process whereby an expert transcriber segments and labels a speech file by hand, referring only to the spectrogram and/or waveform. There is no automatic assistance in segmenting. The manual method is believed to be more accurate. Also, the use of a human transcriber ensures that the segment boundaries and labels (at least at the narrow phonetic level) are perceptually valid. However, there is a need for explicit segmentation criteria to ensure both inter- and intra-transcriber consistency, together with (ideally) some form of checking procedure. Sets of guidelines for manual segmentation have been developed by various projects. One such is Hieronymus et al. (1990), which uses the four levels of underlying phonemic, broad phonetic, narrow phonetic and acoustic. It also retains the same base phonemic symbol even at the acoustic level, to facilitate the automatic determination of boundaries at the phonetic level once the boundaries at the acoustic level have been determined. Much speech data (particularly in English) has been segmented and labelled entirely manually,

One possible measure of accuracy for segmentation and labelling is consistency between transcribers. Barry & Fourcin (1992) quote Cosi & Omologo (1991) as saying that one should not expect more than 90Eisen (1993) investigates inter-transcriber consistency for the separate tasks of segmentation and labelling at three different levels of labelling, and concludes that consistency depends partly on the degree of abstraction of the labelling level, and partly on the type of sound involved. The best results in labelling were achieved at the broad phonetic level, for fricatives, nasals and laterals, which showed greater than 90The best results for segmentation were achieved at the acoustic-phonetic level, for the acoustic features ‘fricative’, ‘voiced’ and ‘vowel-like’.

Recommendation 4

Any accuracy measure based on inter-transcriber consistency must control for the factors ‘level of transcription’, ‘segment type’, and ‘task type’ (whether segmentation or labelling).

5.2.3.4 Automatic and semi-automatic segmentation

Automatic segmentation refers to the process whereby segment boundaries are assigned automatically by a program. This will probably be an HMM-based speech recogniser that has been given the correct symbol string as input. The output boundaries may not be entirely accurate, especially if the training data was sparse. Semi-automatic segmentation refers to the process whereby this automatic segmentation is followed by manual checking and editing of the segment boundaries.

This form of segmenting is motivated by the need to segment very large databases for the purpose of training ever more comprehensive recognisers. Manual segmentation is extremely costly in time and effort, and automatic segmentation, if sufficiently accurate, could provide a short cut. However, it is still necessary for the researcher to derive the correct symbol string to input to the autosegmenter. This may be derived automatically from an orthographic transcription, in which case it will not always correspond to the particular utterance unless manually checked and edited. The amount of inaccuracy that is acceptable will depend on the uses to which the database is to be put, and its overall size.

There will always be a need to verify the accuracy of an autosegmented database, and the obvious accuracy measure is the consistency between manual and automatic segmentation over a given subset of the database. Schmidt & Watson (1991) carried out this evaluation over nearly 6000 phoneme-sized segments, and found that the discrepancy between manual and automatic boundaries varied across segment types. The absolute mean discrepancy was greatest for diphthongs (5.4 ms) and least for nasals (0.37 ms). For 50all segmentations, the discrepancy was less than 12 ms, while for 95less than 40 ms. This falls within the range of just-noticeable differences in duration for sounds of the durational order of speech sounds (Lehiste (1970: 13)) and so one could conclude that the discrepancies are not perceptually relevant.

This means that automatic segmentation for the given data, using the given autosegmenter, was probably sufficiently accurate.

5.2.3.5 Segmentation and labelling in the VERBMOBIL project

In the German VERBMOBIL project, segmentation and labelling of recorded speech data is a fundamental part of the research. The following procedures are adopted. The phonemic labels are based on the SAMPA symbols for German, augmented by extra labels for phonetic segments such as plosive release, aspiration after a plosive, creak, and nasalisation of a vowel as a reflex of a deleted nasal. Hence the labelling is carried out at the narrow phonetic level.

During the labelling process, a label will be aligned with the start of the portion of speech that is considered to represent its chief acoustic correlates. Labels are discrete and non-overlapping, except in the following cases:

1. Labels for creak and nasalisation are always superimposed on other labels, which they modify.
2. A special label (–MA) is used to indicate that the phonetic correlates of one or more deleted segments are present in the surrounding material. For example, where an unstressed rounded vowel has been deleted, labialisation may still be present in a neighbouring consonant, and will be marked in this way.

Inter-labeller consistency is maintained in three ways, as follows:

1. The inventory of possible labels is restricted mainly to the list of German phonemes. This restriction minimises the possibility of error.

2. The labeller works from a citation-phonemic form of the utterance that has been previously prepared. This eliminates gross errors.
3. There are restrictions on the types of modification allowed. The labeller is permitted to mark the following: deletion (where the initially-provided label is marked with a following hyphen, to indicate deletion); insertion (where the new label is prefixed with a hyphen); and substitution (where the new label is inserted after the one initially provided, separated by a hyphen).

The checking of segmented and labelled speech files is carried out partly by a program that detects invalid sequences of symbols, and partly by experienced labellers checking the work of less experienced transcribers. All segmenting and labelling is carried out manually. The initial citation-phonemic transcription is derived from the grapheme-to-phoneme converter of the Infovox TTS system (and then is checked for mistakes). A system of prosodic labelling is currently under development.

5.2.4 Prosodic labelling and annotation

5.2.4.1 Definition and motivation

5.2.4.1.1 Types of approach to prosodic labelling

The above discussion has been in terms of segmental labelling only. It is also possible to annotate a speech database at the prosodic (suprasegmental) level. This is less straightforward than segmental annotation, as there are far fewer clear acoustic cues to prosodic phenomena. The F0 curve will be the relevant acoustic display, possibly augmented by the intensity curve. The waveform is a useful guide to the current location in the speech and is usually displayed together with the F0 curve (as in the WAVES labelling software).

The units segmented will depend on the particular theoretical bias underlying the given research programme. A basic distinction may be drawn between a prosodic labelling system that annotates the boundaries of units (analogous to the method used in segmental annotation) and a system that annotates the occurrence of isolated prosodic events (such as F0 peaks) where the status of the units bounded by these annotations is of little interest.

The first type of method may possibly use the intonational categories proposed by Nespor & Vogel (1986), such as intonational phrase, phonological phrase, phonological word, foot, and syllable. Alternatively, it could mark the more traditional units of 'minor tone-unit' and 'major tone-unit', as in the MARSEC database (Roach et al. (1993)). Or again, it could annotate the perceptual phonetic categories used in the 'Dutch school' of intonation studies, such as rises and falls that are early, late or very late in their timing, fast or slow in their rate of change, and full or half sized ('t Hart, Collier & Cohen (1990)). This type of annotation could be used in conjunction with annotation at the morphosyntactic level to yield information about the relationship between the syntactic and prosodic levels in terms of duration, pauses, etc.

The second type of method, though it may refer to the units mentioned above in its underlying theory, does not in fact annotate them but rather marks the occurrence of high and low tones of various kinds. The recently formulated ToBI transcription system (Silverman et al. (1992)) is the most well-known system of this kind for English.

5.2.4.1.2 Examples of the two types of approach

Prosodic annotation has only recently come into favour in the field of speech and language technology research. Now that a basic level of competence has been achieved as regards the synthesis and recognition of speech segments, researchers have come to realise that much more work is required on the prosodic aspect of speech technology. This is the motivation for the growth in popularity of speech database research, and for the formulation of the ToBI prosodic transcription system. In order for the prosodic transcriptions of various different speech databases to be comparable, and in order to make the best use of existing resources, the originators of ToBI (Silverman et al., op. cit.) proposed a simple system that would be easy to learn and that would lead to good inter-transcriber consistency. To date it has largely been used for English, especially American English, but at least in principle it could be extended to other languages as well. The system has certain severe limitations (eg. it has no way of representing pitch range) but its minimalist formulation was dictated by the need for learnability and consistency in use. The 'British school' type of system used in the MARSEC database of British English (Roach et al. (1993)) contains more phonetic detail but may require more effort in teaching to novice transcribers. The 'IPO' classification of F0 patterns ('t Hart, Collier & Cohen (1990)) has not yet been used systematically in the annotation of large-scale publicly-available speech corpora, but has been used successfully in the development of speech synthesisers.

Prosodic transcription also has obvious uses in basic linguistic research, especially since research into the suprasegmental aspects of language is not nearly as advanced as research into the segmental aspects. As indicated above, a database annotated at the prosodic and morphosyntactic levels can provide information on the relationship between them with respect to duration and pauses. If the segmental level is also annotated, then many possibilities open up for the study of segmental duration in prosodic contexts. This is especially true in the case of languages other than English, where these aspects have received comparatively little attention to date.

The concept of levels of prosodic labelling applies differently to the two different approaches to prosodic labelling outlined above. In the first case, the obvious categories would be those proposed by Nespor and Vogel (op. cit.), comprising levels of non-overlapping units each of which corresponds to one or more units on the level immediately below (eg. phonological phrase, foot, syllable). In the second case, the separate levels have no such intrinsic relationship to one another, but merely deal with different types of phenomena.

For example, in the ToBI system, there are separate levels for tones and inter-word 'break indices'. The ToBI system can be described briefly in terms of its separate levels, and is described in what follows. The MARSEC system will be outlined after that. The 'Dutch school' system of IPO will not be described in much detail, as it has not (yet) been used for annotation of publicly-available speech corpora: however, extensive references are avilable in 't Hart et al. (op. cit.).

5.2.4.2 The ToBI labelling system

A recent experiment (Pitrelli, Beckman & Hirschnerg (1994)) used several prosodic transcribers working independently on the same speech data, comprising both read and spontaneous American English speech. The ToBI system was used, and a high level of consistency across transcribers was found, even given the fact that transcribers included both experts and newly-trained users of the system. This suggests that the system achieves its object of being easy to learn and to apply consistently, at least in the case of American English.

The 'orthographic' level of the ToBI system contains the orthographic words of the utterance (sometimes only partial words in the case of spontaneous speech). It is also possible to represent filled pauses (eg. 'um', 'er' at this level).

The 'miscellaneous' level may be used to mark the duration of such phenomena as silence, audible breaths, laughter and disfluencies. There is no exhaustive list of categories for this level, and different transcription projects may make their own decisions as to what to annotate.

The 'break index' level is used to mark break indices, which are numbers representing the strength of the boundary between two orthographic words. The number 0 represents no boundary (with phonetic evidence of cliticisation, eg. resyllabification of a consonant), and 4 represents a full intonation phrase boundary (usually 'end of sentence' in read speech), defined by the occurrence of a final boundary tone after the last phrase tone. The number 3 represents an intermediate phrase boundary, defined by the occurrence of a phrase tone after the last pitch accent, while the number 1 represents most phrase-medial word boundaries. The number 2 represents either a strong disjuncture with pause but no tonal discontinuity, or a disjuncture that is weaker than expected at a tonally-signalled full intonation or intermediate phrase boundary.

The 'tone' level is used to mark the occurrence of phonological tones at appropriate points in the F0 contour. The basic tones are 'L' or 'H' (for 'low' and 'high'), but these may function as pitch accents, phrase accents or boundary tones, depending on their location in the prosodic unit. In the case of pitch accents (which occur on accented syllables), there may be one or two tones, and the H tone may or may not be 'downstepped'.

Information about the ToBI system and guidelines for transcribing are available by anonymous ftp from kiwi.nmt.edu.

5.2.4.3 The MARSEC labelling system

The MARSEC project (Roach et al. (1993)) is based on the Spoken English Corpus (Knowles, Taylor & Williams (1995)), a corpus of British English that at the time was not time-aligned. The MARSEC project time-aligns the prosodic annotations, the orthographic words, the grammatical tag of each word, and individual segments. The type of prosodic annotation used is the 'tonetic stress mark' type of system. Several types of accent are recognised: low fall, high fall, low rise, high rise, low fall-rise, high fall-rise, low rise-fall, high rise-fall, low level, and high level. These may occur either on nuclear or on non-nuclear accented syllables. In addition, there is a distinction between major and minor tone-unit boundaries, and there is provision for 'markedly higher' or 'markedly lower' in perceived pitch. The tonetic stress mark type of system has been used for many years, and has been applied to many languages apart from English (the same is not true of the ToBI system). However, no extensive attempts have yet been made to apply it in the field of speech technology.

The Spoken English Corpus comprises over fifty thousand words of broadcast British English in various styles, mostly monologues. Two transcribers prosodically annotated it in an auditory fashion, with no access to the F0 curve. They each transcribed half the corpus, but each also independently transcribed certain passages known as 'overlap' passages, the purpose of which was to check on inter-transcriber consistency. Analysis of the overlap passages reveals that the consistency is fairly good, certainly in the case of major aspects such as location of accents and direction of pitch movement (Knowles & Alderson (1995)). This result is especially encouraging in view of the fact that the transcription system used contains far more phonetic detail than the ToBI system.

5.2.4.4 The IPO approach

The phonetically-based analysis of intonation used at IPO ('t Hart, Collier & Cohen (1990)) has the advantage of having proved its usefulness for more than one language, and of having been successfully applied in the field of speech synthesis (neither of these considerations apply to the ToBI system). The analysis proceeds by modelling F0 curves in terms of straight lines that have been experimentally proved to be perceptually indistinguishable from the original ('close-copy stylisations'). This type of representation is then further simplified into 'standardised stylisations' in terms of a small set of available contours for a given language. This type of representation has been experimentally proved to be distinguishable from the original on close listening, but yet not functionally any different from the original (ie. the standardised stylisation is linguistically equivalent).

In the case of Dutch, there are ten basic pitch movements (the model has also been applied to British English, German and Russian). These are the five falls and five rises, varying along the parameters of syllable position, rate of pitch change, and size of pitch excursion. These ten pitch movements are grouped into 'pitch configurations' (of one or two pitch movements each). The pitch configurations are classified into prefixes, roots and suffixes. Sequences of pitch configurations are grouped into valid 'pitch contours', which in turn are grouped into melodic families or 'intonation patterns' (of which there are six in Dutch). These groupings are experimentally verified by listeners. The units of this analysis, at all levels, are based on speech corpora of spontaneous and semi-spontaneous utterances in Dutch. In contrast to the ToBI and MARSEC systems, comparatively little effort has been put into checking inter-transcriber consistency, possibly because the detection and labelling of this kind of phonetic unit is less problematical.

5.2.4.5 Provisional recommendations

It is reasonable to assume nowadays that a prosodic transcriber will have access to at least the waveform and the F0 curve for the speech to be transcribed. In that case, the recommendation is to use either the ToBI or the IPO system (and the MARSEC system if a purely auditory transcription is being carried out). If the language to be transcribed is not English, and especially if the projected application of the prosodic transcription is in the field of speech technology, then it is probably best to use the IPO system if possible (i.e. if the basic 'grammar' of contours has already been researched for that language). However, these can only be provisional recommendations, as little work has been carried out on prosodic labelling in comparison with the great effort that has been expended on segmental labelling. In this situation, it may be that a different system entirely will prove more appropriate to the given language, and it is not possible to make absolute recommendations.

5.2.5 Annotation conventions

Next to the symbolic representation of words (lexical transcriptions), annotations can be used to indicate possible events that occur during speaking. In the following, these events are mentioned. How to represent these events depends on the system one wants to use. A system like the one proposed by TEI is very elaborate, and therefore it demands quite a lot of a transcriber. A very advanced system supplies the means to retrieve all possible information from the transcription. TEI, for example, makes use of SGML which guarantees that existing software can be used. A transcriber, however, has to learn a lot and this will increase the chance that mistakes are made. Other systems, like the ones used in ATIS and Switchboard, are perhaps less advanced, but they are relatively easy to learn by transcribers. One must keep in mind that listening to *all* events that occur during speaking already is a difficult task, and a great number of conventions makes transcribing even more difficult. The conventions used in ATIS, Switchboard, POLYPHONE and the Groningen Corpus consist of different types of brackets with or without additional information. Retrieval software referring to these specific notations must be designed in a more or less ad hoc way. This is perhaps less convenient in comparison with the TEI system. However, it is possible to provide a speech corpus with standard scripts. One must seek a balance between an advanced annotation system and convenience for transcribers.

Notation conventions used in discourse analysis, like speaking turns and initiatives are not mentioned here. As far as we can see now, conventions can always be added to indicate this type of information.

In the following, some occurrences that must be annotated are discussed.

5.2.5.1 Deletions (read text)

Words of the written texts which are not said by the speaker must be indicated. For spontaneous speech it is very difficult to know whether a speaker deletes words which he actually wanted to say. Therefore, this type of deletion is usually not indicated when spontaneous speech is concerned.

5.2.5.2 Verbal deletions or corrections (implicitly or explicitly)

Words that are verbally deleted by the speaker must be indicated. Verbal deletions are words that are actually uttered by the speaker, but which, according to the transcriber, are superseded by subsequent speech. This can be explicitly done, e.g. *Can you give me some information about the price, I mean, the place where I can find...*, or it can be implicitly done, e.g. *Can you give me some information about the price, place where I can find...*. Verbal deletions can be indicated in read as well as spontaneous speech.

5.2.5.3 Word fragments

Word fragments must also be indicated. Word fragments may be single sounds or larger parts of a word. In ATIS, word fragments are for example indicated by means of a hyphen (*please show fli- flights from dallas*).

5.2.5.4 Unintelligible words

Words which cannot be understood must be indicated. Sometimes only part of a word is not understood. In this case the part that is understood can be written down. If a word is completely unintelligible this must be represented, e.g. by putting two round brackets before and after a space *(( ))*.

5.2.5.5 Hesitations (filled pauses)

Filled pauses like *uh* and *mm* must be indicated. Some annotation conventions (e.g. POLYPHONE and Switchboard) choose to indicate only one or two types of filled pauses (*uh*, or *uh* and *mm*). Other systems (e.g. ATIS and Speech Styles) choose for more types of these sounds (*uh, um, er, ah, mm*). The types of filled pauses are not the same for all languages. For example, the English *er* is not used in Dutch. It is recommended to use at least two types, one vowel type *uh*, and one nasal type *mm*.

5.2.5.6 Non-speech acoustic events

Here the distinction can be made between non-speech acoustic events produced by the speaker and non-speech acoustic events produced in another way. Smacks, grunts, laughing, loud breath, coughing, all produced by the speaker belong to the first category. Noises of doors, phone ringing, dog barking, and all kinds of noise produced by other speakers belong to the second category. The Switchboard corpus uses a very extensive, precise list of non-speech acoustic events, ranging from bird squawk to groaning, and from dishes to yawning. It is recommended that non-speech acoustic events are marked at the correct location in a transcribed utterance. This must be done by first transcribing the words and then indicating during which words the acoustic events occur.

5.2.6 Recommendations

For the transcription of dialogues between more than two speakers use a "music score notation".

For orthographic transcriptions, use the standard spelling as much as possible.

Indicate reduced word forms in orthographic transcriptions a) if these forms occur frequently and b) if they involve syllable deletion.

It is recommended to include a list of all non-standard spellings used in the orthographic transcription in the documentation of a speech corpus.

When orthographic transcription is used in a corpus, it is recommended that a list of unique words and word forms is generated on the basis of the transcription. Then, the graphemic forms of the words can be converted to phonemes by means of computerised grapheme-to-phoneme conversion. The result of this is a list of citation forms, also called canonical forms. These forms indicate the pronunciation of words when spoken in isolation. Of course, the great variability in pronunciation cannot be indicated, but this procedure at least gives the standard pronunciation. This is especially relevant if a corpus will be used by other persons than those belonging to the own language community. On the basis of the canonical forms phonetic transcriptions can be made semi-automatically using large vocabulary speech recognisers. At least the attempts made so far are promising.

If there are no clear reasons, do not start to transcribe a corpus phonetically, since the time spent on this will never be paid back. If very specific phonetic details are needed, one is recommended to look for these on the basis of orthographic and/or phonemic transcriptions.

It is recommended that transcribers give information about the process of transcribing and about the speech they transcribed. Some speakers will be easier to transcribe than the speech of other speakers. This will depend on the speech rate, the neatness of articulation and the amount of hesitations, and the amount of dialect words used by the speakers. Some information about the difficulty of the transcription is very useful for later queries. The transcribers of the Switchboard (telephone) Corpus had to indicate on a scale ranging from 1 to 5 the following characteristics of a conversation: difficulty, topicality, naturalness, echo from B (in listening to A separately, B could hardly be heard (1) or was as nearly as loud as A (5)), echo from A, static on A (no static noise (1) or great deal of it (5)), static on B, background A, and background B.

In case of transcriptions at more than one level (e.g. orthographic transcription with some prosodic marks and indications of hesitations, etc.), it is recommended to listen to one level at a time. In everyday life, listeners are used to skip hesitations, false starts, and other imperfections and they do not pay explicit attention to prosody. Transcribers must learn to hear all these events. It seems easiest to listen to the words first and transcribe these, and then to place the prosodic marks and other annotations.

For orthographic transcriptions it is not necessary to find experienced transcribers. However, for phonemic and phonetic transcriptions it is recommended to look for transcribers who are used to listen to speech in a very precise way.

To give some indication about the time necessary to transcribe speech we give some examples. The time that will be necessary to make an orthographic transcription of spontaneous speech is about ten times the duration of the speech itself. The time necessary for an orthographic transcription of read sentences is about three times the duration of the speech and for an orthographic transcription of read texts it is about five times the duration of the speech.

Checking of transcription is always necessary. Checking can be done in different ways. An independent transcriber can transcribe the whole or a sample of the corpus. Another possibility is to let someone check the transcription by reading the transcription and listening to the speech. This is less time consuming. For the last procedure, it is recommended to check the transcription in the reversed order as the first transcriber transcribed the speech. At the end the first transcriber will be more consistent than in the beginning. Differences can occur in the conventions used (spelling and annotation conventions (brackets, etc.)), as well as in what is heard by the two different persons.

5.3 Storage and design of the data base

5.3.1 Data types for a speech data base

The previous chapter and section described the data collection and the representation of the speech. In this part information is given about how speech, speaker information, information about data collection, and transcriptions may be stored in a data base, that is a *structured* speech corpus. The design of the data base again very much depends on its future use.

Many additional sources of information can be gathered apart from the basic acoustic speech signal. Whatever choices of speakers, speech material, and recording conditions are made, it is always of crucial importance that the collecting procedure is documented as elaborately as possible. It is good practice to record all possible details about, for instance, sex and age of speakers, type of speech material (isolated words, sentences, discourse, etc.), place of recording (in a laboratory, on location, etc.), type of microphone and recording medium. Although you may not be interested in specific information at the time, it can turn out to be important at a later stage. In that case it is often difficult or impossible to recollect the information you need. And in the second place, a well documented speech corpus may also be used for other research. The following list summarises the most common information sources that may be useful for a speech corpus:

- Transduced signals
  Examples: The acoustic speech signal, laryngograph signal, X-ray data.
- Analysis results
  Examples: FFT data, LPC data, filter bank data, pitch extraction, formant extraction.
- Descriptors
  Examples: Characteristics of the speakers, or the recording conditions.
- Markers
  Examples: Markers to indicate pitch periods, or the beginning of vowels.
- Annotations/Labels
  Examples: Orthographic, phonemic, or phonetic transcriptions.
- Assessment parameters
  Examples: Test material, assessment results.

All these information sources must be stored in such a way that potential users of the speech corpus can get access to the speech and the speech-related data in an efficient and easy-to-use manner.

5.3.2 Storage demands

The sampled speech signals form the core of the speech corpus. Since these data do not need to be changed, they can conveniently be stored on static media, such as optical disks. Fixed speech-related information, such as speaker descriptors, etc., can also be stored on a static medium. However, speech-related information sources that are subject to change, such as transcriptions, should preferably be stored on, for instance, a high capacity computer disk. Analyses of the sampled speech data, such as FFT, formant extraction, etc., could be done 'on line', provided that special signal processing software would be available. In this way permanent storage of analysis results could be avoided, leading to considerable savings of storage.

5.3.2.1 Storage medium

Nowadays, CD-ROMs are most frequently used for the distribution of speech corpora. A CD-ROM can contain 640 Mbyte. The amount of speech it actually contains depends of course on the used sampling rate and compression techniques. The file system needed for a CD-ROM depends on the computer system that one wants to use. This makes the use of CD-ROMs slightly problematic, since it is impossible to standardise the file system on it. For example, if one plans to use a CD-ROM on a Personal Computer working under DOS, one must take into account that the file names have the appropriate eight.three format. On-line access to speech corpora (often used for text corpora) is not propagated, since this is still very time consuming at this moment. DAT tapes can be used to store the corpus, but these tapes are not safe enough. One can also use exabyte tapes, but this medium is not very accessible.
In table 5.3.2.1 we present an overview of storage media, together with their properties:

In the near future, workstations are expected to be equipped with the same number of devices as today (i.e. still one hard disk, one CD), but of larger size or higher speed. In general, an improvement of 2 to 8-fold can be expected. Optical media, incl. CD-ROM will profit from shorter wave-length lasers currently being developed in research labs, but it will take until '98 for higher capacity devices to reach the market. Not much will change for tapes.

5.3.2.2 Sampling rates

If the original recordings have been made on DAT (which is strongly recommended), and the purpose of the corpus does not demand that the speech is stored with DAT quality at a sampling frequency of 48 kHz, sampling frequency can easily be reduced to 16 kHz. Rate conversion to 20 kHz, which seems to be the de-facto UK/EC standard for distribution on CD-ROM, can also be done without difficulty.

5.3.2.3 Compression

Sampled data files can be compressed and by doing this one can win on average 50% of the storage space. The amount of possible compression depends on the signal-to-noise ratio in the recordings. A high signal-to-noise ratio results in a better compression. Compression of telephone speech (A-law or U-law) results in a compression of about 40%, as was experienced during the storage of the Dutch POLYPHONE material. A compression tool that is often used and is recommended here is *Shorten* by T. Robinson. It is available through anonymous ftp (svr-ftp.eng.cam.ac.uk)

5.3.3 How to combine speech and speech-related data

The simplest way to combine speech and speech-related data is to code information in the name of the sampled data file. For example, the filename "w3_s7.sdf" could be used for the speech recording

Table 5.2: Storage media classification:

	Memory	Disk				Tape		
	RAM	Floppy	Hard	Magneto Optical	Compact	QIC	ExaByte	DAT
volatility	vol.	non volatile						
life time	-	2y[a]	5y	20y	40y	2y	2y	
technology	CMOS	magnetic		mag/opt	optical	magnetic		
access order	random					sequential		
mode	read/write many				read (write once)	read/write many		
seek time	10-9s	10-1s	10-3s	10-3s	10-2s	t/2[b]		
data rate	10+9b[c]	10+2b	10+7b		10+6b	10+6b		
cost[d] device				$500	$200	$600		
medium	$50/M	$1/M	$.5/M	$1/M	$.02/M	$1/M	$0.02/M	
workstation config[e]	16M	1.4M	500M	500M	600M	250M	2000M	
#		2	1	ext[f]	1	external		
future (97)	64M	1.4M	4000M		1200M			
trends		2	1		1			

[a] y is years
[b] t is rewind time for the entire tape
[c] b is bits/s
[d] cost estimates in US$ per device or Megabyte
[e] "workstation config" gives a standard configuration of a high-end PC workstation (Pentium) or middle range UNIX (Sparc) workstation
[f] ext is external devices

of the third word (w3) in a word list, uttered by the seventh speaker (s7). The extension ".sdf" indicates that the file contains sampled data. It will be clear that the possibilities to code information in a filename are very limited. Much more space for additional speech-related information is available in the header of a sampled data file. Several data about, for instance, the speaker (sex, age, weight, height, etc.), or about microphone type, recording medium, etc. can be stored in the file header. Frequently used headers are NIST Sphere headers and SAM headers. Examples of these are presented in Appendix B. The choice for a specific header may depend on future use of the speech corpus. For example, if one plans to use the speech analysis program X-waves, the ESPS header used in this program, can be chosen.

When the amount of speech-related information becomes very large, it is recommended to use a separate file (or perhaps even more than one) for the storage of these data. The drawback of this method is that the speech-related information is no longer physically connected to the speech data it refers to. But with a proper administration in the database structure, this should pose no problems. An example of a data description file, containing all kinds of additional information that could be gathered along with the speech recordings, can be found in Millar et al. (1994).

5.3.4 Database Management Systems

A *Database Management System (DBMS)* is a software system with the following properties:

- Data definition and manipulation in a formal data model,
- Safe storage of large amounts of data,
- Application independent storage of data,
- Controlled access to the data.

In the past, many spoken language (*SL*) corpora have been stored in the file management system provided by the operating system. Data is accessed via shell-scripts, custom programs, and application packages, e.g. spread-sheets, statistics programs, etc.

However, there are many disadvantages to this approach:

- Data is organised in an ad-hoc way,
- Data structures reflect the underlying, physical data structures and the file system organisation
- No automatic protection against data corruption,
- No access control to the data.

Clearly, this approach should not be followed.

5.3.5 Data model

A *data model* is a formal description of *entities* and their *relationships*, and of *operations* allowed on the entities. A data model is independent of a specific DBMS implementation. An *instance* of a data model or *database schema* is a data definition based on the data types and the data manipulation commands provided by the DBMS.

Data modelling is the process of describing the world of interest in the terms of a data model.

The following data models can be discerned:

- Hierarchical,
- Network,
- Relational,
- Object-oriented,
- Deductive.

The development of the data models can be characterised as a) continued abstraction from physical storage, and b) increase in expressive power.

5.3.5.1 Hierarchical data model

In the hierarchical data model, entities (for instance, speakers, recordings, or types of speech) are considered as *record types*. Record types are subdivided into *fields*. A *record* stores the information of one particular entity.

Record types are organised in hierarchical *tree* structures. Except for the topmost record type (the *root*), each record type in the tree, a *node*, has exactly one predecessor, and zero or more successor record types. A node with zero successors is a *leaf*. A record belongs to exactly one record type, and it may be linked to zero or more records in the successor record types.

The hierarchical data model thus considers all relationships between entities as 1:N relationships. Access to record types is possible by *navigation* through the tree, starting at the root; records are selected from the set of records of a record type.
Summarising, the hierarchical data model has the following properties:

- Hierarchical organisation of entities in 1:N relationships.
- Access through navigation starting at the root record type.
- Obsolete data model.
- Implementations: IMS.

5.3.5.2 Network data model

In the network data model record types are organised in a *directed graph* structure. A record type may have zero or more predecessor and successor record types. As in the hierarchical data model, the information on an entity is held in records. Unlike fields in the hierarchical data model, the fields of a record in the network data model may have multiple values.

A set type describes a 1:N relationship between two record types. Set types have names to distinguish the relationships of a record type. Since a record type may be linked to any other record type, complex relationships, e.g. N:M relationships, may be represented .

Access to record types is possible via special "entry point" record types from which dependent record types can be reached through navigation.

Summarising, the network data model has the following properties:

- Directed graph organisation of entities with named 1:N relationships between any two record types.

- Access through navigation starting at special "entry point" record types.
- Obsolete data model.
- Implementations: xxx.

5.3.5.3 Relational data model

The relational model is a data model with a thorough mathematical foundation. An *attribute* is a named set of atomic values, the *domain*. A *relation table* is a subset of the cross-product of the relation attributes.

The relational data model separates the logical data definition from physical storage.

Two data manipulation languages have been developed for the relational data model:

- *Relational algebra* is a procedural language which consists of the set operators 'union', 'intersection', 'difference' and the relational operators 'selection', 'projection', and 'join'. The selection operator selects rows from a relation table that meet a selection condition. The projection operator selects specific attributes from a relation table. The join operator merges relation tables according to a comparison condition over attributes from the original relation tables.

  The result of a relational operation is again a relation, so that the operators can be nested.

- *Relational calculus* is a declarative language based on first-order predicate logic.

  An *expression* is a term of the form $\{x_1, ..., x_n | COND(x_1, ..., x_n, x_{n+1}, ..., x_{n+m})\}$ with x_i variables over a domain and $COND$ a formula that is either true or false. The variable bindings on the left side of the vertical bar are the values returned as the result of a query, and $COND$ consists of atomic formulas which are connected through logical AND and OR operators.

The relational data model makes no assumptions about the relationships between entities. In general, an entity is mapped to a relation table, with the properties of an entity described through the relation attributes. Meaningful relationships between entities are expressed through relational algebra operations or calculus expressions. Hence, complex relationships between entities can be expressed, with the notable expression of recursive relationships.

SQL

Relational algebra and calculus are formal data manipulation languages. As such they lack data definition constructs, and because of their formal nature, they are not well suited for interactive data manipulation. The language SQL (*Structured Query Language*) is a relational database language which covers both data definition and manipulation and is close to English. It has become the de-facto language standard for relational DBMSs, and has been formally standardised by the ISO.

5.3.5.3.1 Example

The following example is a sample speaker and recording database schema implemented in SQL. First, relation tables are created using the data definition command CREATE with the appropriate arguments:

```
CREATE TABLE SPEAKER (
   ID, CHAR(8), PRIMARY KEY
   NAME, CHAR(20)
   FNAME, CHAR(20)
   SEX, CHAR(1)
   DBIRTH, DATE)

CREATE TABLE RECORDING (
   ID, DECIMAL, PRIMARY KEY
   RECDATE, DATE
   MEDIUM, CHAR(8)
   LOCATION, CHAR(20)
   SPK, CHAR(8)
   FOREIGN KEY SPK REFERENCES SPEAKER)
```

The data types available in SQL are restricted to very simple character or number types of fixed length. Bit-stream data and complex data structures are not supported.

Then queries are formulated using the data manipulation language of SQL:

```
SELECT S.ID, S.NAME, S.FNAME, R.RECDATE
FROM SPEAKER S, RECORDING R
WHERE S.DBIRTH > 12/27/60 AND R.SPK = S.ID
```

Besides being a database language for interactive database access, SQL has become, through its standardisation, increasingly popular as a programming language interface. External applications generate SQL code, which is then transmitted to the DBMS and evaluated there. The result relation is returned to the calling application for further processing (either as a whole or one tuple after the other with a cursor mechanism).

Most commercial relational DBMSs support SQL, but add their own extensions to overcome the limitations of the standard data types, e.g. with binary large objects /it BLOBs for image, audio or video data, or complex data structures for graphics objects.

5.3.5.3.2 Summary

- The relational model is a mathematical model.
- Logical data types are completely independent of physical storage.
- Data is manipulated using procedural relational algebra or declarative relational calculus.
- Relational DBMSs are now standard DBMS technology.
- SQL is the de-facto standard relational database language, but each DBMS implementation extends the standard in a proprietary way.
- Implementations: DB2, Oracle, SyBase, Ingres, and many others; available for all platforms (Mainframe, UNIX, PCs, Mac).

5.3.5.4 Object-oriented data model

The object-oriented data model aims at bridging the semantic gap between relation tables and entities of the real world through *objects* that directly correspond to entities. An object has a unique and immutable *object identifier*, and it belongs to a *class*. Classes are object definitions; they comprise attributes and operations over the class or the attributes. Both attributes and operations may be private to a class, i.e. visible only to objects belonging to that class (*encapsulation*), or public, i.e. visible to other classes. Classes are organised in class hierarchies and may inherit attributes and operations from superclasses. Operations are invoked by sending an object messages (*message passing*) which are executed if the required operation is defined for the object, otherwise they are passed on to the superclass.

In general, object-oriented DBMSs are integrated into an object-oriented programming language, giving them the full expressive power of the programming language and persistent object storage. Object-oriented DBMSs have just entered the marketplace, and they are successful in *non-standard* applications which require the full expressive power of programming languages, or complex and highly diverse data structures.

5.3.5.4.1 Summary

- The object-oriented data model is based upon the notion of object.
- An object belongs to a class, and classes are organised in a hierarchy with inheritance of properties from super- to subclasses.
- Object-oriented DBMSs now enter the marketplace.
- There is not yet a standard object-oriented database language.
- Implementations: O_2, ObjectStore, POSTGRES, Starburst, GemStone and many others.

5.3.5.5 Deductive data model

The deductive data model is a restricted first-order predicate logic extension of the relational data model. In the deductive data model, relations are either defined *extensionally* through *facts*, or *intensionally* through *rules*. Rules may be defined recursively.

Various formal languages have been developed for the deductive data model, e.g. Datalog$^+$, a function-free sublanguage of first-order predicate logic, or Datalog which includes negation.

Logic programming languages, e.g. Prolog, extend Datalog with complex data structures. Current Prolog systems feature access to external DBMS or an internal database component for the persistent storage of large amounts of data.

5.3.5.5.1 Summary

- The deductive data model is a logic based extension of the relational data model.
- Data is stored as facts and rules.
- Persistent logic programming languages combine the declarative logic language with persistent data storage.
- Implementations: *LDL*, Eclipse, Prolog.

5.3.6 Safe storage of data

DBMSs provide safety features that prevent the loss of data due to hardware failure. The basic mechanism is that of *transactions*. A transaction is a sequence of data definition or manipulation commands that is considered as atomic by the DBMS; a transaction either succeeds completely, or fails and undoes all data changes (*rollback*).

During a transaction, the DBMS works on a copy of the data; if the transaction succeeds, the copy is written to permanent storage, otherwise it is simply discarded.

Transactions may be visible to the user. Other mechanisms to preserve the physical integrity of data are invisible:

- Commit points, i.e. points in time at which the database contents is dumped to a permanent storage medium in a consistent way.
- Redundant data storage, e.g. through a *RAID*-array (Rapid access to inexpensive drives) organisation of disks where redundant data is distributed over hard disks to allow the reconstruction of a corrupted hard disk, or *mirroring*, where exact duplicates of devices are used.

5.3.7 Application-independent storage of data

Data held in a DBMS has a greater lifetime than the applications using the data. In fact, in most cases data will outlive the DBMS in which it is stored.

This means that data storage has to be organised independently of any application so that different applications (or generations of applications) can access the same data efficiently, and that import and export procedures have to be provided for the migration of data to new DBMSs.

In DBMSs, application independence is achieved by hiding the physical storage of data from the users. The DBMS automatically maps the storage requirements of a database schema to the file system of the operating system, creates a meta-schema which contains all information about the data stored in the DBMS and generates index files to speed up access to the data.

External applications cannot access data stored in a DBMS directly, e.g. by opening a file. On the one hand, this is a further security measure by which tampering with data is prevented, on the other hand this places the burden of determining efficient data storage organisation and optimising access to the data on the DBMS.

5.3.7.1 Controlled access to data

In general, DBMSs are multi-user systems, i.e. many users or applications may access the same data in parallel. In order to prevent the interference of user operations, access to the data is controlled by the DBMS through user-specific *views* and *access privileges*.

A user-specific view or *sub-schema* defines which entities and relationships are visible to the user. The entities and relationships visible to the user must not be identical to those of the global database schema.

Access privileges define what mode of access a user has to the data. Access privileges can be defined for the global database schema, the sub-schema, or individual entity sets (e.g. tuples in a relation table, objects in a class). The access mode is either *read* or *write*. In general, a user can grant his access privileges to other users.

5.3.8 Summary

In SLP, DBMSs should be used for speech corpora of *any* size.

The special requirements of SLP, namely

- the size of signal data,
- the complex datastructures for an adequate representation of data on different levels of abstraction, and
- the permanence of the data

cannot be met with the hierarchical and network data models.

Even the relational data model in its pure form is not well suited for SL-corpora because it lacks support for bit-stream data and complex data structures. However, all current commercial relational DBMS implementations support binary large objects and a richer type system, which makes them candidate DBMSs for SL-corpora. The major advantage of relational DBMSs is that they are now a standard and proven technology.

Object-oriented DBMSs, especially in conjunction with object-oriented programming languages, and persistent logic programming languages provide both a rich inventory of data types, full computational power, and a good support of the basic DBMS functionality. They are thus well suited for SL-corpora.

The major drawback of object-oriented DBMSs is that there is not yet a common standard object-oriented database language. The major drawback of persistent logic programming languages is that there exist only a few implementations of such systems. However, the logic based formalism is basically the same as in phonology and linguistics, and thus it is possible to use a single formalism on all levels from phonetics to linguistics.

The major advantages of a DBMS can thus be summarised as follows:

- Logically coherent data definition in a formal data model.
- Uniform access to data in a database schema.
- Stable database schema, flexible amount of data.
- Interactive and application program interfaces for ad-hoc querying and inter-application communication.

DBMSs are, for instance, applied in the GRECO project (cf. Carre (1992), or the Alvey STA project described in Thomas & Winski (1987)), or the PhonDat-Verbmobil database of spoken German (cf. Draxler (1995)).

It will be clear that only the basic characteristics of DBMSs are mentioned above. The interested reader should consult Elmasri & Navathe (1989), or Ceri, Gottlob & Tanca (1990) or any other recent database introductory text book.

Chapter 6

Spoken Language Lexica

6.1 Introduction

6.1.1 Lexica for spoken language systems

Spoken language systems are becoming increasingly versatile, and a central problem in developing such a system is the creation of its lexicon. In other areas of language processing, such as computational linguistics, the lexicon is also taking on a more and more central role. The lexicon of a spoken language system may be designed for broad or narrow coverage, for specific applications, with a particular kind of organisation, and optimised for a specific strategy of lexical search. Since the construction of a lexicon is a highly labour-intensive and thus also error-prone job, a prime requirement is for automatising lexicon development as far as possible, and in transferring lexical information from previous applications to new applications.

The main object of this chapter is to provide a framework for relating such concepts to each other and for the formulation of recommendations for development and use of lexica for spoken language systems.

In this introductory section, some basic concepts connected with the use and structure of lexica in spoken language systems are outlined; in the following sections, specific dimensions of spoken language lexica are discussed in more detail. Discussion is restricted to spoken language system lexica; non-electronic lexica for human use (e.g. pronunciation dictionaries in book form) are not considered. Features common to spoken and written language lexica, such as syntactic and semantic information in lexical entries, are only mentioned in passing; see the contribution of the EAGLES Working Group on Computational Lexica on these points. The close relation between spoken language lexica and speech corpora entails a minimal amount of overlap with the contribution of the Spoken Language Corpus chapter of this handbook, which should also be consulted.

In the remainder of the introductory section of this chapter, recent results in the development of spoken language lexica are summarised. The following sections of the chapter are concerned with basic features of spoken language lexica, lexical information, lexicon structure, lexical access, and lexical knowledge acquisition for spoken language lexica.

6.1.2 Lexical information as properties of words

At the present time, information about lexica for spoken language systems is relatively hard to come by. One reason for this is that such information is largely contained in specifications of particular systems, technical reports and books on various aspects of speech processing. Another is that there is not a close relation between concepts and terminology in the speech processing field, and concepts and terminology in traditional lexicography, natural language processing and computational linguistics. Components such as Hidden Markov Models for word recognition, stochastic language models for word sequence patterns, grapheme-phoneme tables and rules, word-oriented knowledge bases for semantic interpretation or text construction are all concerned with the areas of word identity and lexical access, lexical disambiguation, lexicon architecture and lexical representation, but these relations are not immediately obvious within the context of speech technology as a whole, and stochastic word models, for instance, would not generally be regarded as providing lexical information (though in the strict sense of the term, they evidently do provide such information). For the purposes of this handbook, the broader view will be taken.

It is customary to distinguish between system lexica and lexical databases. The distinction between the two may, in specific cases, be blurred, and the unity of the two concepts may also be rather loose, if the system lexicon is highly modular or several lexical databases are used. However, the distinction is a useful one. The distinction between lexica and lexical databases will be discussed below. Since the kinds of information in both these types of lexical object overlap, the term "spoken language lexicon" will generally be used in this chapter to cover both types.

6.1.3 Applications of spoken language lexica

With the advent of organisations for coordinating the use of language resources, such as ELRA (The *European Language Resources Association*) and the word-wide LDC *Linguistic Data Consortium*, access to information on spoken language lexica will become more widely available. The following overview is temporary, and therefore necessarily partial.

6.1.3.1 Types of application for spoken language lexica

Lexica for spoken language are used in a variety of systems, including the following:

- Automatic spelling correctors (spelling is determined to a large extent by phonological considerations).
- Medium and large-vocabulary automatic speech recognition (ASR), as in systems such as SPICOS described in Höge et al. (1985), Dreckschmidt (1987), Ney et al. (1988), Thurmair (1986), HEARSAY-II described in Lesser et al. (1975), Erman (1977), Erman & Lesser (1980), Erman & Hayes-Roth (1981), SPHINX described in Lee, Hon & Reddy (1990), ISADORA, (cf. Schukat-Talamazzini (1993)), or, for example in automatic dictation machines such as IBM's TANGORA (cf. Averbuch et al. (1986), Averbuch, Bahl & Bakis (1987), Jelinek (1985)) and DragonDictate by Dragon Systems (cf. Baker (1975a), Baker (1975b), Baker (1989), Baker et al. (1992)).
- Speech synthesis in text-to-speech systems, for example in reading machines, speaking clocks. For further speech synthesis applications, various relevant studies such as Allen, Hunnicutt & Klatt (1987), Bailly & Benoît (1992), Bailly (1994), Van Coile (1989), Klatt (1982), Klatt (1987), Hertz, Kadin & Karplus (1985), Van Hemert, Adriaens-Porzig & Adriaens (1987) can be consulted. See also chapter 4 on spoken language output assessment.
- Interactive dialogue systems, with speech front ends to databases and enquiry systems and synthesised responses (see for instance Brietzmann et al. (1983), Niemann et al. (1985), Niemann et al. (1992), Bunt & al. (1985)); see also the chapter on interactive dialogue systems.
- Speech-to-speech translation systems as developed in the ATR and VERBMOBIL projects, which use various speech recognition techniques, including continuous speech recognition, recognition of new words, word-spotting in continuous speech. For speech translation systems see for instance Rayner et al. (1993) and Woszczyna et al. (1993).

- Lexica and encyclopaedias on CD-ROM with multi-media (including acoustic) output.
- Research and development of spoken language processing systems, in the process of which broader based lexica for written language, coupled with tools such as grapheme-phoneme converters, may be used as sources of information.

6.1.3.2 Spoken language lexical databases as a general resource

Spoken language lexica may be components of systems such as those listed above, or reusable background resources. System lexica are generally only of local interest within institutes, companies or projects. Lexical databases as reusable background resources which are intended to be more generally available raise the question of standardised storage and dissemination. In general, the same principles apply as for Spoken Language Corpora: they are stored and disseminated using a variety of media. In research and development contexts, magnetic media (disk or tape) were preferred until recently; in recent years, local magnetic storage and wider informal dissemination within projects or other relevant communities via electronic networks such as Internet using standard file transfer protocols and electronic mail has become the norm. Large lexica, and corpora on which large lexica are based, are in general stored and disseminated in the form of ISO standard CD-ROMs.

The following brief overview can do no more than list a number of examples of current work on spoken language lexicography. At this stage, no claim to exhaustiveness is made, and no valuation of cited or uncited work is intended. It is intended to include a more detailed overview in later versions of this chapter.

- A number of general lexica with information relevant to spoken language have already been available on CD-ROM for quite some time, including the Hachette and Robert (9 volume) dictionaries for French, the Oxford English Dictionary, the Duden dictionary for German, and the Franklin Computer Corporation Master 4000 dictionary with acoustic output for 83000 words (cf. Goorfin (1989)).
- Several lexica with more restricted circulation have been developed in the context of speech technology research and development. Companies such as IBM, and telecom research and development institutes such as CNET in France have developed large lexica (CNET, for instance, has a 55000 word and 12000 phrase lexicon).
- University and other research institutes have also constructed large lexica; in France, for example, such institutes as ENST in Paris, ICP in Grenoble (cf. Tubach & Bok (1985)), Paris (cf. Plenat (1991), for a pronunciation dictionary of abbreviations) and IRIT in Toulouse (the BDLEX project) have worked on large spoken language lexica. The BDLEX-1 lexicon coordinated by IRIT (cf. Pérennou & De Calmès (1987)) contains 23000 entries, and BDLEX-2 (cf. Pérennou et al. (1991), Pérennou et al. (1992), Pérennou & Tihoni (1992)) contains 50000 entries; a set of linguistic software tools permits the construction of a variety of daughter lexica for spelling correction and lemmatisation, and defines a total of 270000 fully inflected forms.
- The Belgian BRULEX psycholinguistic lexicon contains information on uniqueness points (the point in a letter tree where a word form is uniquely identified), lexical fields, phonological patterns and mean digram frequencies for 36000 words (cf. Content, Mousty & Radeau (1990)).
- In the United Kingdom, the Alvey project resulted in tools and lexical materials (cf. Boguraev et al. (1988)).
- In the Netherlands, the Nijmegen lexical database CELEX (cf. Baayen (1991)), also available on CD-ROM, contains components with 400000 Dutch forms, 15000 English forms and 51000 German forms, together with an access tool FLEX.

- For German, lexical databases for spoken language lexica have been constructed by companies such as Siemens, Daimler-Benz, IBM and Philips, as well as in university speech technology departments (e.g. Munich, Erlangen, Karlsruhe, Bielefeld), and in the VERBMOBIL project (Gibbon (1995), Gibbon & Ehrlich (1995)).
- Work in computational lexicology and computational phonology has led to the development of structured lexicon concepts for spoken language such as ILEX (cf. Gibbon (1992), Bleiching (1992)) based on the DATR lexical knowledge representation language (cf. Evans & Gazdar (1989), Evans & Gazdar (1990)); the DATR language has been applied to word form lexica in the multilingual SUNDIAL project (cf. Andry et al. (1992)) by the German partner Daimler-Benz and in the German VERBMOBIL project (cf. Gibbon (1993)).
- The European Commission has funded a number of projects, particularly within the ESPRIT programme, in which questions of multilingual spoken language system lexica have been addressed, albeit relatively indirectly (POLYGLOT, SUNDIAL, SAM, SAM-A), as well as lexicography projects such as MULTILEX in the ESPRIT programme (cf. Heyer, Waldhur & Khatchadourian (1991)), GENELEX in the EUREKA programme (cf. Nossin (1991)) and AQUILEX, which concentrate on multi-functional written language lexica, though extension of the results to spoken language information has been provided for by the adoption of general sign-based lexicon architectures (see the results of the EAGLES Working Group on Computational Lexica).

6.1.3.3 Lexica in selected spoken language systems

The range of existing spoken language systems is large, so that only a small selection can be outlined, concentrating on older and established systems whose lexicon requirements are fairly well known; the situation is currently in a state of flux and for this reason the most recent developments are not included. Small vocabulary systems are also excluded, as their strong points are evidently not in the area of the lexicon. The concepts referred to in the descriptions are discussed in the relevant sections below.

HARPY is a large-vocabulary (1011 words) continuous speech recognition system. It was developed at Carnegie Mellon University. HARPY was the best performing speech recognition system developed under the five-year ARPA project launched in 1971. HARPY makes use of various knowledge sources, including a highly constrained grammar (a finite state grammar in BNF [Backus Naur Form] notation) and lexical knowledge in the form of a pronunciation dictionary that contains alternative pronunciations of each word. Initial attempts to derive within-word phonological variations with a set of phonological rules operating on a baseform failed. A set of juncture rules describes inter-word phonological phenomena such as /p/ deletion at /pm/ junctures: /helpmi/ — /helmi/. The spectral characteristics of allophones of a given phoneme, including their empirically determined durations, are stored in phone templates. The HARPY system compiles all knowledge into a unified directed graph representation, a transition network of 15,000 states (the so-called blackboard model). Each state in the network corresponds to a spectral template. The spectra of the observed segments are compared with the spectral templates in the network. The system determines which sequence of spectra, that is, which path through the network, provides the best match with the acoustic input spectral sequence.

(cf. Klatt (1977); see also Lowerre & Reddy (1980))

HEARSAY-II also uses the blackboard principle (see HARPY), where knowledge sources contribute to the recognition process via a global data base. In the recognition process, an utterance is segmented into categories of manner-of-articulation features, e.g. a stop-vowel-stop pattern. All words with a syllable structure corresponding to that of the input are proposed as hypotheses. However, words can also be hypothesised top-down by the syntactic component. So misses by the lexical hypothesiser, which are very likely, can be made up for by the syntactic predictor. The lexicon for word verification has the same structure as HARPY; It is defined in terms of spectral patterns.

(cf. Klatt (1977), see also Erman (1977) and Erman & Lesser (1980))

SPHINX is currently regarded as the state of the art system. It is a large-vocabulary continuous speech recognition system for speaker-independent application. It was evaluated on the DARPA naval resource management task. The baseline SPHINX system works with Hidden Markov Models (HMMs) where each HMM represents a phone. The total of phones is 45. The phone models are concatenated to create word models, which in turn serve to create sentence models. The phonetic spelling of a word was adopted from the ANGEL System (cf. Rudnicky et al. (1987)). The SPHINX baseline system has been improved by introducing multiple codebooks and adding information to the lexical-phonological component:

- The most likely pronunciation was substituted for the baseform pronunciation of a lexical item in the pronunciation dictionary, retaining the assumption that each lexical item has only one pronunciation.
- Different models were created for phones that have typically more than one realisation such as released and unreleased /d/ at the beginning of /dIdmaI/ and before /m/, respectively.
- Two subword units were introduced: function-word-dependent phone models and generalised triphone models. Since function words are typically unstressed, phones in function words are very often deleted or reduced, do not serve as proper models for recognition tasks, and account for almost 50% of the errors.

The SPHINX system works with grammars of different perplexity (average branching factor; see the section on word models); the grammars are of a type which can, in principle, be regarded as a specialised tabular, network-like or tree-structured lexicon with probabilistic word-class information:

- A null grammar with a perplexity of 997 (i.e. a vocabulary of 997 words was used); in a null grammar any word can succeed a given word.
- A word-pair grammar with a perplexity of 60; word-pair grammars are lists of words that can follow a given word.
- A bigram grammar with a perplexity of 20; this is a word-pair grammar equipped with word-category transition probabilities.

In word recognition tests, the best results were obtained with the bigram grammar, the most restrictive kind of the grammars mentioned above (96% accuracy compared with 71% for null grammars).

The SPHINX system makes use of various levels of representation of linguistic units:

- phone models (generalised triphones and extra models for function words)
- word models (stored in the pronunciation dictionary with one representation for each word).
- sentence models (for final confirmation)

(cf. Lee, Hon & Reddy (1990); see also Alleva et al. (1992))

EVAR ('Erkennen — Verstehen — Antworten — Rückfragen') is a large-vocabulary continuous speech recognition and dialogue system. It is designed to understand standard German sentences and to react either in form of an answer or a question referring back to what has been said, within the specific discourse domain of enquiries concerning Intercity timetables. The EVAR lexicon has the following properties:

- The lexicon includes not only sublanguage-specific words but also many words of the general vocabulary that occur during a dialogue of this kind.
- The lexicon contains fully inflected word forms.
- The baseforms, so-called *Normalformen*, e.g. infinitive for verbs, nominative singular for nouns, contain information relevant for all grammatical forms, thus reducing redundancy in the lexicon.
- The lexicon contains phonological, syntactic, semantic, and pragmatic information.
- Since the system modules need access only to special lexical knowledge (the *articulation module* makes use of phonological information, while the module in charge of generating the surface structure of an answer also needs syntactic information), access of individual modules to the lexicon is restricted. Preprocessors extract the subset of information relevant for each module.
- The lexical unit in the EVAR lexicon is the graphematic word (*graphematisches Wort*); so-called phonetic words (standard pronunciation) and so-called grammatical words (syntactic categories plus meanings) are assigned to the graphematic words.
- Lexical units are described in attribute-value notation. For example, the attribute WORT takes a graphemic word as its value.
- Graphemic words again have the attributes AUSSPRACHE and SYNTAX-TEIL for which values are defined in the form of a Duden standard pronunciation and morpho-syntactic properties such as the attribute-value pair WORTART-Verb. Numbers keep track of the various entries for different meanings or syntactic variants (e.g. reflexive — non-reflexive), etc. of a lexical item.
- In the baseform entries, information on stem, pronunciation of the stem (in ASCII symbols that replace the standard IPA notation), and the inflection pattern is given under SYNTAX-TEIL.
- Semantic information includes specifications of semantic features and valence properties as well as selectional restrictions. Fillmore's system of deep structure cases as suggested in Fillmore (1968) has been expanded to 28 cases.

A lexicon administration system is being developed which uses tools for extracting words according to specified criteria, such as "Look for nouns that express a location" or "Look for prepositions that express a direction".

(cf. Ehrlich (1986), Brietzmann et al. (1983), Niemann et al. (1985), Niemann et al. (1992))

6.1.4 Recommendations on resources

The following recommendations should be seen in conjunction with recommendations made after the following more specialised sections of this chapter.

1. For information on resources in specific areas of spoken language systems, consult the other chapters and appendices in this handbook.
2. For information on resources which relate to written language, consult the other EAGLES working groups.
3. For general information on resources consult the organisations ELRA (*European Language Resources Association* and LDC *Linguistic Data Consortium*
4. In preparation for decisions on the use of resources, distinguish between *lexical database* and the *system lexicon*

5. Identify the types of information required for the lexical database and the system lexicon.
6. Consider the relevant lexical database models and system lexicon architectures.
7. Develop a systematic concept for the tools required in producing and accessing a lexicon or a lexical database.

6.2 What is a spoken language lexicon?

6.2.1 Basic features of a spoken language lexicon

A spoken language lexicon may be a component in a system, a *system lexicon*, or a background resource for wider use, a *lexical database*, in each case containing information about the pronunciation, the spelling, the syntactic usage, the meaning and specific pragmatic properties of words; lexica containing subsets of this information may also be referred to as spoken language lexica, though the simpler cases are often simply referred to as wordlists. Where there is little danger of confusion, the term *spoken language lexicon* will be used to refer indifferently to either a *spoken language system lexicon* or a *spoken language lexical database*

A *spoken language lexicon* is defined as a list of representations of lexical entries consisting of spoken word forms paired with their other lexical properties in such a way as to optimise lookup of these properties. This definition covers a wide range of specific types of spoken language lexicon, from lists in which orthography provides a more or less indirect representation of a spoken word form pronunciation through tables and dictionaries based on phoneme representations to declarative knowledge bases from which details of lexical information are inferred from specific premises (entries) about individual lexical items and general premises (rules) about the structure of lexical items. It also covers application directed special lexicon types based, for instance, on the different requirements for pronunciation tables for recognisers and for synthesisers.

The different kinds of spoken language lexicon are primarily orientated towards the forms of words rather than towards their distribution in larger units or their meaning. The current state of the art prefers, where possible, to deal with finite sets of fully inflected words which are actually attested in corpora, rather than with the composition of words on morphological principles. Questions of word composition are increasing in importance in projects aimed at the recognition of spontaneous continuous speech, in which unknown words from a wider vocabulary, or ad hoc coinages (nonce forms) are encountered, and require intelligent treatment in a robust processor.

6.2.2 Lexical databases and system lexica for spoken language

The distinction between lexical databases and system lexica is a useful one, though in practice more complex distinctions are required. The main characteristics of the two kinds of lexical object are outlined below.

Lexical database: A spoken language lexical database is often a set of loosely related simpler databases (e.g. pronunciation table, signal annotation file, stochastic word model, main lexical database with syntactic and semantic information).

- Purpose:
  - Resource for system development (training, evaluation; construction of stochastic language models)
  - Definition of vocabulary coverage
  - Basis for vocabulary consistency maintenance

  - Reference point for integrating different kinds of lexical information
  - Source of information for investigation of vocabulary structure

- Structure:
  - Generally fixed record structures, with fields for different types of lexical information, and strings as values in fields
  - Often identification of *lexical key (lexical identifier)* with *orthographic word form* (Disadvantage: problem with orthography, particularly with large vocabularies: homographs are different lexical items, with different pronunciations and meanings, and are thus an artefactual source of "lexical noise")
  - Alternative for larger databases with more complex linguistic information: Unique identification of word as a more abstract unit with an abstract formal identifier, with specific properties including orthography, pronunciation, etc., on an equal footing.
  - Implementation generally conforming to local laboratory standards as a database of ASCII strings, created and accessed by means of standard UNIX tools and UNIX shell scripts, or C programmes; in more complex environments perhaps with a commercial database such as *Oracle*

- Content:
  - Main lookup key attribute (in general orthographic representation, perhaps supplemented by numbering of homographs)
  - Database entries may be *fully inflected forms*, *uninflected stems*, or *morphemes* (generally *morphs*, i.e. the phonemic forms of morphemes), or all of these; other inventories containing units such as phonemes, diphones or syllables, may be included
  - Pronunciation (in canonical phonemic representation, perhaps including pronunciation variants
  - Sub-word boundaries between units such as syllables, morphs (phonemic forms of affixes, lexical roots), derived stems and constituents of compound words
  - Syntactic category (e.g. Noun, Adjective, Article, Pronoun, Verb, Adverb, Preposition, Conjunction, Interjection) or subcategory (e.g. Proper vs. Common Noun, Intransitive vs. Transitive vs. Ditransitive vs. Prepositional, etc., Verb)
  - Semantic categories (domain or application specific)
  - Corpus information: frequency statistics; concordance information (i.e. list of contexts of occurrence for each word); signal annotations
  - Further information: Translation information (in

System lexicon: Lexical information (i.e. properties of words) referred to during the speech recognition or synthesis process may not be concentrated in one identifiable lexicon in a given system.

- Purpose: Definition of those properties of words required for recognition, parsing and understanding, or for planning, formulation and synthesis.
- Structure: In general separate modules for different properties of words with different functions within the system (which are often not regarded as having anything at all to do with a lexicon)
  - In speech recognition: Modules such as the word recogniser (typically based on *Hidden Markov Model* technology), which identifies word forms, i.e. recognition oriented lexical access keys, often in orthographic representation), the stochastic language model (which defines statistical properties of words in their immediate contexts as *bigrams*, *trigrams*, etc.), and the main lexicon with syntactic and semantic information, perhaps linked to an application-specific database or knowledge-base.

  - In speech synthesis: Modules defining word structure in terms of morpheme sequences, word-prosody (e.g. accentuation), and pronunciation (in terms of phonemes), supplemented by detailed rules for phoneme variants in different contexts and for timing information.

- Content: Application specific; subsets of information defined in the lexical database resource, as outlined under "Structure".

6.2.3 Spoken language and written language lexica

Spoken language lexica differ in coverage and content in many respects from lexica for written language, although they also share much information with them. Written language lexica are generally based on a stem, neutral or morphologically canonical form (e.g. nominative singular; infinitive), or headword concept, in which generalisations over morphologically related forms may be included, leading to fairly compact representations. Spoken language lexica for speech recognition are generally based on fully inflected word forms, as in the dictation system TANGORA with about 20000 entries. Depending on the complexity of inflectional morphology in the language concerned, the number of entries is larger than the number of entries in a corresponding conventional dictionary based on stems or neutral forms by a factor between 2 or 3 and several thousand. Speech synthesis systems for text-to-speech applications generally do not rely on extensive lexica, but use rule-based techniques for generating pronunciation forms and prosody (speech melody).

In an orthographically oriented lexicon, it is generally sufficient to include a canonical phonemic transcription, based on the citation form of a word (the pronunciation of a word in isolation) which can be utilised, for example, in sophisticated tools for automatic spelling correction. However, this is inadequate for the requirements of speech recognition systems, in which further details are required.

A spoken language lexicon contains information about pronunciation and pronunciation variants, often including prosodic information about syllable structure, stress, and (in tone and pitch accent languages) about lexical tone and pitch accent, and morphological information about divisions into stems and affixes. Spoken language lexica are in general much more heavily orientated towards properties of word forms than towards the distributional and semantic properties of words.

It may happen that a canonical form or a canonical pronunciation does not actually occur in a given spoken language corpus; this would be of little consequence for a traditional dictionary, but in a spoken language dictionary it is necessary to adopt one of the following solutions:

1. Use the canonical form, but mark it as non-occurring.

2. Adopt an attested form as canonical form (e.g. nouns occurring only in the plural such as French 'ténèbres' darkness, English 'trousers', German 'Leute', people).

At a more detailed level, orthography (the division of word forms into standardised units of writing) and phonology (the division of word forms into units of pronunciation) are related in different ways in different languages and also to the morphology (the division of word forms into units of sense) of the language. The orthographic notion of 'syllable' serves, in general, in written language lexica for defining hyphenation at line-breaks and certain spelling rules; for this purpose, morphological information about words is also generally required. In spoken language, however, the phonological notion of 'syllable' is different; it refers to units of speech which are basic to the definition of the well-formed sound sequences of a language and to the rhythmic structure of speech, and forms the basis for the definition of variant pronunciations of speech sounds.

Morphology is often described at an abstract level which permits generalisations over spoken language and written language forms. However, when complex word forms are put together from combinations of smaller units, different *alternations* of orthographic units (letters) and phonological units (phonemes) occur at the boundaries of the parts of such words. Furthermore, additional kinds of lexical unit are required in the lexicon of a spoken language dialogue system, such as discourse particles, hesitation phenomena, and pragmatic idioms, or so-called *functional units* (sequences of functional words which behave as a phonological unit) and *clitics* (functional words which combine with lexical words to form a sequence which behaves as a phonological unit).

6.2.4 Basic lexicographic coverage criteria

Criteria for the coverage of lexica for spoken language processing systems are heavily corpus determined, and differ considerably from criteria for coverage of lexica for traditional computational linguistics and natural language processing. In theoretical computational linguistics, interests are determined by systematic fragments of natural languages which reveal interesting problems of representation and processing. In natural language processing, broad coverage is often the goal. In spoken language lexica as currently used in speech technology, lexica are always oriented towards a particular well-defined corpus, which has often been specifically constructed for the task in hand. When speech technology and natural language specialists meet, for instance in comprehensive dialogue oriented development projects, these differences of terminology and priorities are a potential source of misunderstanding and disagreement, and joint solutions need to be carefully negotiated.

The main coverage criteria for spoken language lexica may be summarised as follows.

- Completeness (all word types in the training and test corpora).
- Minimality (only word types in the training and test corpora).
- Consistency (with respect to the training and test corpora and other related data types).
- Generality (projection of the word form set on to related forms not in the corpus).

These criteria pertain to words; if other units are used, the criteria apply analogously to these.

The first three criteria are essentials for the current state of speech technology. Conventional expectations in written language processing (in Computational Linguistics and in Natural Language Processing) are widely different, and are expressed in the fourth criterion. Clearly the second and fourth criteria clash; the relation to relevant corpora must therefore be carefully flagged in a spoken language lexicon. The degree of coverage (which for a speech recognition system generally has to be 100%) is expressed for lexica in general in terms of the notions of degree of static coverage (ratio of words in a corpus which are contained in a given dictionary to the number of words in the corpus) and the degree of dynamic coverage (the probability of encountering words which have previously been encountered); the latter value is generally higher than the former (cf. Ferrané et al. (1992)). On the basis of corpus statistics for typologically different languages such as English (cf. Averbuch, Bahl & Bakis (1987)) and French (cf. Mérialdo (1988)), two languages which differ widely in their inflectional structure (English with few verbal inflections, French with a rich verbal inflection system), interesting quantitative comparisons can be made.

	Vocabulary (no. of forms)	Static coverage	Dynamic coverage
English	5000	92.5%	
English	20000	97.6%	
French	20000	94.9%	98.2%
French	200000	97.5%	99.5%

6.2.5 The lexicon in spoken language recognition systems

In a spoken language recognition system, the lexicon is generally divided into two components: the recognition component and the search component. In the recognition component, intervals of the speech signal are mapped by probabilistic systems such as Hidden Markov Models, Neural Networks, Dynamic Programming algorithms, Fuzzy Logic knowledge bases, to word hypotheses; the resulting mapping is organised as a *word lattice* or *word graph*, i.e. a set of word hypotheses, each assigned to a temporal interval in the speech signal. The term *word* is used here in the sense of 'lexical lookup key'. The keys are traditionally represented by orthography, but would be better represented in a spoken language system by phonemic transcriptions. The search component has the task of finding the correct word form from the set of hypotheses delivered by the recognition component, and the correct lexical properties of this word form.

In spoken language recognition system development, a corpus based lexicon of orthographically transliterated forms is used as the basis for a pronunciation lexicon; the lexicon is often supplemented by rules for generating pronunciation variants. This lexicon is required in order to tune the recognition system to a specific corpus by statistical training: frequencies of distribution of words in a corpus are interpreted as the a priori probabilities of words in a given context. These a priori probabilities may be based on the absolute frequencies of words, or on their frequencies relative to a given context, e.g. *digram* (*bigram*) frequencies. The functionality of spoken language lexica may be summarised in the following terms.

- Offline functions
  - Fully inflected word form list construction
  - Pronunciation table construction
    - Synthesiser development
    - Recogniser (forced alignment, stochastic training)
    - Transliteration checking
    - Integration of word and sentence prosody
    - Integration of morphology for
      - describing new words
      - sparse data training with stems and word classes
  - Frequency table construction
  - Word distribution frequency tables
  - Language models (n-gram tables)
  - Coverage definition for inter-project coordination
- Online functions
  - Definition of forms for word graph interface
    - orthographic (conventional technology)
    - phonological (new architectures)
  - Definition of criteria for lexicon architecture and lookup

6.2.6 Recommendations on defining spoken language lexica

1. Define the anticipated functions of the offline lexical database in spoken language system development and the online system lexicon, bearing in mind the differences between written text and speech, the intended application, and possible economies to be gained in using or creating reusable lexical resources.

2. Speficy the quantitative coverage (size) of the lexical database and of the system lexicon with reference to the available system components and in terms of .

3. Specify the qualitative coverage (content) of the lexical database and of the system lexicon with reference to the application domain and in terms of standard types of lexical information.

6.3 Types of lexical information in spoken language lexica

6.3.1 Lexicon models and lexical representation

A given system lexicon or lexical database is based on a *lexical information model* or a *data model*; often the model is intuitively constructed, or based on notions taken from traditional school grammar, but scientifically motivated models are becoming available. A model of lexical information will make at least the following distinctions:

- Lexical objects: The basic objects (such as words) described in a lexicon. It is becoming customary in lexicography and computational linguistics to refer to the *lexical sign*, i.e. an object associated with attributes denoting orthogonal kinds of lexical information. A second kind of lexical object is the *lexical sign class* in which similar lexical objects are grouped together, each characterised by subsets of the lexical information required to characterise specific lexical signs. These class-based generalisations may be organised in terms of *implication rules (*redundancy rules*), or* type hierarchies, or *default inheritance hierarchies*.

- Lexical information: In a theoretically well-founded lexicon which satisfies formal criteria of consistency and coverage criteria such as empirical completeness and soundness, types of lexical information are *orthogonal*, i.e. of different types which complement each other. These orthogonal types of lexical information are often labelled with *attribute* names, and the items of information regarded as the *values* of these attributes. The types include orthography, pronunciation, syntactic distributional properties, meaning, and pragmatic properties of use in context (e.g. speech act type, stylistic level).

Lexicon models for lexical databases and system lexicons are part of the overal conceptual framework required for lexicon development. Modern approaches to lexicon development provide suitable lexical representation languages for formulating and integrating the different kinds of lexical information specified in a lexicon model and assigning them to lexical objects, and implementations for these representation languages (cf. Andry et al. (1992)). In recent work, the following useful distinctions are sometimes made:

- Lexicon formalism: A specially designed logic programming language such as DATR, though high level languages such as LISP or Prolog are often used, with compiler concepts for translating these languages into conventional languages for efficient processing. Particular styles of programming in more traditional languages such as C are sometimes used directly for smaller lexicons.

- Lexicon theory: A coherent and consistent set of expressions formulated in a well-defined formalism and interpreted with respect to a lexicon model.
  - General lexicon theory: A general theory of lexical objects and information, for instance a theory of lexical signs and their representation.
  - Specific lexicon theory: A given lexicon formulated in a lexicon formalism on the basis of a lexicon model.

- Lexicon model: Specification of the domain denoted by a lexicon theory, conceptually independent of the theory itself (cf. the notion of a *data model* for a database). A different definition is also common: the general structure of the objects and attribute-value structures in a formal lexicon. A lexicon model specifies the following kinds of information:
  - Types of lexical object and structure of lexical entries.
  - Types of lexical information associated with lexical objects in lexical entries.
  - Relations between lexical objects and structure of the lexicon as a whole *lexicon architecture.*

The aspects of representation and architecture will be dealt with in a later section. The following subsections are concerned with the main kinds of lexical information required for spoken language lexical entries.

6.3.2 A simple sign model for lexical properties

Lexical information is often regarded as a heterogeneous collection of idiosyncratic information about lexical items. An assumption such as this makes it hard to discuss lexical information systematically and, moreover, from the point of view of contemporary lexicography, it is wrong. For this reason, a simple unifying informal model of lexical signs, related to a view which is current in computational linguistics and computational lexicography, is introduced for the purpose of further discussion.

In general terms, a sign is a constituent of a pattern in communication with identifiable form and meaning. Lexical signs are *words* or perhaps also *phrasal idioms* and other items such as *dialogue control particles* (*er*, *uhm*, *aha* etc.); it may also be argued that even smaller units such as morphemes also have sign structure. Possible lexical signs range, essentially, over fully inflected word forms, morphs (roots, affixes), stems (roots or stems to which affixation has applied), lemmas (or lemmata), often represented by an orthographic form, and phrasal items (idioms).

Signs are characterised by the following four basic types of information:

1. Surface properties: *orthographic* and *phonological* representation; for pronunciation, several different levels of transcription are possible (morphophonemic, phonemic, phonetic).
2. Semantic properties: *semantic* and *pragmatic* representation.
3. Distributional properties: *syntactic category* and *subcategory* (e.g. *Verb*, *Transitive Verb*).
4. Compositional properties: *head* and *modifier* (*complement* or *specifier*) constituents; word formation is recursive:

The first two types are referred to as *interpretative* properties, since they interpret the basic sign representation in terms of the real world of phonetics (or writing) and meaning, while the second two types are referred to as *structural* (or *syntactic*, in a general sense of the term) properties. Complex signs are constructed compositionally from their constituent signs and derive their properties compositionally from these.

The following sections will be devoted to the four main types of lexical information, referring to them as *surface*, *content*, *grammatical* and *morphological* information, respectively.

The model used here for describing lexical information is based on the widespread current view of linguistic units as language signs and sign components. A number of complexities occur in the relations between the types of information, which will be discussed in a later section.

In the examples below, a basic computer-readable attribute-value syntax is used, based on the kind of spoken language lexical representation in DATR used by Andry et al. (1992). The name of the lexical sign is written with an initial upper case letter and followed by a colon, attribute names can be either word-like atoms or sequences of atoms (in the latter case, permitting an indirect representation of more complex attribute structures); they are enclosed in corner brackets and separated from their values by an equality sign, and the lexical sign is terminated by a period. The SAMPA notation is introduced below and is defined in the Appendices; see also the chapter on Spoken Language Corpora.

```
Table:
  <surface orthography>      = table
  <surface phonetics sampa> = teIbl
  <semantics>                = artefactual horizontal surface
  <distribution>             = noun common countable
  <composition>              = simplex z_plural.
```

In the case of complex signs, the meaning of the sign is a function of the meanings of its parts and the pronunciation of the sign is a function of the pronunciations of its parts. These functions may be somewhat idiosyncratic with lexical signs; this is shown by the pronunciation and meaning of words like English "dustman".

```
Dustman:
 <surface orthography>       = dustman
 <surface phonetics sampa>   = dVsm@n
 <semantics>                 = 'municipal garbage collector'.
```

The *pronunciation* and *meaning* of this complex lexical sign are not in all respects a general compositional function of its parts:

```
Dust:
  <surface phonetics sampa> = dVst
  <semantics>                = 'just visible particles of solid matter'.

Man:
  <surface phonetics sampa> = m{n
  <semantics>                = 'male adult human being'.
```

In contrast, the *spelling* and the *distribution* of the complex sign are perfectly regular functions of the spellings of the parts and the distribution of the head (i.e. Man) of the sign.

In perfectly regular cases, there is no necessity to include complex words in the lexicon. Such cases are practically non-existent, however, since complex words are in general partially idiosyncratic; in a comprehensive spoken language lexicon, both these and their parts need to be included. For most current practical purposes, in which *potential words* (unknown words or ad hoc word formations) do not need to be treated in addition to *actual words* (those contained in a lexicon), complex words can be listed in full as unanalysed forms.

Modern lexicographic computational lexicographic practice attempts to reduce the redundancy in a lexicon as far as possible; fully regular information in compounds can be inherited from the parts of the compounds, while idiosyncratic information is specified locally. In a case like this, a lexical class is specified for defining the structure of compounds, and "inheritance pointers" are included. The result is a *hierarchical lexicon structure*. Reference should be made to the literature cited in this section and in the section on *Lexicon structure* for further information on this approach to lexical lingware development.

6.3.3 Lexical units

6.3.3.1 Kinds of lexical unit

Intuitively, the prototypic lexical unit is a word. This definition has a number of catches to it, however, because words are not as simple as they seem. The intuitive notion of *word* has "fuzzy edges", as in the following cases.

1. Words may contain other words (e.g. compound words such as *database*, *Sprachtechnologie*).

2. Words have different status in respect of their phonetic realisations and their meaning; compare the difference between function words, e.g. *to*, *for* with reduced pronunciations and structural meanings, and content words, e.g. *word*, *spell*, which refer to real world objects, properties, event types.

3. Words may be merged with other words in informal speech (*cliticisation*). Examples of clitics are English *'is* in *he's* — /hi:z/, French *l'* in *il l'a vu* — /il la vy/, German *'m* in *auf'm Tisch* — /aUfm tIS/.

4. Particular types of word formation such as spelling and acronym formation may require special attention: *ecu* — /ik"ju/, /isi"ju/.

5. Words may be *inflected word forms*, making *sound* singular and *sounds* plural into different words; or words may be regarded as classes of inflectionally related forms (a *paradigm*), i.e. *sound* and *sounds* belong to the same word, which may be characterised by a canonical inflected form (e.g. nominative singular), or by the stem shared by the forms and identified by linguistic analysys, or by a number or other abstract label. In speech technology, the *inflected word form* is the standard definition; in standard dictionaries, the *paradigm* definition of word is used, and represented by a *headword* or *lemma*, generally the canonical inflectional form such as nominative singular, in orthographic representation.

6. Lexical units may need to be larger than the word (e.g. phrasal idioms).

7. Lexical units may need to be smaller than the word: semantically oriented morphological word sub-units (word constituents) include:

   - word stems minus inflections; indivisible word stems are *lexical morphemes*).
   - Constituent words words formed by compounding (composition)
   - Constituent prefixes, stems and suffixes in words formed by derivation

   Pronunciation oriented phonological word sub-units include syllables and their parts; phonological sub-units do not necessarily correspond closely with morphological sub-units.

8. The precise definition of the intuitively given concept of a word is theoretically more differentiated: linguistic textbooks distinguish between several different views of word-level lexical units, depending on which kind of information is regarded as primary:

   - The *phonological word* (based on its conformity to the phonological structure of a language).
   - The *prosodic word*, based on its conformity to the accentuation and the rhythm patterning of the language.
   - The *orthographic word* (for instance, as delimited by spaces or punctuation marks).
   - The *morphological word* (based on the indivisibility and fixed internal structure of words).
   - The *syntactic word* (based on its distribution in sentences).

9. The lexical word as a *type*, as opposed to an *occurrence* of the type in larger units, and a *token* of the type in a corpus of speech or writing.

The central meaning for the purpose of spoken language lexica will be taken to be the *morphological word*.

Lexical units (entries, items) are assigned sets of properties which identify lexical signs and determine the organisation of the lexicon. In practical contexts, the choice of lexical unit for a given lexicon is decisive for procedural reasons, i.e. in determining ways in which a lexicon may be most easily accessed: through orthography, pronunciation, meaning, syntactic properties, or via its morphological properties (stem, inflection). The application-driven decision on the kind of lexical unit for a given purpose is a non-trivial one. However, for many practical purposes fairly straightforward guidelines can be given. The form of a lexical item, in particular its orthography, is often used as the main identifying property for accessing the lexicon. However, access on phonetic grounds, via the phonological form, is evidently the optimal procedure for speech recognition, and access on conceptual semantic or syntactic grounds is evidently the optimal procedure for speech synthesis. The use of orthography as an intermediate stage in speech recognition is a useful and widespread heuristic which generally does not introduce significant numbers of artefacts into the mapping from speech signals to lexical items, but is not recommended for complex systems with large vocabularies, except as a means of visualisation in user interfaces. For text-to-speech applications, of course, orthography is the optimal lexical access key.

6.3.3.2 Fully inflected form lexica

It has already been noted that fully inflected form lexical and lexical databases are fairly standard for speech recognition. Where a small closed vocabulary is used, and new, unknown or ad hoc word formations are not required (as in most current applications of speech synthesis and recognition), fully inflected word forms are listed. This procedure is most convenient in languages with very small inflectional paradigms; for languages of the agglutinative type, in which large numbers of inflectional endings are concatenated, the procedure rapidly becomes intractable. In other applications, too, such as speech synthesis, it may be more tractable to generate fully inflected word forms from stems and endings.

An example of a language with few inflections is English, where (except for a few pronouns) only nouns and verbs are inflected, and even here only three forms for nouns (uninflected, genitive and plural) and four for verbs (uninflected, third person singular present, past, and present participle; irregular verbs in addition have a different past participle form). English is therefore not a good example for illustrating inflectional morphology (in other areas of morphology, i.e. in word formation, languages appear to be equally complex).

French is much more complex, including inflections on adjectives and large verb paradigms; orthographic inflection in French has more inflectional endings than are distinguished in phonological inflection.

German also has a complex inflectional morphology, with significantly more endings on all articles, pronouns, nouns, adjectives and verbs, increasing the size of the vocabulary over the size of a stem-oriented lexicon by a factor of about 4.

In extremely highly inflecting languages such as Finnish, the number of endings and the length of sequences of endings multiply out to increase the vocabulary by a factor of over 1000. Special morphological techniques have been developed (e.g. two-level morphology) to permit efficient calculation of inflected forms and to avoid finite but intractable explosion of lexicon size for highly inflecting languages (cf. Koskenniemi (1983) and Karttunen (1983)). These techniques have so far not been applied to any significant extent in speech technology.

The figures cited refer only to the sets of forms. When the *form-function mapping*, i.e. the association of a given inflected form with a morphosyntactic category, is considered, the figures become much worse. A single inflected adjective form such as *guten* in German has 44 possible interpretations which are relevant for morphosyntactic agreement contexts (cf.: Gibbon (1995)), with 13 *feminine* readings, 17 *masculine* readings, and 14 *neuter* readings, depending on different cases (*nominative*, *accusative*, *genitive* and *dative*) and different determiner (article) categories (*strong*, *weak* and *mixed*). It is possible to reduce the size of these sets by means of default-logical abbreviations in a lexical database, but for efficient processing, they ultimately need to be multiplied out. Similar considerations apply to the other word categories, and to the other highly inflecting languages.

The inflectional properties of many languages imply that, for these languages, large vocabulary systems with complex grammatical constructions require prohibitively large fully inflected form inventories. Although the sets of mappings involved can be very large, the inflectional systems of languages define a finite number of variants for each stem, and therefore it may make sense in complex applications to define a rule-based "virtual lexical database" or a "virtual lexicon" which constructs or analyses each fully inflected word form on demand using a *stem* or *morph lexicon* with a *morphological rule component.*

6.3.3.3 Stem and morph lexica

Lexica based on the morphological parts of words, coupled with lexical rules for defining the composition of words from these parts, are not widely used in current speech recognition practice. They are useful, however, in expanding lexica of attested forms to include all fully inflected forms, and in tools which verify the consistency of corpus transcriptions and lexica.

Terminology in this area is somewhat variable. In the most general usage, a *stem* is any uninflected item, whether morphologically simple or complex. However, intermediate stages in word formation by affixation, and in the inflection of highly inflected languages, are also called stems. The smallest stem is a *phonological morph* or an *orthographic morph*, i.e. the phonological or orthographic realisation of a *lexical morpheme*. Since stems may vary in different inflectional contexts, as affixes do, it is necessary to include information about the morphophonological (and morphographemic) alternations of such morphemes:

```
Knife:
  <surface phonology singular>   = naIf
  <surface phonology plural>     = naIv + z
  <surface orthography singular> = knife
  <surface orthography plural>   = knive + s.
```

The use of morphological decomposition of the kind illustrated here has been demonstrated to bring advantages in medium size vocabulary speech recognition (Geutner (1995)).

In a stem lexicon, the basic lexical key or *lemma* is the stem, which is represented in some kind of normalised notation. The most common kind of normalised or canonical notation has the following two properties:

1. Canonical inflected form: With morphologically inflected items, a "normal form" such as the *infinitive* for verbs or the *nominative singular* for nouns is used.

2. Canonical orthography: A standardised orthographic representation of the canonical inflected form is used.

For specific purposes, in which lexical entries need to be accessed on the basis of a specific property, indexing based, for instance, on the canonical phonemic representation, either of a fully inflected form or of the canonical inflected form, or even of the stem itself, may be used. This question is dealt with in more detail below.

6.3.3.4 The notion of 'lexical lemma'

As in the *knife* example, one particular form, for instance orthographic, of an entry is often used as a headword or lemma. From a technical lexicographic point of view, this form then has a dual function:

1. It names the entry.
2. It also represents one of its properties, namely its spelling.

In modern spoken language lexicography, this distinction is central, and ignoring it may lead to confusion. This applies particularly in the context of spoken language lexicography, where the primary criterion of acces by word form is phonological. The concept of an abstract lemma, deriving from recent developments in computational linguistics and their application to phonology and prosody, may be used in order to clarify the distinction (cf. Gibbon (1992)): an *abstract lemma* may have any convenient unique name or number (or indeed be labelled by the spelling of the canonical inflected form, as already noted); all properties have equal status, so that the abstract lemma is neutral with respect to different types of lexical access, through spelling, pronunciation, semantics, etc. The examples of lexical entries given so far are based on the concept of an abstract lemma. The neutrality of the abstract lemma with respect to particular properties and particular directions of lexical access make it suitable as a basic concept for organising flexible lexical databases. A lexicon based on a neutral abstract lemma concept is the basic form of a *declarative lexicon*, in which the structure or the lexicon is not dictated by requirements of specific types of lexical access (a *procedural lexicon*, but by general logical principles. The distinction between declarative and procedural lexica is a relative one, however, which is taken up in the section on spoken language lexicon architectures.

6.3.4 Lexical properties and lexical relations in spoken language

The relations between orthographic, phonological, syntactic and semantic properties of lexical units are complex, and make a theoretically satisfying definition of "lexical sign" quite elusive. Lexical relations are either *paradigmatic*, and define classes of similar items, or *syntagmatic*, and define whole items in terms of relations between their parts.

Present discussion is restricted to the main paradigmatic relations in traditional terms; the expression of these relations in terms of *semantic features*, *semantic markers* or *semantic components* is not dealt with explicitly, though it figures implicitly in the notion of attribute-value structures which is referred to in the examples.

The syntagmatic relations (roles; collocational relations) are more controversial, and are not dealt with here. For further information on these and on semantic components, reference should be made to the results of the EAGLES Computational Lexica Working Group and to standard semantics textbooks such as Lyons (1977) or Cruse (1986).

The following systematised versions of traditional definitions express the main paradigmatic relations between lexical signs.

1. The main relations of *form* between lexical signs:

   Homonymy: Two words with the same orthographic and phonological forms, but different syntactic categories and/or meanings are homonyms. Example: *mate* /meIt/ 'friend' or 'state of play in a chess game'.

   Homography: Two words with the same orthographic form and different phonological forms are (heterophonic) homographs. Example: *row* /roU/ 'horizontal sequence', /raU/ 'noise, quarrel'.

Homophony: Two words with the same phonological form and different orthographic forms are (heterographic) homophones. Example: *meet* /mi:t/ 'encounter' — *meat* /mi:t/ 'edible animal tissue'.

Heterography: Two orthographic forms of the same word are heterographs. Example: *standardise* — *standardize* /st{nd@daIz/.

Heterophony: Two phonological forms of the same word are heterophones. Example: *either* /aID@/ — /i:D@/ 'disjunction'.

2. The main relations of *function* between lexical signs:

Hyperonymy: If the meaning of one word is entailed by the meaning of another, it is a hyperonym of the other (a superordinate term relative to the other). Example: *book* is a hyperonym of *manual* as the meaning of *book* is implied by the meaning of *manual* (in one of its meanings).

Hyponymy: The converse of hyperonym. If the meaning of one word entails the meaning of another, it is a hyponym of the other (a subordinate term relative to the other). Example: *manual* is a hyponym of *book* as the meaning of *manual* implies the meaning of *book*.

Co-hyponymy: Two words are co-hyponyms if and only if there is a word which is a hyperonym of each (in the same meaning of this word). Example: *manual* and *novel* are co-hyponyms in relation to *book*.

Synonymy: Two words are synonyms if and only if they have the same meaning (or at least have one meaning in common), i.e. if the meaning of each entails the meaning of the other. They are partial synonyms if either has additional readings not shared by the other. They are full synonyms if they have no readings which is not shared by the other. Example: *manual* and *handbook* are partial synonyms (*manual* is also, among other things, a term for a traditional organ keyboard). Full synonyms are rare. By implication, synonyms are also co-hyponyms.

Antonymy: Two words are antonyms (a) if they are co-hyponyms with respect to given meanings, and (b) if they differ in meaning in respect of those details of the same meaning which are not shared by their hyperonym. Example: *manual* and *novel* are antonyms. Note that the term is sometimes restricted to binary oppositions, e.g. *dead — alive*.

In addition to these lexical relations, there are a number of structural complexities which hold between different types of information.

- Semantically, recursion in word formation is unrestricted, with left- or right branching, or centre-embedding.
- Morphologically, recursion is restricted to flat, linear concatenation, as in *Spracherkennungsevaluationsmethode* or *operationalisation*, which can be efficiently described and implemented by finite state devices.
- So-called *bracketing paradoxes* occur because of the different structures determined by semantics and phonology; the most well-known example is *transformational grammarian*, semantically bracketed as ((*transformational grammar*) *ian*), morphologically bracketed as (*transformational (*grammar ian)).
- Note, too, that morphological (lexical, semantic oriented) bracketings may not correspond with non-lexical, phonologically motivated bracketings, as in *operate − tion* and *o − pe − ra − tion*; the latter cuts across the former.

6.3.5 Recommendations on types of lexical information

1. Distinguish between the declarative aspect of *types of lexical information* in a lexicon or a lexical database (e.g. orthography, pronunciation, syntactic category) and the procedural aspects of how to use this information in *lexicon production* and *lexical access* of a lexicon or a lexical database.

2. Specify an explicit lexicon model for the structure of lexical entries and the overall structure of the lexicon, both for the lexical database and for the system lexicon.

3. Define the basic *lexical entry types* or *lexical objects* (e.g. fully inflected forms, morphological stems, conceptual units), and the notation for lexical keys.

4. Distinguish clearly between the status of *orthography* as a property of written language, *canonical phonemic representation* as a property of spoken language, and the *lexical key* as a unique identifier for lexical entries (which may often be the orthography of a canonical inflected form, supplemented by a serial number).

5. Define explicit specifications for the notation of orthography and for canonical phonemic representation; for the latter, the SAMPA alphabet is recommended (see the following section, and Appendix A).

6. First specify the lexicon model suitable for the intended applications, then select formalisms appropriate to the model, i.e. representation conventions for the lexical database and the system lexicon.

7. Do not "programme" the lexical information directly in an existing formalism without first explictly specifying the lexicon model.

6.4 Lexical surface information

6.4.1 Orthographic information

Orthography has been used in several different roles in the context of spoken language lexica:

1. Convenient general reference labels for words, due to the high level of awareness of, familiarity with and standardisation of orthography in literate societies.

2. Convenient identifying names for lexical entries, for 'normal lemma' forms, and for headwords in complex lexicon entries for related words.

3. Convenient identifying names for word hypotheses in word lattices, as lexical lookup keys.

4. Visualisation of word hypotheses in a development system.

5. Representation of the orthographic properties of words.

Each of these functions is distinct and needs to be kept conceptually separate in order to avoid confusion. On the side of word forms, the functions (1) and (2) are not particularly problematic. Function (3) is traditionally a feature of speech recognition systems for relatively small vocabularies. The larger the vocabulary, however, the greater the danger of introducing unnecessary *orthographic noise*, i.e. intrusive artefacts due to homography (words with identical spelling and different pronunciation); for this reason, in new architectures, phonological (e.g. phonemic) representation in word graphs is preferred. Function (4) is unproblematic, though similar reservations as with (3) are to be noted. Function (5) is obviously essential for written output of any kind; however, it is often confused with both functions (2) and (3).

Orthography has the advantage of being highly standardised, except for certain regional variants (British and American English; Federal, Swiss, and Austrian German) and variations in publishers' conventions (e.g. British English *ise/-ize* as in *standardisation/standardization*, capitalisation of adjectives in nominal function in German, as *die anderen / die Anderen*, or variations in hyphenation conventions). Orthography is given further attention in the section on lexical representation.

A standard orthographic transliteration is often used for convenience as a means of representing and accessing words in a spoken language lexicon. This has several reasons:

1. Familiarity to all educated speakers of the language.
2. High level of standardisation in comparison with theory-influenced phonological transcriptions.
3. Sufficient proximity to phonological form ensures a close mapping to pronunciation at the level of whole words (not necessarily in the details of grapheme to phoneme mapping) in small vocabularies in some languages: French and English are notorious exceptions.

Most European languages have highly regulated orthographies, the use of which is associated with social and political rewards and punishments. Variation is found particularly in the treatment of derived and compound words (e.g. separation and hyphenation) and in the use of typographic devices such as capitalisation.

For use in spoken language lexica, particularly in word lists used for training and testing recognisers, consistency is essential and often additional conventions are required in order to meet the criterion of general computer readability in the case of special letters and diacritics. Although it cannot be regarded as a standard, it is becomming common practice to use the ASCII codings of LaTeX adaptations for specific countries. For example, a standard computer-readable orthography for German has become widely accepted for German speech recognition applications which marks special characters, in particular those with an *Umlaut* diacritic, as shown in the table.

Standard orthography	Computer readable orthography
Äpfel	`''Apfel`
ändern	`''andern`
Öl	`''Ol`
östlich	`''ostlich`
Überzug	`''Uberzug`
über	`''uber`
heiß	`hei''s`

The results of the EAGLES Working Groups on Text Corpora and Lexica should be consulted on orthographic and other matters pertaining to written texts.

6.4.2 Pronunciation information

Pronunciation information is much more application specific (and indeed theory specific) than orthographic information. Standardly, information about phonemic structure is included in the form of a *phonemic transcription* of a standard *canonical* or *citation form* pronunciation, i.e. the pronunciation of a word in isolation in a standard variety of the language. Often the phonemic transcription is enhanced by including prosodic information such as the stress position (Dutch, English, German), type of tonal accent (Swedish), syllable, word boundaries in compound words, and word and phrase boundaries in phrasal idioms. Morphological information (morph boundaries, as well as the boundaries of words and phrases) is relevant to stress patterning, and is sometimes also included.

A particularly thorny question is the inclusion of information about pronunciation variants. The following rules of thumb can be given:

- Pronunciation lexica for synthesis generally require only one standard (canonical) pronunciation; however, variants of these with different prosodic contexts are often used.
- Pronunciation lexica for recognition require a distinction to be made between variants of the same word, and variants which are associated with the same spelling but different words (heterophonic homographs).
- Strictly speaking, pronunciation lexica for recognition require only variants to be listed which are idiosyncratic and cannot be predicted by rule (e.g. English *either* /aID@/ — /i:D@/). Variants which are general and regular (such as the reduction of schwa + liquid or nasal to a syllabic liquid or nasal) can be calculated using pronunciation rules (phonological rules): English *running* /rVnIN/ — /rVnIn/, German *einem* /aIn@m/ — /aIn=m/ — /aIm/).

Although *phoneme* is a technical term with somewhat different definitions in different theoretical contexts, and although there are technical arguments due to Generative Phonology (cf. Chomsky & Halle (1968)) which show that the notion of phoneme leads to inconsistencies, the core of phoneme theory is relatively standard. In linguistics handbooks, the phoneme is commonly defined as the minimal distinctive (meaning-distinguishing) unit (temporal segment) of sound. In the following fairly standard definition, the distinctiveness criterion is implicit in the concept of a *system*; the concept of a *sound* covers possible variants of a phoneme (e.g. English aspirated word-initial /p/ as opposed to unaspirated /p/ in the context /sp.../ (Crystal (1985: 228)):

> **phoneme (phonemic(s))** The minimal unit in the sound SYSTEM of a LANGUAGE ... Sounds are considered to be members of the same phoneme if they are phonetically similar and do not occur in the same ENVIRONMENT.

A fairly complete definition is thus based on distinctiveness, minimality, phonetic similarity and distributional complementarity. Phoneme definitions are differential or relational definitions, illustrated by the notion of minimal difference between two words in a minimal pair (such as the items in the set of English words *pin–tin–kin–fin–thin–sin–shin–chin–bin–din–gin*, or in standard SAMPA computer readable phonemic transcription: /pIn–tIn–kIn–fIn–TIn–sIn–SIn–tSIn–bIn–dIn–dZIn/). Phonemes defined as in this way are further classified as bundles of distinctive features. Operationally, phonemes are defined by procedures of segmentation and classification (reflected, for example, in the recognition and classification components of automatic speech recognition systems):

- Segmentation is the procedure of isolating the minimal distinctive temporal phonetic segments (*phones*).
- Classification is the procedure of classifying phones as *allophones* (phonetic *alternants* of the same *phoneme*, mainly on the grounds of their phonetic similarity and their complementary distribution (i.e. their occurrence in complementary contexts as contextual variants of that phoneme).

In contrast to orthographic representations, which for social and cultural reasons, are highly standardised common knowledge, direct lexical representations of pronunciation count as highly technical, and are consequently theory and application specific. The most widely used representations in pronouncing dictionaries for human use, foreign language teaching, and in spoken language systems are *phonemic transcriptions.*

Phonemic descriptions are available for several hundred languages, and phonemic transcriptions based on these are suitable for constructing roman orthographies for languages which have orthographies based on different principles or no orthography at all. For a given language, phonemic descriptions differ peripherally (for instance: Are diphthongs and affricates one phoneme or two?). On these grounds, phonemes are in general the units of choice for phonological transcriptions in spoken language system lexica. Other, more specialised types of representation such as feature matrix representations required by all modern phonological descriptions, and even autosegmental directed acyclic graph representations, or metrical tree graph and histogram representations (cf. Goldsmith (1990)) are increasingly finding application in experimental systems (cf. Kornai (1991), Carson-Berndsen (1993), Church (1987a), Church (1987b)) because of their richness and their more direct relation to the acoustic signal. However at the lexical level, they can generally be calculated from the more compact phonemic representations. Because of the widespread use of phonemes, the concept is discussed in more detail below; for fuller explanations, textbooks in phonology should be consulted.

The central question in phonological lexical representation in cases where the notion of phoneme alone is not fully adequate, is that of the level of representation (or level of abstraction). There are three main levels, each of which is an essential part of a full description, and which need to be evaluated for a given application:

Morphophonemic: The morphophonemic level provides a simplification of phonological information with respect to the phonological level; the simplifications utilise knowledge about the morphological structure of words, and permit the use of *morphophonemes*, (a near-synonym is *archiphoneme*) which stand for classes of morphologically and phonologically related phonemes.

There are no standard conventions for the representation of morphophonemes, whether computer readable or not (but see the SAMPA alphabet for French, Appendix A), though capital letters are often used in linguistics publications. Note that this use of capital letters at the morphophonemic level should not be confused with the use of capital letters in the SAMPA alphabet at the phonemic level.

Citations of *morphophonemic* representations are often delimited with brace brackets {...}.

A standard example of a morphophoneme is the final obstruent in languages with final obstruent devoicing, including Dutch and German. For example, the phonemic representation German *Weg* /ve:k/ 'way' — *Wege* /ve:g@/ 'ways' corresponds to a morphophonemic representation {ve:G} — {ve:G+@}, which simplifies the description of the stem of the word. The morphophoneme {G} stands for the phoneme set {/k/, /g/}, and selection of the appropriate member of the set (the appropriate feature specification) is triggered by the morphological boundary and neighbouring phonological segments. Alternatively the morphophoneme may be said to consist of the underspecified feature bundle shared by /k/ and /g/, or more technically, the feature bundle which *subsumes* the feature bundles of /k/ and /g/.

An example from English is the alternation /f/ — /v/, as in *knife* /naIf/ — *knives* /naIvz/, which can be represented morphophonemically as {naIV} — {naIV+z}. The morphophoneme {V} stands for the phoneme set {/f/, /v/}. Here, too, selection of the phoneme (specification of the underspecified subsuming feature bundle) is determined by the morphological boundary and the phonological properties of neighbouring segments.

A corresponding level is necessary for the description of spelling: cf. variations such as English *y-ie* in *city* — *cities*, or German *s-ss-ß* as in *Bus* — *Busse*, *Kuß* — *Küsse* and *Fuß* — *Füße*.

Morphophonemic representations augmented by realisation rules are a powerful *compression technique* for reducing lexicon size:

- Lexica can be stem-based, and thus have fewer entries, and all inflections can be automatically calculated by rule for any stem in the lexicon.
- Morphotactic and morphophonological rules can be used for extending lexica of fully inflected attested forms, and for checking such lexica for consistency.

For requirements such as these, the use of morphophonemic representations, supplemented by morphological construction rules and morphophonemic mapping rules is recommended (cf. Koskenniemi (1983), Karttunen (1983) and Ritchie et al. (1992) for descriptions of standard technologies).

Phonemic: The phonemic level is a standard intermediate level corresponding to criteria outlined in more detail below. The standard European computer readable phonetic alphabet is SAMPA (Appendix A): this alphabet is used for the main languages of the European Union, and is recommended for this purpose. The internationally recognised standard alphabet for phonemic representations is the International Phonetic Alphabet (IPA); this alphabet is shown in Appendix A. One of the main functions of the International Phonetic Association since its inception over 100 years ago has been to coordinate and define standards for this alphabet.

Until relatively recently, the special font used for the IPA has made it difficult to interface with spoken language systems, and for this reason a number of computer-readable encodings of subsets of the IPA have been made for various languages (cf. Allen (1988), Esling (1988), Esling (1990), Jassem & Lobacz (1989), Ball (1991)). The standard computer phonetic alphabet for the main languages of the European Union is the SAMPA alphabet, developed in the ESPRIT SAM and SAM-A projects (cf. Wells (1987), Wells (1989), Wells (1993b), Wells (1993a), Llisterri & Mariño (1993)); see also Appendix A. SAMPA is widely used in European projects, both for corpus transcription and for lexical representations (see also the chapter on Spoken Language Corpora).

However, there is a standard numerical code for IPA symbols (cf. Esling (1988), Esling (1990); Appendix A), and developments in user interfaces with graphical visualisation in recent years are leading to the increasing use of the IPA in its original form, particularly in the speech lab software which is used in spoken language system development.

Citations of *phonemic* representations are standardly delimited by slashes /... /.

Phonetic: At the phonetic level further details of pronunciation, beyond the phonemically minimal, are given. Since the relation between the phonemic and the phonetic level can be described by general rules mapping phonemes to their detailed realisations (allophones) in specific contexts (cf. Woods & Zue (1976)), it is strictly speaking redundant to include these regular variants in a lexicon. However, for reasons of efficiency, detailed word models may be calculated using phonological rules and stored. Essentially this is a software decision: whether to use tables (for efficiency of lookup) or rules (for compactness and generality) for a given purpose.

A specific version of the phonetic level of transcription is *phonotypic* transcription, defined as a mapping from the phonemic level using regular phonological rules of assimilation, deletion, epenthesis (cf. Autesserre, Pérennou & Rossi (1989)); this level is frequently used for generating additional word models to improve speech recognition. Since the amount of phonetic detail which can be processed depends heavily on the vocabulary size and the number of phonological rules which are considered relevant, no general recommendation on this can be given.

There is no widely used standard ASCII encoding of the entire IPA for computer readable phonetic representations and therefore no recommendations can be given on this. Currently, individual laboratories use their own enhancements of phonemic representations. However, the fuller encodings mentioned in connection with the phonemic level of transcription are eminently suitable for interface purpose at the phonetic level, and will no doubt be increasingly used where more detailed phonetic information is required.

Citations of phonetic forms are standardly delimited by square brackets [...].

The chapter on Spoken Language Corpora should also be consulted in respect of levels and types of representation.

6.4.3 Prosodic information

The area of word prosody, and, more generally, the description of other prosodic units which have quasi-morphemic functions, is gradually emerging as an important area for spoken language lexica. For present purposes, prosodic properties are defined as properties of word forms which are larger than phonemes; further specification in phonetic terms (e.g. F0 patterning) and in semantic terms (e.g. attitudinal meaning) may also be given but are not essential for present purposes.

One type of lexical information on prosody pertains to phonological or morphological properties of words, such as Swedish pitch accents, or stress positions in words (for German cf. Bleiching (1992)). Some aspects of word prosody are predictable on the basis of the regular phonological and morphological structure of words, but some are idiosyncratic. Examples in English where word stress is significant include the noun-verb alternation type as in *export* - /"EkspOt/ (Noun), /Eksp"Ot/ (Verb). In German, word stress is significant for instance in distinguishing between compound separable particle verbs and derived inseparable prefixed verbs as in *"ubersetzen* - /"y:b6z

It has been shown (cf. Waibel (1988)) that taking word prosody into account in English can produce a significant improvement in recognition rate.

In addition, there is lexical information associated with prosodic units which occur independently of particular words, and therefore may themselves be regarded as lexical signs and be inventarised in a prosodic lexicon (cf. Aubergé (1992)). To give a highly simplified example in a basic attribute-value notation, a prosodic lexicon for an intonation language might have the following structure.

```
Terminal_1:
  <phonetics pitch> = fall
  <semantics>       = statement or instruction.

Terminal_2:
  <phonetics pitch> = rise
  <semantics>       = question or polite instruction.
```

This kind of information, in which prosodic categories function as a kind of *morpheme* with an identifiable meaning, is generally not regarded as lexical information, but treated as a separate layer of organisation in language. Intonation is being taken increasingly into account for *prosodic parsing* in two main senses of the term:

1. Analysis of speech signal in respect of the *fundamental frequency* (*F0*, *F-zero* trajectory, for speech recognition.

2. Analysis of sentence structure for the generation of intonation patterns in speech synthesis.

Prosodic representation in the lexicon is in general restricted to the prosodic properties of words, such as stress position in English, Dutch, and German words, or tonal accent in Swedish words, or to rhythmically relevant structures such as the syllable and the foot. For spoken language processing purposes where prosody plays a role, it is also necessary to include an inventory of prosodic forms, and their meanings, which play a role at the sentence level, independently of specific words: i.e. a prosodic lexicon.

The IPA defines means of representing both types of information, and a subset (for word prosody) has been encoded in the SAMPA alphabet. However, the state of knowledge in the area of prosody is less stable than in the area of segmental word structure, and a range of different conventions are available (cf. Bruce (1989)); in this area, there are even SAMPA 'dialects', for instance replacing SAMPA " and % for primary and secondary stress by the more iconic ' (single quote) and " (two single quotes) or " (double quote).

For American English, the ToBI (Tones and Break Indices) transcription has been extensively applied, but this system is not immediately applicable to other languages, though experimental applications and adaptations are being developed; see also the chapter on Spoken Language Corpora).

In more theoretically oriented spoken language lexicography within the VERBMOBIL project, attribute-based formal representations of prosodic features in the lexicon have been developed using the ILEX (Integrated Lexicon) model and the lexical knowledge representation language DATR (cf. Bleiching (1992), Gibbon (1991)).

There is an increasing tendency no longer to regard prosodic representations as totally exotic and quite unlike anything else, and a degree of convergence can be observed. But there is still insufficient consensus on lexical prosodic features to permit generally valid recommendations to be made for prosodic representations in the lexicon. For most purposes, plain SAMPA-style symbols will be adequate, though an extended set of conventions is available in the SAMPROSA alphabet (see Appendix). For covering new ground with extended lexica for use with discourse phenomena at the dialogue level, a lexical knowledge representation language with a more general notation, as illustrated above, may be more appropriate.

6.4.4 Recommendations on lexical surface information

1. Define the basic lexical entry type, as recommended in the previous section, and its notation.
2. Establish a machine-readable orthography convention.
3. Define the phonological level of representation (*morphophonemic*, *canonical phonemic*).
4. Specify requirements in respect of word prosody representation.
5. Specify requirements in respect of phrasal and discourse prosody (intonation) information.
6. Use the European standard machine readable phonemic alphabet SAMPA (See Appendix).
7. Select an appropriate prosodic representation, such as SAMPA, SAMPROSA, or ToBI notation, where required.
8. Ensure that the relation between notations and representations used in the *lexical database* and the *system lexicon* are well-defined, and that they are completely consistent with notations and representations in other resources such as *corpora* and in the different parts of the system, such as the *word lattice* produced by the speech recogniser, operated on by a *stochastic language model*, and further processed by a sentence parser.

6.5 Morphological information

6.5.1 Types of morphological information

Morphological information is information about semantically relevant word structure; the smallest morphological unit is the *morpheme*, often defined as the *smallest meaningful unit* in a language.

Morphemes should not be confused with units such as the *syllable* and its constituents, which are used for describing the structure of words from a pronunciation point of view, without reference to meaning.

The domain of morphology may be divided in terms of the *functions* of morphological operations, i.e. *agreement* vs. *word formation*, or in terms of the *structures* defined by morphological operations, i.e. *affixation*, (*prefixation*, *suffixation*, *infixation* or *prosodic modification*) vs. compounding (*concatenation* of *stems* or *words*). These two dimensions can be represented

	Agreement	Word formation
Affixation	INFLECTION	DERIVATION
Stem/word concatenation		COMPOSITION

There is an apparent gap in the use of stem or word concatenation for agreement purposes; however, so-called *periphrastic constructions*, typically with auxiliary verbs and participles or infinitives, may be assigned to this slot, though these constructions cannot be compared directly with standard compounds: compare English *John will come* with French *Jean viendra*. English lacks an inflectional future, but has periphrastic modal or infinitive complement future forms such as *John will come tomorrow*, *John is going to come tomorrow*, as well as the present tense as a general or neutral tense form, as in *John comes tomorrow*.

There are other intermediate cases which sometimes present difficulties in classification and where the solution is not always immediately obvious.

- Are the degrees of comparison of adjectives *inflections* or *derivations* (i.e. positive *loud*, comparative *louder*, superlative *loudest*)?
- Are participles *inflected forms* of verbs or *derivations* as deverbal adjectives?
- Are infinitives of verbs *inflected forms* or *derivations* (cf. *They want an answer* vs. *They want to know*; *To be or not to be He begged to come*, *He decided to come*)?

Traditional treatments often treat these forms together with inflections, presumably because of their regularity and the involvement of suffixation. They are generally better treated as derivations, however, because they have different syntactic distributions from other inflections of the same stems, and may be additionally inflected as adjectives or nouns (cf. the perfect participle in French: *On l'a vue*, where it can be deduced from the feminine inflection on the participle and the rules of *inflectional agreement* (*inflectional congruence*) that *l'* refers to a *feminine* item.

6.5.2 Applications of morphology

Morphological structuring is required for the following tasks:

- The treatment of large vocabularies for speech recognition and synthesis by means of rule-based assignment of inflected forms to stems.
- The prediction of new (unattested, unknown) words for speech recognition on the basis of known principles of word composition, and known attested parts of words.
- Rule-based assignment of stress patterns.
- Word recognition by stem spotting.
- Construction of sub-word language models for speech recognition.

There are two main ways of structuring words internally into word sub-units (word constituents):

1. *Semantic orientation.* On morphological grounds, word forms may be decomposed into smaller meaningful units, the smallest of which are morphs, the phonological forms of morphemes; an intermediate unit between the morph and the word form is the stem.

2. *Phonological orientation.* On phonological grounds, word forms may be decomposed into smaller pronunciation units, the smallest of which are phonemes; an intermediate pronunciation unit is the syllable.

It is important to note that *decomposition into syllables is not isomorphic with decomposition into morphs.* For example, *phonological* has the syllable structure /fO . n@ . lO . d3I . k@l/ and the morph structure /fOn + @ + lOd3Ik + @l/, which are quite different from each other.

In addition to phonological decomposition, in the written mdoe word forms may be decomposed into smaller spelling units, graphemes, each consisting of one or more characters; an intermediate orthographic unit is the orthographic break (orthographic syllable), which is in general only needed for line-breaks and does not correspond closely to either syllable or morph boundaries but combine phonological, morphological and orthographic criteria.

For the core requirements of speech recognition, in which a closed vocabulary of attested fully inflected word forms are generally used, morphological structuring is not necessary. Phonological structuring into syllables, demisyllables, diphone sequences or phonemes is widely used in order to increase statistical coverage and to capture details of pronunciation (cf. Browman (1980), Ruske & Schotola (1981), Ruske (1985)).

In many languages, syllables and morphs do not always coincide; morphs may be smaller than or larger than syllables.

A brief outline of the main concepts in morphology, as they affect spoken language lexica will be useful in developing spoken language lexica; for more detail a textbook in linguistics should be consulted (e.g. Akmajian (1984)).

Morphology: Morphology is the definition of the composition of words as a function of the meaning, syntactic function, and phonological or orthographic form of their parts. The morphology of spoken language is fundamentally the same as the morphology of written language in respect of meaning, syntactic function, and the combinability of *morphemes*. It differs in respect of morphophonological alternations, which differ from spelling alternations, and word prosody (for instance word stress patterns). General definitions are given here; examples are given below.

Morphotactics (word syntax) is the definition of the composition of words as a function of the forms of their parts.

Inflection is that part of morphology which deals with the adaptation of words to their contexts within sentences.

Word formation is that part of morphology which deals with the construction of words from smaller meaningful parts.

Derivation is that part of word formation which deals with the construction of words by the concatenation of stems with affixes.

Compounding (composition) is that part of word formation which deals with the construction of words by concatenating words or stems.

Simple morphological units: Traditional terminology varies in this area; a standard but incomplete definition of a *morpheme*, for instance, is that it is 'the minimal meaning-bearing unit of a language'. This definition is not entirely satisfactory, however, and for present purposes the sign-based model and the unit of *word* will be used as the starting point.

A *morpheme* is the smallest abstract sign-structured component of a word, and is assigned representations of its meaning, distribution and surface (orthographic and phonological) properties. More informally, morphemes are parts of words defined by criteria of form, distribution and meaning; i.e. they have meanings and are realised by orthographic or phonological forms (morphs).

Traditionally, the two main kinds of morpheme are:

- *Lexical morphemes*, characterised by membership of a large, potentially open class, with meanings such as properties and roles of objects, states and events.
- *Grammatical morphemes*, characterised by membership of a closed class, defined by their distribution with respect to larger units such as sentences or complex words (e.g. inflectional and derivational endings; function words such as prepositions, articles).

Morphs are, in traditional linguistics, the orthographic and phonological forms (realisations) of morphemes. *Orthographic morphs* consist of graphemes (either single letters or fixed combinations of letters); in traditional phonology, *phonological morphs* consist of phoneme sequences with a prosodic pattern (e.g. word stress).

Roots (lexical morphs) are the morphs which realise lexical morphemes and inflectable grammatical morphemes, and function as the smallest type of *stem* in *derivation* and *compounding. Affixes (prefixes, suffixes)* are morphs which realise the inflectional and derivational beginnings and endings of words.

A *free morph* is a morph which can occur on its own with no affixes as a separate word; a *bound morph* is a morph (generally an affix) which always occurs together with at least one other morph (typically a *stem* in the same word.

Complex morphological units: The structure of words is, like the structure of sentences, defined recursively, since the vocabulary of a language (including new coinages) is potentially unlimited. The functional and formal classification of morphological word structure (compounding and derivation, see above) takes this into account. Where words which are not in a given lexicon are likely to be encountered, *morphotactic rules* and a *morphological parser* or *morphological generator* may be required in order to supplement the lexicon. The condition of recursive structure does not apply to inflection, which, given a finite set of stems, defines a finite set of words (though in agglutinative languages, an extremely large finite set):

Inflectional affixation: A word (fully inflected word) is a stem morphologically concatenated with a full set of inflectional affixes, e.g. English *algorithm* + *s* = *algorithms* or German *ge* + *segn* + *et* + *en* 'blessed' (plural participle or adjective).

Derivational affixation: A stem is

- either a root (i.e. lexical morph), e.g. *tree*, *algorithm*
- or a stem morphologically concatenated with a derivational affix, e.g. *algorithm* + *ic*, *algorithm* + *ic* + *al* + *ly*, *non* + *algorithm* + *ic* + *al* + *ly*, etc.

Compounding: A compound word is a word morphologically concatenated with a word.

Morphophonological and orthographic alternations: The operation of *morphological concatenation* is defined for present purposes as 'concatenation and modification of segments at morph boundaries by boundary phenomena.' The details of pronunciation and spelling are altered in morphologically complex items. An example of morphophonological alternation is /f/ — /v/ in *knife* /naIf/ — /naIvz/. An example of orthographic alternation is *y* — *i* — *ie* in *fly*, *flier*, *flies*. These alternants can be described by rules:

1. Morphophonological rules are rules (analogous to spelling rules) which describe morphophonological alternations, i.e. the differences between pronunciations of parts of composite words and pronunciations of corresponding parts of simplex words.
2. Spelling rules are rules which describe spelling alternations, i.e. the differences between spellings of parts of composite words and the spellings of corresponding parts of simplex words.

A standard technology for formulating spelling rules and morphophonological rules is *Two-Level Morphology* (cf. Koskenniemi (1983), Karttunen (1983); cf. Ritchie et al. (1992)).

6.5.3 Recommendations on morphology

1. Decide whether word sub-units may have a role to play in the intended application.

2. For a large vocabulary (¿ 5000 words) spoken language lexical database of a highly inflecting language as a general resource, consider using a *morphology component* will be needed to generate fully inflected forms, either on demand or as a (very large) table.

3. For specific problem areas such as the identification of *new words* or increasing the robustness of a recogniser, consider using morphological units in speech recognition.

4. In highly inflecting languages, consider the use of stochastic language models based on word stems as an alternative to fully inflected words.

6.6 Grammatical information

6.6.1 Statistical language models

Language models are a major area of research and development, and crucial for the success of a speech recognition system. The chapter on language modelling should also be consulted on this topic.

In speech recognition systems, the mapping of the digitised acoustic forms of words on to symbolic representations for use as lexical lookup keys is performed in primarily statistical approaches by pattern recognition, and in primarily knowledge-based approaches by phonemic rule systems. The *word models* used for matching with the acoustic analysis of the speech signal are accordingly very different. The literature can be consulted for details of current practice (cf. Waibel & Lee (1990)).

In written language processing, a comparable task is Optical Character Recognition (OCR), and in particular, handwriting recognition; there is no comparable task in conventional Natural Language Processing or Computational Linguistics, where letters are uniquely identified by digital codes, and dictionary access may be trivially optimised by encoding letter sequences into tries (letter trees, letter-based decision trees). However, in linguistic terms, in each case the task is the identification of word forms as lexical keys.

Current statistically oriented technology is based on dynamic programming, on Hidden Markov Models, or on Neural Networks. At the present time, Hidden Markov Models are fairly generally considered to provide the highest recognition and accuracy rates. For the restricted tasks which dominate much previous and current work on Spoken Language Input Systems, HMM technology is consequently recommended.

Statistical language models are used to reduce the search space for the correct hypotheses in the word lattice output of a speech recogniser (cf. Oerder & Ney (1993)). They provide top-down (a priori) information on the distribution of words relative to immediately preceding (strings of) words, and may be *n-gram models* (sequences of n words, including the target word), stochastic grammars, or probabilistic networks and automata. Since statistical language models are based on examination of a finite corpus, they also presuppose a finite vocabulary. Hence, the statistical distribution properties expressed by a statistical language model constitute a form of probabilistically weighted syntactic information about lexical items and may be considered a highly specialised form of spoken language lexicon. Statistical language models for spoken language are corpus-specific, and it is hard to consider them as general resources in view of the fairly small size of the corpora. For this reason, they are best defined in terms of the standard principles on which they are based, with references to the basic components of a language model:

1. Input data formats for the construction of statistical language models (e.g. with fixed and consistent transcription conventions).

2. An output data structure for the language model (a table, tree, network, etc.).

3. A construction function mapping the input data to a statistical language model, which can be digram table construction with UNIX tools or more sophisticated procedures for the construction of tree, network or other database structures.

4. A set of access functions for utilising the language model in a spoken language recognition system, yielding such values as statistical properties of words (for instance the negative logarithm of the conditional probability) or the perplexity of the model (its average branching factor, when interpreted as a transition network).

The simplest n-gram models are based on the following lexicographic concepts.

- A *word list* is an ordered list of word form *types* (each represented by one token only) extracted from the possibly numerous occurrences of word form *tokens* in a corpus.

- A *word frequency list* is a word list in which each type is paired with the number of occurrences of corresponding tokens in a given corpus.

- The *type/token ratio* and the rate of its increase with the size of a corpus is a mark of the representativity of a corpus.

The definition of type is not trivial, and in general it is not very explicitly defined. In practice, this means that frequencies of words in a given training corpus need to be tabulated; these frequencies are divided by the corpus size (in word tokens) for conversion into probability estimates.

Often an orthographic form is taken to represent a type, even though this may introduce loss of information through *orthographic noise*, introducing artefactual ambiguities in the case of homographs. A better procedure, though more cumbersome in practice, is to treat a lexical entry for the purposes of a word list as a pair of orthographic and pronunciation types.

The ranking of hypotheses in speech recognition (indirectly: a form of disambiguation of the speech input) calls for the calculation of both a posteriori (currently observed) and a priori (previously observed and generalised) probabilities in accordance with Bayes' Law:

> The probability of a word, given a set of observations, is the probability of the observations, given the word, multiplied by the a priori probability of the word, and divided by the probability of the observations. Since the last factor is common to all observations for a given system and scenario, it can be factored out:

```
p(W|O) = P(O|W) * P(W) / P(O)
```

More specific descriptions are given in the literature such as Bahl et al. (1989).

In order to evaluate results, analysis of a corpus is required. Common practice in development is to partition a corpus into 90% training data and 10% test data, and to run tests on, for instance, 10 different partitionings in order to ensure homogeneity. On the question of evaluation, the results of the EAGLES Working Group on Evaluation should be consulted.

More generally, statistical language models are used in order to enhance the a priori predictions and thus enhance the hypothesis selection process. Statistical language models are also in a sense word lists, but lists in which the types are themselves paired with the contexts in which the tokens occur and with the frequencies of these relative occurrences. If the context is one preceding word, the result is a *digram table* (*bigram table*). This technique can be used with any size of linguistic unit, for example with phonemes, from which diphone or diphoneme tables are constructed. If the context is more than one preceding word, the result is a *trigram* (*triphone* etc.) table. In the general case, a distribution table of this kind is an *n-gram*, *n-phone* etc. table, where n is the size of the context including the target unit itself.

A basic constraint on frequency information and probability estimates in the lexicon is the inverse relation between size of unit and number of occurrences: clearly, smaller units such as phonemes will occur more frequently in a corpus than words, words will occur more frequently than sentences, digrams will occur more frequently than trigrams, and so on. The larger the unit, the greater becomes the *sparse data* problem, i.e. the problem of having too few occurrences (in particular, zero occurrences) of a given unit to be statistically significant.

The basic item of syntactic lexical information required in spoken language systems is the *part of speech* (POS, see below). The information is required in order to enhance language models in the case of sparse data; classes of items judged to be similar in distribution occur more frequently than their members; thus, *di-class* (in the general case, *n-class*) statistics hold promise of better statistical generalisations than word-based digram statistics. Experiments with semantic classes are also being made (cf. Yarowsky (1991)). For further information, standard texts such as Waibel & Lee (1990), Sagerer (1990), Haton et al. (1991) should be consulted.

In respect of general recommendations, the chapter on language modelling should be consulted. It is hard to make a general recommendation for future developments from a lexicographic point of view, in view of much current basic research work on the equivalence of HMMs with certain types of Neural Networks and probabilistically weighted knowledge-based systems based on finite state automata (cf. Kornai (1991), Carson-Berndsen (1993)).

6.6.2 Sentence syntax information

Syntactic information is required not only for parsing into syntactic structures for further semantic processing in a speech understanding system, but also in order to control the assignment of prosodic information to sentences in *prosodic parsing* and *prosodic synthesis*.

Syntactic information is defined as information about the distribution of a word in syntactic structures. This is a very common, indeed 'classical', but specialised use of the words 'syntax' and 'syntactic' to pertain to phrasal syntax, i.e. the structure of sentences. Other more general uses of the terms for linguistic units which are larger or smaller than sentences are increasingly encountered, such as 'dialogue syntax', 'word syntax' (for morphotactics within morphology).

Within this classical usage, the term *syntax* is sometimes opposed to the term *lexicon*; the term *grammar* is sometimes used to mean syntax, but sometimes includes both phrasal syntax and the lexicon.

Strictly speaking, a statistical language model is a form of sentence syntax defined for a finite language, since it also defines the distribution of words in syntactic structures; however, the notion of syntactic structure here is very elementary, consisting of a short substring or window within a string of words. It is also used with quite a different function from the classical combination of sentence syntax and sentence parser.

Sentence syntax defines the structure of a (generally unlimited) set of *sentences*. Syntactic lexical information is traditionally divided into information about *paradigmatic* (classificatory; disjunctive; element-class, subclass-superclass) and *syntagmatic* (compositional; conjunctive; part-whole) relations. The informal definitions of these terms in linguistics textbooks are often unclear, metaphorical and inconsistent. For instance, temporally parallel information about the constitution of phonemes in terms of distinctive features is sometimes regarded as paradigmatic (since features may be seen as intensional characterisations of a class of phonemes) and sometimes as syntagmatic (since the phonetic events corresponding to features occur together to constitute a phoneme as a larger whole). The relation here is analogous to the relation between intonation and sentences, which are also temporally parallel, and in fact treated in an identical fashion in contemporary computational phonology. From a formal point of view, this is purely a matter of perspective: the internal structure of a unit (syntagmatic relations between parts of the unit) may be seen as a property of the unit (paradigmatic relation of similarity between the whole unit and other units). In lexical knowledge bases for spoken language systems it is crucial to keep questions of syntagmatic distribution and questions of paradigmatic similarity apart as two distinct and complementary aspects of structure.

The *part of speech* (*POS*, *word class*, or *category*) is the most elementary type of syntactic information. One traditional set of word classes consists of the following: Noun or Substantive, Pronoun, Verb, Adverb, Adjective, Article, Preposition, Conjunction, Interjection. Two main groups are generally identified:

1. *Lexical categories* are the open classes which may be extended by word formation: Noun, Verb, Adjective, Adverb.

2. *Grammatical categories* are the closed classes which express syntactic and indexical relations: Pronoun and Article (anaphoric and deictic relations), Preposition (spatial, temporal, personal relations etc.), Conjunction (propositional relations), Interjection (dialogue relations).

The granularity of classification can be reduced by grouping classes together in this way (this particular binary division is relevant for defining stress patterns, for example), or increased by defining subclasses, depending on particular requirements. For further information, introductory texts on syntax, e.g. Sells (1985) or Radford (1988) may be consulted.

6.6.3 Recommendations on grammatical information

1. Consult the chapter on language models for further details on the application of language models in speech recognition.

2. Consult the chapter on speech synthesis for further details on grammar in the production of speech.

3. Consult the results of the EAGLES Working Groups on Formalisms and on Computational Lexica for additional information about the use of grammatical information in sentence parsing.

4. Ensure that a thorough grammatical analysis is performed on the application-relevant data in order to ensure that no more grammatical complexity is introduced than is actually needed.

6.7 Lexical content information

6.7.1 Lexical semantic information

The treatment of semantics in lexica for spoken language systems is far less uniform than the treatment of word forms; decisions in this area are highly application specific:

- In single word control systems, a direct mapping to an operation or action is sufficient.
- In a dictation system, a mapping to text is needed, with or without deeper analysis.
- Conversely, in a text-to-speech system, a full semantic analysis is also not generally necessary.
- In a database front end application, a translation into a database query language is required.
- In a spoken language understanding system a sophisticated system of semantic representation such as artificial intelligence representation languages (semantic networks or frame structures), or more general formalisms such as predicate calculus or attribute logics are required.

An outline of some of the basic lexical semantic relations between lexical units was given at the beginning of this section in the context of the discussion of relations between lexical units (lexical relations).

Reference should be made to the results of the EAGLES Computational Lexica Working Group for further information on lexical semantics, and to standard texts such as Lyons (1977) or Cruse (1986).

6.7.2 Pragmatic information

The central area of application for pragmatic information from the point of view of spoken language is in dialogue situations, in which prosody (intonation, emphasis, accentuation) is required in addition to word-based information: prosodic information is typically associated with speaker-centred pragmatic information such as topic focussing, speaker attitudes, and dialogue turn-taking. For further details of dialogue structure, the chapter on Interactive Dialogue Systems and the results of the EAGLES Working Group on Computational Lexica should be consulted.

Because spoken language corpora are generally highly application oriented, and therefore bound to a particular speaker or set of speakers, with relatively homogeneous properties of register, speech style, and dialect, including pronunciation, vocabulary, grammar, and intentions with respect to specific actions, the kind of pragmatic information required for particular lexical items is restricted, as the same information applies to all lexical items.

Spoken language lexica differ in this way from large-scale general coverage lexica, though the need for such lexica in the spoken language area is growing. The kinds of pragmatic information required are generally limited to information about a few speech act types (question, answer, instruction, etc.). The advent of spoken language dialogue systems, however, is making more sophisticated approaches necessary: treatment of discourse particles, including hesitation markers, and of word fragments is becoming necessary.

6.7.3 Idiomatic information

Pragmatic idioms such as greetings are generally treated as holistic lexical items, i.e. as 'canned text', and included in the lexicon in full. The same applies to fixed idioms such as *Come to think of* *it, ...* in the meaning 'I just thought of another relevant point, namely ...'; variants like *I will just come to think of it, ...*, *Come to consider it, ...* etc. do not have the idiomatic meaning.

The most complex problems arise in the case of idioms with variant forms, such as *If you twist my arm, then ...* , i.e. 'If you give me a really good reason, then ...', where forms such as *Twist my arm!* or even *would you mind twisting Fred's arm?* have the idiomatic meaning, *Don't twist my arm!*, *It was my arm that Tony twisted* etc. tend to have the literal, rather than the idiomatic meaning.

Very often, idioms are associated with a specific range of prosodic patterns (intonation patterns; for example, in *How do you do?*, the words *how*, *do* and *you* might be emphasised as a joke, but not with the standard pragmatic idiom connotation. In the example given above, if the word *twist* is given a noticeably stronger accent than *arm*, the same also applies.

¿From a more general perspective, so-called *functional units* (sequences of functional words which behave as a phonological unit) and *clitics* (functional words which combine with lexical words to form a sequence which behaves as a phonological unit), share a number of properties with idioms in the more traditional sense of the term. Characteristic of these units is that they have special phonological characteristics, with deletion and assimilation of segments to their neighbouring environment.

An example of a functional unit in English is *I c'n ...* /aIkN/ for *I can* /aI k{n/ in informal, fast speech or particularly unstressed contexts.

An example of a clitic in English is he's /hi:z/ or even /hIz/, for *he is* /hi: Iz/.

Some cliticised sequences have become *lexicalised* (i.e. independent lexical items) in informal styles, e.g. *can't* /ka:nt/ and *cannot* /k{nOt/ for *can not* /k{n nOt/.

A general solution to the lexical treatment of idioms is not currently available either for written language processing or spoken language processing, and further research and development is needed in this area in view of the frequency of idioms in actual corpora.

6.7.4 Recommendations on semantic information

1. Ensure that the application domain is as precisely specified and modelled as possible.
2. Determine the size of vocabulary which can be processed, and consider ways in which to minimise the vocabulary needed for this domain.
3. Consult the results of the EAGLES Working Group on Computational Lexica; this is an area in which Spoken Language and Written Language processing overlap considerably.

6.8 Lexicon structure

The area of *lexicon structure* deals with the organisation of lexical information in lexica. Models for lexical information, and types of lexical information, are dealt with in the preceding sections. Terminology varies considerably in this area. The structure of a spoken language lexicon may be seen from the following points of view:

Lexical formalisms, lexicon representation languages: Representation conventions of various types (symbolic notations, programming languages, database languages, logical formalisms, purpose-designed knowledge representation languages), which are suitable for formulating *lexical models.*

Lexicon architecture: The choice of basic objects and properties in the lexicon, and the structure of the lexicon as a whole, such as a table of items, a trie (decision tree), an inheritance hierarchy, a semantic network, a database.

6.8.1 Spoken language lexicon formalisms

Spoken language lexicon formalisms (representation languages) may be broadly classified according to their use:

1. Linguistically and phonetically based working notations.
2. Implementation languages for the operational phase.
3. Algebraic and logical formalisms for formal definition.

In Speech Technology contexts, often no distinction is made in practice. Where an ad hoc solution is required and where lexicon structure is simple, a lexicon may be written directly in a standard programming language suitable for high-speed runtime applications, traditionally Fortran but more recently C, or in a higher level language such as LISP or Prolog. Recent developments are moving towards knowledge representation languages which are specifically designed to meet all three of the above criteria equally well, in that they are useful working notations, have efficient implementations, and are formally well-defined.

A variety of lexical formalisms are currently used in Spoken Language lexica; some of these are also used for general written language lexica. A more detailed classification of formal representation systems may be given as follows:

1. General data structures (lists, tables or matrices, tree structures designed for optimal lexical access).
2. Programming languages (C for efficiency; LISP or Prolog for flexibility).
3. Database systems.
4. General text markup languages such as SGML.
5. Knowledge representation languages (inheritance networks, semantic networks, frame systems).
6. Linguistic knowledge representation languages, commonly based on attribute-value logics.
7. Lexical knowledge representation languages (attribute based inheritance formalisms) such as DATR.

General data structures definitions are required for specifying the general processing properties of a system, and are relevant for developers and for theoretical work on the complexity and efficiency of lexica and lexicon processing; standard textbooks on data structures and algorithms should be consulted for this purpose.

Conventional programming languages are generally used for performance reasons in runtime systems. They may also be used to implement small or simple lexica directly, in particular for rapid prototyping of these; this is not optimal software development practice, however, and not to be recommended for developing large or complex lexica, in particular those with highly structured linguistic information.

Database management systems (DBMSs) are widely used for general resource management, including large-scale lexica with rich information which needs to be accessed flexibly and efficiently. In the SAM project, an ORACLE database management concept for spoken language corpora and lexica was developed (cf. Dolmazon, Caërou & Barry (1990)).

General text markup languages are used for integration with large, pre-analysed written corpora in the development of natural language processing systems and in statistical basic research in computational linguistics, and so far are not used in spoken language system development (cf. the results of the EAGLES Working Group on Text Corpora). Implementations of SGML are readily available.

Knowledge representation languages (KRLs) are used for developing complex semantic and conceptual knowledge representations, and for integrating spoken language front ends with knowledge based systems; see Schröder, Sagerer & Niemann (1987), Sagerer (1990); more generally, cf. Bobrow & Winograd (1977), Brachman & Levesque (1985), Charniak & McDermott (1985), De Mori et al. (1984), Young et al. (1989).

Linguistic formalisms in general are discussed in the results of the EAGLES Working Group on Grammar Formalisms, which should be referred to in this connection.

Lexical knowledge representation languages (LKRLs) are a relatively new development. They are coming to be used in knowledge acquisition for integrated lexica which contain a variety of complex lexical information from phonology through morphology and syntax to semantics and pragmatics. They provide a means of bridging the gap between complexity of lexical information and easy-to-read representations using sign-based lexicon models. A LKRL which has been used in several natural language processing and language and speech projects is DATR (cf. Evans & Gazdar (1989), Cahill (1993), Cahill & Evans (1990), Andry et al. (1992), Gibbon (1991), Gibbon (1993), Bleiching (1992), Langer & Gibbon (1992)). This is the language whose syntax conventions for basic attribute-value representations is used in this chapter. A number of public domain DATR implementations are available.

6.8.2 Lexicon architecture and lexical database structure

Lexicon architecture pertains to the *choice of basic objects and properties* in the lexicon, and to the *overall structure* of the lexicon. More formally, it defines the relation which assigns lexical properties to lexical entries.

The term "architecture" generally refers to the structure of system lexica, but the term is also justified in connection with lexical database structure, particularly when more complex relational or object-oriented structures are concerned.

The basic objects in terms of which an architecture may be defined were discussed in the section on lexical information for spoken language.

The overall structure of a spoken language lexicon is determined by a wide range of declarative, procedural and operational criteria such as the following:

- The complexity of the information assigned to lexical entries.
- The complexity of the relations defined between lexical entries.
- The particular subset of objects and properties defined for a given application lexicon.
- Linguistic and logical compression techniques such as redundancy rules or, more generally, inheritance hierarchies.
- Task driven directionality of access.
- Variety of information required for access (from phonological to pragmatic).
- Performance requirements of software (including lingware) size and speed of access.
- Techniques of acquisition and maintenance (with respect, for instance, consistency).

At the one extreme is the ideal notion of a fully integrated sign-based model with non-redundant specification of entries and property inheritance; in between is the efficient database management system used for large scale lexical databases, and at the other extreme is the simple pronunciation table which is the starting point for the training of speech recognition devices.

The choice of lexicon architecture on the basis of parameters such as those listed above, and taking into account practical constraints from the actual working environment, is application specific. There is no single principle of organisation which applies to all lexica.

The closest approximation to a neutral form of spoken language lexicon organisation is a sign-based general background lexicon organised as a database with flexible access. Such a lexicon is basically knowledge acquisition oriented, and can function as a source for the specialised lexica required for different speech synthesis and recognition applications. Specialised models for sublexica which are optimised for particular applications can then be formulated, and sublexica can be automatically compiled out of the main lexicon into notations and structures.

The organisation of a lexicon determines the general properties of the formalism to be used with the lexicon. Conversely, available formalisms determine tractable forms of lexicon organisation in terms of data structures, algorithms and programming environments (cf. Knuth (1973), Carbonell & Pierrel (1986), Rudnicky et al. (1987), Lacouture & Normandin (1993)). Object-oriented system architectures, with local encapsulation of all aspects of representation and processing, permit the construction of hybrid systems with functionally optimised components; by analogy, the lexicon itself can be conceived as a hybrid object system if required.

This is in effect the situation in current speech recognition technology, in which a more or less large set of HMMs representing words, for instance, can be seen as a procedurally sophisticated lexicon with acoustically driven lookup of keys which are then used to access the main lexicon. Although the standard perspective is to see the two components as separate, they can be seen as objects which are both located in an overall hybrid spoken language system lexicon.

Current research on new object-oriented interactive incremental spoken language system architectures raises many questions about the role of a lexicon in such a system. One major question is whether the lexicon is an object (or system of objects) in its own right, or whether the lexicon is distributed over the system components and is thus a *virtual lexicon*. Questions such as this are very much within the domain of ongoing basic research, and it would be premature to make specific recommendations at this point.

For a broader discussion of lexicon architectures, the work of the EAGLES Working Group on Computational Lexica should be consulted. Further discussion here will be restricted to the elementary case of pronunciation tables.

6.8.3 The architecture of spoken language system lexica

It has already been pointed out that at the current state of the art, spoken language systems tend to deal with the properties of words in a distributed but modular fashion. The following components of a speech recognition system, for example, may be said to constitute a lexicon in the strict sense of the term:

Word recognition as lexical key identification: The task of identifying lexical keys in written language is straightforward, once the characters have been identified by some analogue input device such as a keyboard or a scanner with OCR (Optical Character Recognition) software and converted to discrete symbolic strings which act as identifiers for lexical access. The strings may be given an optimised representation for the purpose of lexical lookup, for example as a letter discrimination tree, which may be determiniatically processed, i.e. without any search in the general sense. Traditionally, this aspect of written language is not in the centre of attention. In spoken language word recognition, the situation is quite different: the main emphasis is on the process of converting from a continuous analogue signal to a set of symbolic strings, which represent hypotheses about lexicon keys, i.e. identifiers for lexical access; the mapping to a set means a non-deterministic situation which needs to be handled by sophisticated search techniques in order to find the best hypothesis. The techniques, such as signal transformations and Hidden Markov Models, which perform the identification of lexical keys, are, in the strict sense of the term, just as much a part of spoken language lexicography as ASCII codes and letter trees are a part of written language lexicography.

Stochastic language models as syntactic category and subcategory information: The search process for identifying the best lexical key uses top-down knowledge about the distribution of words in sentence contexts. Distributional properties of words range from general linguistic categories such as *Verb*, *Ditransitive Verb*, etc., to the immediate string context consisting of n words, e.g. *bigrams*, *trigrams*. This information is not seen in speech technology as being lexical information; from a general lexicographic point of view, however, distributional information about is lexical information, and a repository of such information is part of a lexicon.

Domain information: The application-specific and, in particular, domain-specific vocabulary and grammar of a speech recogniser is used to restrict search further. In more general terms, vocabulary and grammar restrictions define a *sublanguage*, *technical language*, or *register* of the language, which imposes general top-down constraints on search; in contrast, search involving unrestricted general vocabulary and syntax would be intractable. From a general lexicographic point of view, register constraints constitute pragmatic information about words which has its place in the lexicon.

The lexicon of a speech recogniser may consequently be said from a general lexicographic point of view to be *modular*, consisting basically of the *Lexical Key Identification Module*, the basic word recogniser, the *Distributional Disambiguation Module*, the stochastic language model which is statistically tuned to a given corpus, and the *Lexical Lookup Module*, which defines the lexical relation between the lexical key and its meaning).

6.8.4 The structure of lexical databases

The main features of spoken language lexical databases have already been discussed. In practice, a spoken language lexical database is often a set of loosely related simpler databases (e.g. pronunciation table, signal annotation file, stochastic word model, and a main lexical database with syntactic and semantic information), and is part of a larger complex of databases involving speech signal files, transcription files (orthographic and phonemic), and annotation (labelling) files which define a function from the transcriptions into the digitised speech signal. However, in the interests of consistency it is helpful to take a more general lexicographic point of view, and see a lexical database for spoken language development as a single database, in which relations between lexical items and their properties at all levels, from acoustics through word structure to syntax, semantics and pragmatics are defined.

The major problem in deciding how to organise a lexical database is the *ambiguity* of word forms. In a spoken language system, the focus is on the *pronunciation*, i.e. on *phonemic word forms* (not the orthography, though this is often used as a conveniently familiar form of representation). The key issue here is *homophony*, i.e. a phonemic word form associated with at least two different sets of lexical information, and thus logically involving a *disjunction* in the database.

In a simple traditional database model based on fixed-length records, in which each field represents a specific attribute of the entity which the record stands for, there is a record for each lexical entry associated with a homophone, uniquely identified by a serial number. However, for specific applications such as the training of a speech recogniser it is convenient to have just one record for each word form. In a database which is optimised for this application, the disjunction required by the homphone is *within a single record*, rather than distributed over alternative records which share the same value for the pronunciation attribute. Disjunctive information of this kind within the lexical database corresponds to nondeterministic situations and the use of complex search algorithms in actual spoken language systems.

6.8.4.1 A simple database type: pronunciation tables

Pronunciation tables hardly correspond to the intuitive concept of a lexical database, which implies a fairly high degree of complexity, but they are nevertheless a useful example of a basic kind of simple lexical database structure for the purpose of illustrating practical points of representation, modelling and organisation.

Pronunciation tables define the relation between orthographic and phonemic representations of words. Often they are defined as functions which assign pronunciations (frequently a set of variant pronunciations) to orthographic representations; this is an obvious necessity for text-to-speech lexical, but in speech recognition applications in which orthographic transcriptions (which are easier to make and check than phonemic transcriptions) are mapped to phonemic representations for the purpose of training speech recognisers, the use of a pronunciation table of this type is clearly relevant.

A pronunciation table which involves pronunciation variants (see below) provides a simple illustration of the ambiguity problem, with disjunctions within the database.

Pronunciation tables have to fulfil a number of criteria, in particular the criterion of unambiguous notation, of consistency with transliterations and transcriptions of a particular corpus, and of simple and fast processing.

General proposals for the interchange of lexical information about word forms, including morphological, phonological and prosodic information, have been made for different languages; they do not have standard status at the current time, but they are sufficiently similar to justify recommendation. A standard for French has been described (cf. Pérennou & De Calmès (1987), Autesserre, Pérennou & Rossi (1989)), containing the following features:

- Boundaries:
  - morpheme: +
  - word: #
  - liaisonless group: ##
  - phonological syntagma: § (in phrasal entries)
- Phonemes (in IPA or SAMPA notation), including a notation for the French archiphonemes.
- Phonological diacritics:
  - latency mark " for consonants pronounced in liaison contexts or morphological linking
  - consonant deletion mark ' (e.g. for final consonants)

For the spoken language lexicon in the German VERBMOBIL project the same basic principle has been adopted (cf. Bleiching & Gibbon (1994)), with extensions for incorporating prosodic information:

- Boundaries:
  - morpheme: +

- word in compounds: #
- word in phrases: ##
- syllable: .
- primary stress: '
- secondary stress: '' (two single quotes)

- Additional conventions:
  - The boundaries # and ## each imply coextensive + and . boundaries.
  - Where + and . boundaries are coextensive, . is written before +.
  - The stress marks ' and '' are written immediately before the vowel, not before the syllable.

The following extract from the VERBMOBIL pronunciation table illustrates the VERBMOBIL WIF (Word form Interchange Format) convention; following current practice, it is organised according to orthographic keys.

Transliteration	Transcription
Angst	`?'aNst`
Annahme	`?'an#n''a:.m+@`
Apparat	`?a.pa.r'a:t`
April	`?a.pr'Il`
Aprilwoche	`?a.pr'Il#v''O.x+@`
Arzttermin	`?'a6tst#tE6.m''i:n`
Aschermittwoch	`?''a.S6#m'It#v''Ox`
Auf_Wiederhören	`?aUf##v'i:.d6#h''2:.r+@n`
Auf_Wiederschauen	`?aUf##v'i:.d6#S''aU.+@n`
Auf_Wiedersehen	`?aUf##v'i:.d6#z''e:.+@n`
August	`?aU.g'Ust ?'aU.gUst`
Augustwoche	`?aU.g'Ust#v''O.x+@`
Ausweichmöglichkeit	`?'aUs#v''aIC#m''2:k.+lIC.+kaIt`

The convention has been designed to permit the removal of information which is not required, or the selection of useful subsets of the table using simple UNIX tool commands; the use of 'for primary and secondary stress permits simple generalisation over both. The plain SAMPA representation is given in the following table.

Transliteration	Transcription
Angst	`?aNst`
Annahme	`?anna:m@`
Apparat	`?apara:t`
April	`?aprIl`
Aprilwoche	`?aprIlvOx@`
Arzttermin	`?a6tsttE6mi:n`
Aschermittwoch	`?aS6mItvOx`
Auf_Wiederhören	`?aUfvi:d6h2:r@n`
Auf_Wiederschauen	`?aUfvi:d6SaU@n`
Auf_Wiedersehen	`?aUfvi:d6ze:@n`
August	`?aUgUst ?aUgUst`
Augustwoche	`?aUgUstvOx@`
Ausweichmöglichkeit	`?aUsvaICm2:klICkaIt`

6.8.4.2 A more complex lexical database

In a complex project, lexical information from several sources may need to be integrated in a fashion which permits flexible further development work even when the information cannot easily be reduced to a logically fully consistent and well-defined system. A situation such as this will arise when alternative modules, based on different principles, are to be made available for the same system. For instance, two different syntactic components will define different forms of syntactic ambiguity and be associated in different ways with semantic ambiguities. And morphological ambiguities arise with inflected forms in highly inflecting languages. In order to achieve any kind of integration, at least the word form representations will need to be consistent. The hybrid information sources will have to be represented as conjunctions of the values of independent attributes (i.e. fields within a record), with separate disjunctions, where relevant, within fields.

In general, spoken language projects have been based on the idealised notion of a *single*, *well-defined*, *consistent* and *complete*; this situation might reasonably be expected to correspond to the reality of a system developed in a single laboratory at one specific location. However, larger scale projects need to be able to cope with hybrid lexical information of the kind just outlined. A project of this type is the VERBMOBIL project funded by the German government, with international participation.

An example of a database structure designed for hybrid lexical information in this kind of context is given here.

1. Internal database structure (standard UNIX database format):
   - database: header records followed by body records
   - header: header_record_1 header_record_2 header_record_3
   - body: body_record_1 ... body_record_n
   - header_record_1: (record containing attribute names, i.e. field names)
   - header_record_2: (record defining internal conjunctive/disjunctive structure of attribute values, i.e. field contents)
   - header_record_3: (record containing source of information)
   - body_record_i: (record containing values for a given entry)
2. Example of record structure:
   - Header: (the designations A3 etc. refer to projects delivering particular types of information)

     Record 1: Orth A3 B1 C1 D1

     Record 2: Orth A3.Phon B1.Wortart,B1.Kasus,B1.Genus,B1.Num,B1.Detagr, B1.Definit,1.Semobj,B1.Semattr C1.Syncat1_C1.Syncat2 D1.Syncat

     Record 3: reference.ort a3joha.lex b1naeve.lex c1jung.lex d1peters.lex
   - Body:
     Mutter mU!t6 nomen_akk,fem,sg,@empty@,@empty@,Raute,@empty@; nomen,nom,fem,sg,@empty@,@empty@,Raute,@empty@ Nom,OBJEKTTYP nom
   - Note that the spaces designate conjunction (i.e. field separators), while the semicolons designate disjunction
3. Example of human-readable formatting

```
Entry 372: Mutter
  Orth: Mutter
  A3:   mU!t6
  B1:   nomen,akk,fem,sg,@empty@,@empty@,Raute,@empty@
        nomen,nom,fem,sg,@empty@,@empty@,Raute,@empty@
  C1:   Nom,OBJEKTTYP
  D1:   nom
```

4. Example of UNIX tool for human-readable formatting (transformation of selected named attributes of a database record into the attribute format given above):

```
#!/bin/sh
# dbviewr
# Prettyprint of single entries
# and attributes in lexicon database
# with regular expression matching
# Uses UNIX tools:
#  gawk (i.e. gnu awk), sed, tr
# (Note: sed and tr are used for illustration, and would
#       normally be emulated in gawk)
# Database structure:
# Header: Record 1: Fields containing attribute names.
#         Record 2: Other information.
# Body:   Records >2: Database relation.

if [ $# -lt 3 ]
 then
 echo "Usage: dbview dbname attribute* regexp"
 echo "dbview V2, 15.1.1994, D. Gibbon"
 exit
fi

gawk '

# Transfer the keyword from the command line to a variable.
BEGIN {keyword = ARGV[ARGC-1]}

# Identify the attributes in the first record whose values
# are to be queried.
NR == 1 {{for (i=2 ; i < ARGC ; i++)
        {for (j=1 ; j <= NF ; j++)
        if (ARGV[i] == $j) {attrib[j] = "yes"; attname[j] = $j}}}
        {for (i = 2 ; i < ARGC ; i++)
        ARGV[i]=""}}

# Find required keyword entry/entries in body of database
# Print required values and set 'found' flag
$1 ~ keyword && NR > 2  {print "\nEntry " NR-2 ":", $1
        {for (i=1 ; i <= NF ; i++)
        if (attrib[i] ~ "yes") {print "  " attname[i] ":\t" $i
                                found="yes"}}}

{last=NR}
```

```
# Print message if no entry was found for the keyword.
END {if (found!="yes") {print "No entry found for",keyword,"in",ARGV[1]}}
 ' $* |

# Translate all sequences of two colons into a slash, and all single
# colons into a single colon followed by eight spaces.
sed -e "s/;;/\//g
        s/;/&        /g" |

# Translate all single colons into a linefeed (newline).
tr ";" "\012"
```

For an overview of useful UNIX tool programming techniques, see Aho, Kernighan & Weinberger (1987), Dougherty (1990), Wall & Schwartz (1991).

6.8.5 Recommendations on lexicon structure

1. The development of re-usable resources is likely to be best supported by the use of a LKRL with an efficient and flexible implementation, and the use of formalisms in this category can therefore be recommended for development purposes where re-usable resources are concerned.
2. Whatever the choice of formalism for a spoken language lexicon, the interface between the lexicon and other components of a spoken language system needs to be given close attention.
3. A given single formalism is highly unlikely to provide optimal features for the development and construction of all components of a given system in the near future.
4. Object-oriented database structures and system architectures are likely to be used increasingly in spoken language systems in order to permit the encapsulation of module-specific optimal formalisms, implementations and information.
5. Weigh the advantages of the use of specialised commercial databases against the practical convenience and portability of ASCII database standards associated with the UNIX operating system and UNIX ASCII processing tools for database format conversion and database access.
6. Consider whether a system lexicon should be a single component, or, as is more usual in speech recognition applications, a "distributed lexicon" with lexical information associated with the *speech recognition*, *stochastic language model* and *lexical search* components.

6.9 Lexical knowledge acquisition for spoken language

The most general declarative perspective on a spoken language lexicon which is required for lexicon acquisition is that of the 'omniscient lexicographer': the lexicographer 'knows', in principle, all the possible categories and relations which may hold between a sign, its meaning and pronunciation, and other signs; all properties of a sign have equal status in terms of possible access. This idealised view, while useful for general lexical databases, is not appropriate for the construction of more specialised spoken language lexica for the classical types of spoken language system, unless these are derived from a more general lexical database as sublexica.

6.9.1 Stages in lexical knowledge acquisition

The first stage in lexical acquisition is to define the following items:

- An application dependent lexicon model which defines relevant types of lexical information.
- A lexical database model which defines the storage structure for lexical information.
- A characterisation of the sources of lexical data (e.g. other databases; manual information provided systematically or *ad hoc* by a lexicographer)
- Definition of procedures for constructing a lexicon from the data on the basis of these models and sources.
- Definition of procedures for validating the consistency of the database.
- Definition of procedures for extracting complex lexical information (e.g. coverage statistics for attributes and attribute combinations).
- Tools for implementing the procedures.

An example of on particular type of application is the construction of word-oriented information required for word recognition. In respect of a stochastic, and the three lexical information types represented by pronunciation table, frequency list, digram language model, the following steps are required (the first two of which are dealt with in more detail in the chapter on Spoken Language Corpora).

1. Pre-recording phase: Define the situational, linguistic, and physical features of the application scenario.
2. Recording phase: Record, post-process, store and structure the digitally recorded speech.
3. Post-recording phase:
   (a) Transliteration. Make an orthographic transliteration of the speech signal; for traditional data types of utterances read aloud in the laboratory, the transliteration is already given. For spontaneous speech, the transliteration must be made according to carefully defined orthographic criteria.
   (b) Frequency list. Reduce the orthographic token representations of words to an ordered set, with one token per word (in the process, make a frequency count for each word).
   (c) Digram language model. Make a list of adjacent pairs (digrams) of orthographic token representations and reduce the list to a set, with one token pair per digram (in the process, make a frequency count for each digram), and convert to an optimal data structure such as a decision tree.
   (d) Pronunciation table. Using a pronunciation dictionary or grapheme-phoneme conversion rules (statistical or knowledge-based or both), construct a table of pronunciations for the orthographic token representations.
   (e) Label alignment. The pronunciation table may then be used for forced alignment (semi-automatic label alignment, SALA) of whole words, syllables, phonemes, etc. with the speech signal, for the purpose of training a stochastic speech recogniser. The word and digram frequency information can be used to weight recognition hypotheses in terms of Bayes' Law, as discussed in connection with *language models* and *word models*.

6.9.2 Types of knowledge source

The types of lexical knowledge source for a spoken language system depend largely on the application. There are few general sources of lexical spoken language material (for instance with pronunciation and general frequency information) for any language. The construction of such a source is a major task which requires concerted action on a large scale by specialists of a whole language engineering community. It is a formidable task for many theoretical and practical reasons, but nevertheless one which will require a great deal of effort in the coming years. The two major sources of lexical knowledge for spoken language lexical systems are:

1. Existing dictionaries (to some extent).
2. Application specific corpora (to a large extent).
3. Results of descriptive, theoretical and computational linguistics (to some extent).

There is a definite lack of general resources in the area (cf. the introduction to this chapter), and the construction of application-derived, generalisable resources will be a major task for any project and for the entire spoken language community in the coming years.

General lexical material is required for the lexical knowledge in general coverage text-to-speech systems, as well as for broad application pronunciation tables for speech recognition.

6.9.2.1 Dictionaries

Useful sources of information are generally available dictionaries, particularly pronouncing dictionaries, provided that they adhere to accepted standards of consistency and expressiveness of notation, and are available in electronic form. An overview of some sources was given at the beginning of this chapter, and reference should be made to the results of the EAGLES Working Group on Computational Lexica for further examples.

6.9.2.2 Corpora

Spoken language lexica are usually application specific; they are necessarily application specific when corpus-derived frequency information is needed. An example of a corpus-derived lexicon type for speech recognition was given above. Another type of corpus-derived lexicon is the diphone word list widely used in speech synthesis technology; for this, phoneme label alignment with the speech signal is required, with the aid of which diphones are defined in the signal for further processing. The chapter on Spoken Language Corpora contains detailed information on procedures of corpus treatment, and the results of the EAGLES Working Group on Text Corpora should also be consulted.

6.9.2.3 Acquisition tools

At the current state of the art, there are few generally available tools for constructing spoken language lexica, either by extraction from existing dictionaries or from corpora. Lexicon construction usually takes place 'in house' in individual laboratories or project consortia; lexicon formats consequently vary greatly.

For information on general acquisition tools in the sense of lexicographers' work benches, reference should be made to the results of the EAGLES Working Group on Computational Lexica.

Of greatest practical use for the development of spoken language lexica in the area of word forms are the tools required for creating different kinds of word form list and word form table from corpora;

the general parameters associated with acquiring syntactic, semantic and pragmatic information are not unique to spoken language lexica (though the details, for instance of spoken language dialogue, indeed differ greatly from spoken to written language).

Standard practice is either to write custom-made programmes in C, or to use standard UNIX tools (where speed of processing is not at a premium). Neither of these procedures is particularly difficult, because of the relatively straightforward and well-understood procedures and associated algorithms.

The simplest approach for many applications where processing time is not critical, for instance with small lexica, or where batch-style processing is acceptable, is to use UNIX tools such as *grep*, *tr*, *sed*, *uniq*, *cut*, *tail*, *spell* and *awk*. For descriptions of these tools, a UNIX manual or textbook, or the man page on-line information on a UNIX system should be consulted; techniques for specific database oriented UNIX tools are described by Aho, Kernighan & Weinberger (1987), Dougherty (1990), Wall & Schwartz (1991).

An example of database formatting was given above. Simple examples of UNIX tool applications are illustrated in simplified form below in order to convey an idea of the sort of corpus pre-processing required for spoken language lexicon acquisition.

- Transliteration to word list:

```
#!/bin/sh
# Simple wordlist generator
echo Wordlist generator
tr -sc 'A-Za-z' '\012' < $1 | sort | uniq > wordlist.srt
echo Wordlist in file 'wordlist.srt'
```

- Transliteration to frequency list:

```
#!/bin/sh
# Simple word frequency generator
echo Word frequency generator
tr -sc 'A-Za-z' '\012' < $1 | sort | uniq -c > wordlist.frq
echo Wordlist in file
'wordlist.frq'.
```

- Transliteration to digram frequency table:

```
#!/bin/sh
# Simple digram table generator
echo Digram generator
tr -sc 'A-Za-z' '\012' < $1 > lines.txt
tail +2 lines.txt > tailed.txt
paste lines.txt tailed.txt | sort | uniq -c > digrams.tab
echo Digram frequency table in file 'digrams.tab'.
```

  Digram frequency information of this type is the basis for the construction of statistical *language models.*

6.9.3 Recommendations on lexicon construction

1. Identify and acquire the data resources needed for construction of the lexical database and, in the second instance, the system lexicon.

2. Construct a basic wordlist from the available data and decide on lexical coverage for the lexical database and the system lexicon.

3. Design the lexicon database to be acquired so as to maximise the information available for the development of the system lexicon.

4. Consider establishing a UNIX tool library for convenient database construction, transformation and transfer, in addition to using standard database software.

5. Results of the EAGLES Working Group on Computational Lexica should be consulted for the design of lexicon models and lexical resources for written language; some of the techniques are also applicable to spoken language lexica and lexical databases.

6. Do not underestimate the time and effort needed for constructing and validating software for lexicon construction and lexical lingware: lexicon construction is labour intensive.

In this area, there is an urgent need for the development of standardised lexicon construction tools and the provision of lexical resources.

6.10 Outlook

In this chapter, an overview of applications, requirements, lexical information, lexical representation and lexicon structure for spoken language systems is given, with particular reference to the differences between lexica designed for spoken language as opposed to written language processing. Specific recommendations were provided at the relevant points.

The state of the art for spoken language lexica is such that at the level of word forms, many basic standard techniques as discussed in the relevant sections have been developed and are widely used. Large-scale spoken language lexical resources, from reference sources of standard, stylistic and regional pronunciations through vocabularies which are characteristic of spoken language, including, for instance, discourse particles, are required for current research and development with both statistical and knowledge-based technologies.

These spoken language lexical resources in the form of actual lexical databases and tools for constructing them are, however, sadly lacking — even more so than for written language, partly because of the specific complexities of spoken language.

The major requirement in this area is simply that this gap urgently needs to be filled. Local and national project work can go some way to meeting the need. In view of the labour-intensive character of lexicon construction and, the feasibility of constructing lexical representation languages and lexical acquisition tools which can be deployed in multifunctional and multilingual environments, cooperative international initiatives are essential.

An important area which has not been touched on is the question of how spoken language lexica relate directly to the human user. This question may be answered in terms of two main areas:

1. Cognition and the spoken language lexicon.

2. Ergonomics and the spoken language lexicon.

The first answer is a theoretical perspective, pertaining to the theoretical sense of the mental lexicon, lexical access and lexicon acquisition strategies of the user, which the lexicon of a spoken language system, in some sense, models (cf. Marslen-Wilson (1989)).

The second answer touches on a more practical area, namely how the role of the lexicon in a spoken language system can be defined in such a way as to permit the most ergonomic, user-friendly development and deployment of complex spoken language systems. The former concept is an essential source of ideas and criteria for the latter.

Treatment of these areas, which would take extensive work in experimental psychology and psycholinguistics as well as user requirements into account, would be premature at present, but may be predicted to play a prominent role in future research and development.

Chapter 7

Language models

Chapter 8

Dialogue

Chapter 9

Physical characterisation

Chapter 10

Assessment methodologies and experimental design

10.1 Introduction

10.1.1 How to read this chapter

This chapter is about methodology for assessing various components involved in language engineering. Some of the sorts of things it should give you are guidelines how to go about ensuring that you have sampled enough speakers to ensure that you can make claims about how likely the results are to generalise to a speaker population at large (where population refers to your target market and will vary from application to application), how to compare performance of your recogniser or synthesiser with other that are on the market, how many speakers to include in benchmark tests of speaker verification systems to appraise performance and so on. For these purposes, an understanding of how to analyse your data statistically is needed.

At other times a user might need to test some very specific idea about, for example, what is going on with his recogniser, whether some gambit for mimicking other people's voices will allow impostors to break into a speaker verification device, what the critical acoustic attributes are that govern the perceptibility of a message in order to improve the systems and how to set up experiments with dialogue systems to check whether they will work adequately for some purpose before committing design engineers to their implementation. The way of approaching the latter group of questions calls for an understanding of the steps involved in setting up and analysing experiments.

The information provided is, then, going to cover general techniques from many diverse areas both in terms of techniques (statistics and experimentation) and applications (including the above examples and many more). However, this chapter cannot hope to be exhaustive in terms of its coverage nor choose an example for assessment which is directly applicable to all needs. However, though there will not be an example for every application encountered, the methodological tools provided should offer some idea of the way to approach many problems that will be encountered. The particular examples chosen for illustration have been raised in consultation with authors of some of the other chapters.

Three further things need stressing:

1. That although in some cases broad guidelines can be given (for example, in selecting how many speakers to choose to record), a trade off is often involved between what is ideally needed and cost (in terms of money and time). Though the procedures tell you, for instance, how

many subjects are needed for a data corpus to afford a representative sample of a population according to some criterion, if the numbers needed are too large, they can be cut down by relaxing the criteria. The amount of relaxation in the criteria that is permissible will involve an element of experience in prioritirising what is most important: So if, for example, your speech verification system is to be used for large business accounts it would be important to ensure that the system has been tested on stringent criteria. Even in cases where cost is no object, statistical techniques cannot give any absolute guarantees that the system cannot be broken into.

Where large scale sampling from a population is not possible because of prohibitive costs, but where it is necessary to report performance of the system for infrequent events, experiments may provide an alternative approach. In the speaker verification example, for instance, impostors may try and break in to the speaker verification system. This would be costly to check in three senses:

(a) To collect examples of attempted break-ins if they are infrequent.

(b) To verify that the system has been broken in to.

(c) The possible financial costs of the break-in. Here, an alternative approach would be to run experiments with individuals likely to be able to mimic voices such as actors or ventriloquists. In making the judgment whether to approach system assessment from an experimental perspective or via statistical analysis of the system in use, again the investigator will need to draw on his or her experience. If an experimental approach is decided on, the skill and experience of the investigator will again be taxed in selecting what vocal disguises an impostor might take and in the setting up of appropriate tests for checking the vulnerability of the system.

2. What will not be presented here are statistical analyses of, for example, the statistical corpora available in other chapters. What is presented here is some of the background that will allow users of the handbook to do the work for themselves.

3. Statistical development, experimental techniques and engineering products and techniques are advancing at a rapid pace. However, statistical and experimental know-how has not featured to any great extent in engineering, and statisticians and experimentalists usually have not drawn on engineering examples or considered engineers' concerns. Thus many of the "recommendations" made here are a first stab at these issues. There are often many ways of achieving a particular goal and the limited number of options that it is possible to consider here can only give a narrow perspective. As these ideas are tried out, other preferred alternatives will undoubtedly arise. Thus, at least some of the recommendations are likely to be short-lived.

10.1.2 Role of statistical analysis and experimentation in Language Engineering Standards

In talking about procedural considerations in language engineering, it will help to make things concrete. Let us assume that a client has commissioned development of a speech recognition system (System A) from scratch where expense is no object (sic). It is to be employed in a European country where all inhabitants might want to use it. At the end of the day the client wants to have some idea about how its performance compares with another system on the Market (System X). The company is given a free hand when developing the system and would prefer, for convenience purposes, to develop it on the basis of read speech though, as noted, it will eventually have to operate with spontaneous speech. The team assigned to the project decided to develop a system based on Artificial Neural Networks (ANNs). Some of the questions the team commissioned to do the work may decide to address (no claims for exhaustiveness) are:

1. How to check whether there are differences between spontaneous and read speech, then make a decision whether they can use read speech.
2. If they find differences between read and spontaneous speech that require them to use the latter, how can they check whether language statistics on a sample of recordings they make to train and test the ANNs is representative of the language as a whole? Whether read or spontaneous speech is used, segments need labelling for training the networks and judges need to be brought in for this purpose.
3. What procedures are appropriate for the tasks of labelling and classifying the segments?
4. How can the accuracy of segment boundary placement and category classification by the judges be assessed?
5. How can improvement during development stages be monitored? This usually involves correct recall of training data by the ANNs. Here segmentation and classification differences between judges (see 2 and 3) might affect recogniser performance. The preceding tests are vital to ensure that the training data is good and that changes in recogniser performance reflect improvements in the architecture, not artefacts of poor training data: An improvement in recogniser performance can be due to a genuine improvement that has been effected or a judge might have made errors and some change allows the system to make the same ones which appear to be correct (i.e., the two errors cancel themselves out). Without appropriate assessment of judges' performance, the latter can never be ruled out.
6. How do you choose appropriate test data?

The preceding highlights some of the statistical analysis and experimental procedures that need to feature in language engineering. Moreover, the specific questions raised, though appertaining to a particular issue of concern, are illustrative of many similar problems that language engineers encounter. Now we will set about attempting to provide answers to these (and other) questions. The reader should be able to employ the materials to answer related questions.

The remainder of the chapter is organised in four main sections (10.2–10.5). These are (10.2) statistical and (10.2.7) experimental techniques to ensure that the corpora employed for training and testing are representative, (10.3) assessing speech recognition, (10.4) speaker verification and (10.5) dialogue systems. Sections 10.2 and 10.2.7 introduce an understanding of statistical analysis and experimentation. Thus, the material presented in that section should be read by everyone. The materials in sections 10.2, 10.2.7 and 10.3 are specifically focussed on the hypothetical scenario outlined above. A final warning: Though the organisation of materials into these sections is convenient, note that the sub-division into sections is to some extent artificial: The relationship between setting up corpora and testing recognisers is a case of the proverbial chicken and egg — apparently poor performance of a recogniser can be due to training and testing on a poor corpus. In turn, speaker verification and dialogue systems depend to an extent on speech recognition.

10.2 Statistical and experimental procedures for analysing data corpora

In the first part of this section, some fundamental ideas in statistics will be illustrated through selected LES examples. In the subsequent parts, the steps in the scenario that has just been run through will be recapped to consider those issues that cannot be addressed until all components introduced are understood. The issues that involve experimentation will then be dealt with.

10.2.1 Statistical analysis

Statistics is the acquisition, handling, presentation and interpretation of numerical data. Language engineers have considerable experience acquiring and presenting data, but less in interpretation. The following will, then, be concerned with the latter.

10.2.2 Population/s, samples and other terminology

A *population* is the collection of all objects that are of interest for the task in hand. In the earlier example, all members of the country are the population. Here everyday use of the term *population* corresponds with its use in a statistical sense. Though *population* in a statistical sense can have the same meaning as the geographical sense, it need not necessarily have. Thus, for instance the population of users of a speaker verification system of a bank would only comprise the clients of the bank. *Population* does not only refer to humans — for example, the population of /p/ phonemes of a speaker would be all of the instances of that phoneme a speaker ever produces.

A variable is a numerical value associated with each unit of the population. Variables are divided into independent and dependent variables. An independent variable is one that is controlled or manipulated by the experimenter. So, for example, when setting up a corpus, the experimenter might deem it necessary to ensure that as many females are recorded in the test data as males. Sex would then be an independent variable (independent variables are also referred to as *factors*, particularly in connection with the statistical technique *Analysis of Variance* discussed later). A dependent variable is a variable that is measured on each experimental unit (each speaker in the preceding example) and the investigator is interested whether its value is affected by independent variables. Thus if fundamental frequency is to be the dependent variable, it is highly likely that it will be affected by the independent variable of speaker's sex.

When a variable is measured on all units of a population, a census has been taken. If it were always possible to obtain census data, there would be no need for statistics. However, since most language engineering applications (and, indeed, in many other aspects that require measurement), involve virtually infinite populations (such as those illustrated earlier of speakers or phonemes), it is not possible to measure variables on all units: In these circumstances, a finite sample is taken. This sample is used to study the variable of concern in the population. So, if you wanted an idea of the average voice fundamental frequency of men, you might make measurements on a sample of 100 men. This sample is then studied as if it is representative of the population. The statistician is able to offer an idea about the relationship between variables measured on the sample (here its mean) and, what the investigator is really interested in, the mean voice fundamental frequency of the population.

10.2.3 Sampling

The main problem in treating data statistically is how to ensure that you can obtain reliable information about the population from a sample. For this, the sample must be representative. The main requirement to achieve this is to take a simple random sample: A sample is a simple random sample if every member of the population has the same chance of being selected as every other member. Thus, if in setting up the ANN recogniser, speakers from the lab are used to obtain training data, the sample would not be simple random: It is unlikely that the members of the lab give females, all social strata, or people outside working-age a fair chance of being selected.

10.2.4 Biasses

Selection of a sample that is not a simple random sample is one of the main sources of bias in experiments. Biass can be defined as a systematic tendency to misrepresent the population. So, if the ANN recogniser is intended to be used by all members of the population, you cannot select an unbiassed sample of speakers from a sample of speakers recorded just between 9 a.m. and 5 p.m. This would exclude speakers who are at work; so, if you do this, you have obtained a biassed sample which is not necessarily representative of the target population.

If you take a sample, how sure can you be that if you measure variables such as the mean of the sample is close to the mean of the population? This sort of problem is termed estimation and is considered in the following sub-sections.

10.2.5 Estimating sample means, proportions and variances

Estimation is used for making decisions about populations based on simple random samples. A truly random sample is likely to be representative of the population; this does not mean that a variable measured on a second sample taken will be the same as the first. The skill involved in estimating the value of a variable, is to impose conditions which allow an acceptable degree of error in the estimate without being so conservative as to be useless in practice (an extreme case of the latter would be recommending a sample of the same order of magnitude as the population). The necessary background skill is to understand how quantities like sample means, variances and proportions are related to their counterparts in the population.

10.2.5.1 Estimating means

A fundamental step towards this goal is to relate the sample statistic to a probability distribution: What this means is: if we repeatedly take samples from a population, how do the variables measured on the sample relate to those of the population? To translate this to an empirical example: How sure can you be about how close your sample mean lies to the population mean? Even more concretely, if we obtain the mean of a set of samples, how does the mean of a particular sample relate to the mean of the population. As has already been said, the value of the mean of the first of two samples is unlikely to be exactly the same as the second. However, if repeated samples are taken, the mean value of all the samples will cluster around the population mean. The fact that when a number of samples are taken, the values of the mean of all the samples cluster around the same value as the mean of the population is usually described as the mean is an unbiased estimator of the population mean.

The usual way this is shown is to take a known distribution (i.e., where the population mean is known) and then consider what the distribution would be like when samples of a given size are taken. So, if a population of events has equally likely outcomes and the variable values are 1, 2, 3 and 4, the mean would be 2.5. If all possible combinations are taken (1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4, 3 and 4), the mean of the mean values for all pairs is also 2.5 (taking all pairs is a way of ensuring that the sample is simple random). An additional important finding is that if the distribution of sample means (the sampling distribution) are plotted as a histogram, the distribution is no longer rectangular but has a peak at 2.5 (1 and 4, and 2 and 3 both have a mean of 2.5 and none of the means of the other pairs has the same mean). Moreover the distribution is symmetrical about the mean and approximates more to a normal (Gaussian) distribution even though the original distribution was not. As sample size gets larger the approximation to the normal distribution gets better. Moreover, this tendency applies to all distributions, not just the rectangular distribution considered (the tendency of large samples to approximate the normal distribution is, in fact, a case of the Central limit theorem). This particular result has far-reaching implications when testing between alternative hypotheses (see below). As a rule of thumb (Recommendation 1) sample sizes of 30 or greater are adequate to approximate the normal distribution.

Another important aspect of the situation described is that the sample means have a *standard deviation* (*sd*). The *sd* of the sample means (here the *sd* of all samples of size two for the rectangular distribution) is related to the *sd* of the samples in the original distribution by the formula:

$$S.E. = \frac{population\ sd}{\sqrt{(n)}}$$

This quantity is given a particular name to distinguish it from the *sd* — it is called the *standard error* (*se*). This quantity is used in the computation of another quantity, the *z score* of the sample:

$$z\ score = \frac{sample\ mean - population\ mean}{se}$$

The importance of this quantity is that the measure can be translated into a probabilistic statement relating the sample and population means. Put another way, from the z score, the probability of a sample mean lying that far from the population mean can be computed.

To show how this is used in practice: if a sample of size n = 200 is taken, what is the probability that the mean is within 1.5 ses of the population mean? Normal distribution tables give the desired area directly as 0.8664 (see Appendix 1). Thus, approximately 86.6% of all samples of size 200 will have means within 1.5 ses of the population mean. If, as in any real experiment, one sample is taken, we can assign a statement about how likely that sample is being within the specified distance of the mean.

Before leaving this section, it is necessary to consider what to do when wanting to make corresponding statements about small-sized samples which cannot be approximated with the normal distribution. Here computation of the mean and sample error proceeds as before. Since the quantity z is used in conjunction with the normal distribution tables, it cannot be used. Instead the analogous quantity t is calculated:

$$t = sample\ mean - population\ mean/sample\ sd/SQRT\ (\mathrm{n})$$

The distribution of t is dependent on sample size and so (in essence) the t value has to be referred to different tables for each size of sample. The tables corresponding to the t distribution are usually collapsed into one table and the section of the table used is accessed by a parameter related to the sample size (the quantity used for accessing the table is n-1 and is called the degrees of freedom). Clearly, since several different distributions are being tabulated, some condensation of the information relative to the standard normal deviates is desirable. For this reason, t values corresponding to particular probabilities are given. Consideration of t tables emphasises one of the advantages of the central limit theorem can be seen insofar as one table can be used for a wide variety of issues rather than is the case for t.

10.2.5.2 Estimating proportions

Here the problem faced is similar to that with means: A sample has been taken and the proportion of people meeting some criterion and those not meeting that criterion are observed. The question is with what degree of confidence can you assert that the proportions observed reflect those in the population? Once again the solution is directly related to that discussed when estimating how close a sample mean lies to the population mean using z scores. In essence the z score for means measures:

$$z = \frac{estimated\ mean - population\ mean}{se}$$

The only difference here is that binomial events are being considered (meet/not meet the criterion). Since the mean of a binomial distribution is np (number tested x population proportion) and the se is $\sqrt{\frac{pq}{n}}$ where $q = 1 - p$), the z score associated with a particular sample is:

$$z = \frac{\hat{p} - p}{\sqrt{(pq/n)}}$$

Normal distribution tables can again be used to assign a probability associated with this particular outcome.

To illustrate with an example: Suppose that it is expected that as many men will use the ANN system as will women (p=0.5). What size of sample is needed to be 95% certain that the proportion of men and women in the sample differs from that in the population by at most 4%?

$$1.96(95\%) = \frac{.04}{\sqrt{(.5)(.5)/n}}$$

Solving for n gives 600.25. Therefore, a sample of size at least 601 should be used.

Now what are the effects if we want to be more than 4% confident, say if the difference is reduced to 2%. The required sample size jumps to 2401.

10.2.5.3 Estimating variance

The relationship between the variance of a sample and that of the population is distributed as χ^2 with n-1 degrees of freedom.

$$\chi^2 = \frac{(n-1) \textit{variance associated with samples of size n}}{\textit{variance of the normal distribution}}$$

Thus, if we have a sample of size 10 drawn from a normal population with pop variance 12, the probability of its variance exceeding 18 is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{9x18}{12} = 13.5$$

This has associated with it 9 degrees of freedom. Because χ squared values are only tabulated for particular probabilities (as with t), the probability can only be estimated for limited probabilities. In this case χ^2 lies between 0.2 and 0.1.

10.2.5.4 Ratio of sample variances

If two independent samples are taken from two normal populations with variance $\sigma 1^2$ and $\sigma 2^2$, the ratio of the two variances ($S1^2$ and $S2^2$) has the F distribution:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

If the two samples (which can differ in size) from the same normal population are taken, then the ratio of the variances will be approximately 1. Conversely, if the samples are not from the same normal population, the ratio of their variances will not be 1 (the ratio of the variances is termed the F ratio). The F tables can be used to assign probabilities that the sample variances were or were not from the same normal distribution. The importance of this in the Analysis of Variance will be seen later.

10.2.6 Hypothesis testing

10.2.6.1 Simple hypothesis testing

Many applications in Language Engineering require testing of hypotheses. An example from the scenario given in section 10.1.2 was testing whether there were differences between read and spontaneous speech with respect to selected statistics. If the statistic was mean tone unit duration in the two conditions where speech was recorded, we have a situation calling for simple hypothesis testing. This situation is called simple hypothesis testing since it involves a parameter of a single population.

Following the approach adopted so far, the concepts involved in such testing will be illustrated for this selected example. The first step is to make alternative assertions about what the likely outcome of an analysis might be. One assertion is that the analysis might provide no evidence of a difference between the two conditions. This case is referred to as the null hypothesis (conventionally abbreviated as H0) and in verbal form for the current case asserts that the mean tone unit duration in the read speech is the same as that in the spontaneous speech.

Other assertions might be made about this situation. These are referred to as *alternative hypotheses*. One alternative hypothesis would be that the tone unit duration of the read speech will be less than that of the spontaneous speech. A second would be the converse, i.e. the tone unit duration of the spontaneous speech will be less than that of the read speech. The decision about which of these alternate hypotheses to propose will depend on factors that lead the Language Engineer to expect differences in one direction or the other. These instances are referred to as *one-tailed* (*one-directional*) *hypotheses* as each predicts a specific way in which there will be a difference between read and spontaneous speech. If the Language Engineer wants to test for a difference but has no theoretical or empirical reasons for predicting the direction of the difference, then the hypothesis is said to be two-tailed. Here, large differences between the means of the read and spontaneous speech, no matter which direction they go in, might constitute evidence in favour of the alternative hypothesis.

The distinction between one and two-tailed tests is an important one as it affects what difference between means is needed to assert a significant difference (i.e., support the null hypothesis). In the case of a one-tailed test, smaller differences between means are needed than in the case of two-tailed tests. (Basically, this comes down to how the tables are used in the final step of assessing significance (see below). There are no fixed conventions for the format of tables for the different tests, so there is no point in illustrating how to use them. The tables usually contain guidance as to how they should be used to assess significance.)

Hypothesis testing involves asserting what level of support can be given in favour of, on the one hand, the null, and, on the other, the alternate hypotheses. Clearly no difference between the means of the read and spontaneous speech would indicate that the null hypothesis is supported for this sample. A big difference between the means would seem to indicate that there is a statistical difference between these samples if the direction in which the means differs is in the same direction as hypothesised for a one-tailed hypothesis or if a two-tailed test has been formulated. The way in which a decision whether a particular level of support (a probability) is provided is described next.

In the read-spontaneous example that we have been working through, we are interested in testing for a difference between means for two samples where, it is assumed, the samples are from the same speaker. The latter point requires that a related groups test as opposed to an independent groups test is required (see Figure 10.2). In this case, the t statistic is computed from:

$$t = \frac{\text{mean of condition 1} - \text{mean of condition 2}}{se}$$

Thus if the read speech for 15 speakers had a mean tone unit duration of 40.2 centiseconds and the spontaneous speech 36.4 centiseconds and the standard deviation of the difference between the means is 2.67, the t value is 1.42. The t value is then used for establishing whether two sample means lying this far apart might have come from the same (null hypothesis) or different (alternate hypothesis) distributions. This is done by consulting tables of the t statistic using n-1 degrees of freedom (here n refers to the number of pairs of observations).

In assessing a level of support for the alternate hypothesis, decision rules are formulated. Basically this involves stipulating that if the probability of the means lying this far apart assuming that the samples are from the same distribution, is so low that a more likely alternative is that the samples are drawn from different populations. The "stipulation" is done in terms of discrete probability levels and, conventionally if there is a less than 5% chance that the samples were from the same distribution, then the hypothesis that the samples were drawn from different distributions is supported (i.e. the alternative hypothesis at that level of significance). Conversely, if there is a greater than 5% chance that the samples are from the same distribution, the null hypothesis is supported. In the worked example, with 14 degrees of freedom, a t value of 1.42 does not support the hypothesis that the samples are drawn from different populations, thus the null hypothesis is accepted.

It should be noted that support or rejection of these alternative hypotheses is statistical rather than absolute. In 1/20 (5%) cases where no difference is asserted, a difference does occur (referred to as a Type II error, accepting the null hypothesis when it is false) and in cases where a 5% significance level is adopted and differences found, 1 occasion out of 20 will also lead to an error (referred to as a Type I error, rejecting the null hypothesis when it is in fact true).

10.2.6.2 Analysis of variance

This chapter of the Handbook is not going to cover all statistical tests that might be encountered, only offer a background and point to relevant material. However, some comments on Analysis of Variance (ANOVA) are called for as it is a technique that has a widespread use in Language Engineering.

ANOVA is a statistical method for detecting factors that produce variability in responses or observations. The approach is to control for a factor by specifying different values or treatment levels for it to see if there is an effect. It can be thought of as having sampled a potentially different population (different in the sense of having different means). Factors that have an effect change the variation in sample means where "factor" means a controlled independent variable. Another term encountered is *treatment level* which means "controlled values of factors".

Just to give an idea about the ANOVA approach, two estimates of the variances are obtained. The two variance estimates are the means for each factor about the overall mean (referred to as *between group variance*). The second is an estimate of the scores at a factor level about the mean factor level (referred to as *within groups variance*). If the treatment factor has had no effect, then variability between and within groups should both be estimates of the population variance. So, as discussed earlier when the ratio of two sample variances from the same population was considered, if the F ratio of between groups to within groups is taken, the value should be about 1 (in this case, the null hypothesis is supported). Statistical tables of the F distribution can be consulted to ascertain whether the F ratio is large enough to support the hypothesis that the treatment factor has had an effect resulting in different variance of the between group to the within groups means (the alternative hypothesis is supported). Another way of looking at this is that the between groups variance is affected by individual variation of the units tested plus the treatment effect whereas the within groups estimate is only affected by individual variation of the units tested.

ANOVA is a powerful tool as it has been developed to examine treatment effects involving several factors. Some examples of its scope are that it can be used with two or more factors, factors that are associated with independent and related groups can be tested in the same analysis, and so on. When more than one factor is involved in an analysis, the dependence between factors (interactions) come into play and have major implications for the interpretation of experiments. A good reference covering many of the situations where ANOVA is used is Winer (1971). Though statistical texts cover how the calculations are performed manually, these days the analyses are almost always done with computer packages. These are easy to use if ANOVA terms are known and how the output should be interpreted. Indeed the statistical manuals for these programmes (Minitab, SPSS and BMDP) are important sources about how to conduct analyses and should be consulted.

10.2.6.3 Non-parametric tests

Parametric tests cannot be used when sample sizes are small or when ranks rather than continuous measures are used since the central limit theorem does not approximate the normal distribution in these instances. In these cases non-parametric (aka distribution-free) tests have to be used. The computations involved in these tests are straightforward and covered in any elementary text book (Siegel (1956)). A reader who has followed the material presented thus far should find it easy to apply the previous ideas to these tests. To help the reader access the particular test needed in selected circumstances (parameteric and non-parametric), a tree diagram for the different decisions it is necessary to make is given in Figure 1.

A number of representative questions a language engineer might want to answer were considered at the start of this section. Let us just go back over these and consider which ones we are now equipped to answer: First there was how to check whether there are differences between spontaneous and read speech. If the measures are parametric (such as would be the case for many acoustic variables), then either an independent or related t test would be appropriate to test for differences. An independent t test would be needed when samples of spontaneous speech were drawn from different speakers to the read speech and a related t test when the samples are drawn from the same speakers. If the measures call for a non-parametric test (e.g. ratings of clarity for the spontaneous and read speech) should be used. A *Wilcoxon test* would be used when the read and spontaneous versions of the speech are drawn from the same speaker and a *Mann-Whitney U test* otherwise.

If the investigators find differences between read and spontaneous speech that require them to use the latter for training data (see application described), how can they check whether language statistics on their sample of recordings is representative of the language as a whole — or, what might or might not be the same thing, how can you be sure that you have sampled sufficient speech? The background information provided to estimate how close sample estimates are to population estimates is appropriate for this.

The next questions in our list given in the introduction lead on to the second major theme which we want to cover: How to set up a well-controlled experiment. The particular experiments that will be considered concern assessing the procedures for segmenting and labelling the speech for training and testing the ANNs though the lessons concerning good experimental design, etc. will apply to many more situations. Once we have got an idea what the experimental data would look like, we can consider how to treat the data statistically (which involves hypothesis testing again).

10.2.7 Experimental procedures

The principles of good experimental design will be illustrated by considering what procedures are appropriate to provide phonemically-labelled data which is to be used for training and, subsequently, testing a recognition system. Aspects of the procedures and considerations which will feature will apply to a wide variety of experimental situations encountered in connection with Language Engineering so the information provided will have some generality. Finally, it ought to be noted that the performance of a recogniser (to be dealt with in section 10.3) can only be as good as the data it is trained on. Also, assessment results for recognisers will be misleading if the data used for training and testing are in error or whether sampling biasses are inadvertently introduced by adoption of inappropriate sample selection. Thus, there is a close relationship between the topics discussed here and those in section 10.3.

10.2.8 Experimental selection of material to employ

The issue of what material to employ applies to speech synthesis as well as recognition (the latter being the focus here). Before concentrating on some issues relevant to recognition, a brief comment is necessary about selecting samples for basing synthesis on as these have different requirements to those of recognition: The issues are about how to select the speaker (should it be male, female, should clear speech be required etc.

If an ANN-based recogniser is to be trained and tested, different requirements are imposed: A statistically representative sample of exemplars is needed in this case. The checks for the representativeness of the sample will depend on what speech units the implementation is to be based on (e.g. are phones, demisyllables, etc. to be recognised?).

Once this decision has been made, the next question is how to elicit a representative sample of the selected speech units. Some of the materials currently employed, though motivated for positive reasons, do not fulfill the need for representativeness. Phoneticians use the term *phonetically balanced* to highlight the controls they consider necessary. A dictionary definition of "balanced" would convey the idea of 'well arranged for representativeness' and this has the same sorts of nuances that are required of a phonetically balanced passage. However, the representativeness will depend on what units the system is based on: For instance, the phonetically balanced text *Arthur the rat* is too short to include a representative sample of all the 900 or so demi-syllables that occur in samples of spontaneous speech. Thus, this text does not offer representative examples of all demi-syllables.

The lack of representativeness is more acute for other materials that have been used for training Hidden Markov Model recognisers: The SPAR sentences were developed to obtain an instance of each phoneme of English in a small amount of material. However, inspection of some of the sentences show that they look like tongue-twisters:

- *George made the girl measure a good blue vase.*
- *Be sure to fetch a file and send theirs off to Hove.*
- *Why are you early you owl?*

A speaker is likely to experience difficulty on phonemes in tongue-twister material that he would not encounter when these same phonemes occur in other sentences. Moreover, the difficulty encountered is more acute for certain classes of phonemes (consonants and most of all with plosives) than others (the vowel sounds). A person who still wants to use this material might reply that it is conceivable that these sentences could have been uttered, which is true. However, the discussion of sampling (above) illustrates that they cannot be considered a simple random sample. If it is necessary to use the phonemes in them as instances for training particular phone models, the variables measured on them should be checked statistically against other groups of sentences that also contain these phonemes. This analysis would establish whether there are differences between, say, their acoustic properties in the wider samples and these compressed versions. To my knowledge, these tests have not been conducted (the point made about tongue twisters suggests that they will not).

The issue of inappropriate control material can lead to misleading conclusions about the performance of a recogniser. This constitutes a major topic of investigation: If the recogniser is learning about phones that are produced atypically, when it fails to recognise a "typical" phone, has it made an error or not? For now, it is assumed that selection of appropriate speech samples has been made and checked using the statistical procedures outlined above. The next issue is how to label the data set for use for recognisers. There are two parts to this: to indicate where the segments start and finish, and what the identity of each segment is. We will deal with segmentation and classification in turn as illustrative examples of experimental design and statistical analysis. Note that the issue of comparing two sets of segmentations or classifications is identical, whether the two classifications are from two humans or from a human and a speech recogniser. Therefore, the procedures outlined are relevant to marking corpora as well as assessing recognisers.

10.2.9 Segmentation

Making segmentation judgments separate from classification judgements

One basic precept when conducting experiments is that the subject should have a clear idea about what decision is being made. In the case of segmenting and labelling when these are conducted together, a situation is encountered where the decisions are mixed up (or confounded). This makes it difficult during analysis to disambiguate whether, when an error occurs, the error is associated with one decision (say segmentation) or the other (classification). This problem is inherent to Signal Detection measures (see section 10.3.2). To avoid this (recommendation) judges should be required to do segmentation and categorisation as separate tasks. Segmentation needs to be conducted first and the accuracy of these decisions reported in circumstances where the judge is not able to classify the sound. The segments can then be taken and played for categorisation and the accuracy of categorisation reported (see section 10.2.10).

In the absence of any data reporting thorough analyses of segmentation accuracy that is not confounded with classification accuracy, some suggestions how this can be done are outlined. The guidance concerns:

- How to make segmentations without confounding effects of classification.
- Suggested guidelines concerning comparative analysis between human judges or between human judges and automatic algorithms.
- Recommendation about what judges to use, and
- Sample size for assessment of performance.

10.2.9.1 How to make segmentations without confounding effects of classification

In experiments, like has to be compared with like. Thus, if an automatic segmentation algorithm is available that works on acoustic patterns, to check the accuracy of the algorithm human judges should be required to do the same thing as the algorithm. To ensure this, when performing segmentations, it is advisable that the humans should not hear the speech to ensure that classification is not influencing segmentation positioning. (As an aside, it should be noted that the hypothesis whether having both audio and visual modalities available actually makes judgments of segment boundaries more reliable). Oscillographic or spectrographic representations presented visually are closer to the information supplied for a segmentation decision to a machine. Therefore, this way of presenting material for segmentation by humans avoids confounding segmentation judgments with classification decisions and allows comparison between human and machine performance. It has been demonstrated that Victor Zue (a highly experienced human judge) can make what is likely to be a more difficult decision (classification) based on this information. Consequently, ideally inexperienced judges should be used and a check made that they do not have this skill to ensure a segmentation, not a classification, decision is being made.

10.2.9.2 Suggested guidelines concerning comparative analysis between human judges or between human judges and automatic algorithms

I am not aware of any work how accuracy of segmentations compares between two human judges on the one hand with human-auto-segmentation on the other. A study conducted by my own research group suggests that segmentation accuracy might have marked effects. In this study, it was shown that different segment boundaries (labelled by different judges) can result in substantially different accuracy in an ANN recogniser indicating that disagreements about boundary placements have significant effects on recogniser classifications.

To perform a comparative test between human judges' performance and automatic algorithms ostensibly doing the same thing, it is recommended that segmentation performance be obtained from at least two judges as well as the algorithm. Parameters that might be measured are mean difference in boundary location between humans and between humans and the algorithm or correspondence between boundary placements. The data could then be analysed with ANOVA where the null hypothesis would be that there would be no difference between inter-human differences and human-algorithm differences with respect to these parameters.

Note that it is not a foregone conclusion that segmentation will be comparable between humans nor between humans and the algorithms. If, for instance, laryngeal vibration is used as a basis for segmentation, the point where it starts is often not clear cut.

10.2.9.3 Recommendation about what judges to use

It has already been mentioned that the judges should not be experienced else they might be able to classify segments based on visually presented information. A minimum recommendation would be to check with phonetically naive judges. Also, the judges who do the segmentations should be different to those who do the classifications so that there is no carry over of segmentation decisions to classification decisions.

10.2.9.4 Sample size for assessment of performance

The factors to take into account when selecting what sub-sample to perform the segmentation assessment are whether to choose sections from all speakers in case judges or speakers show specific difficulties or whether to do complete assessments on selected judges and speakers. Other factors to consider (which can be answered with the statistical techniques outlined earlier) are what length of sample to take — the sample ought to be at least long enough to contain examples of all phones of interest if phones are going to be used in the recogniser. This will ensure that where speakers have specific difficulties speaking certain phones, or judges have difficulties in locating some phones can be identified.

10.2.10 Classification

The main issue to be covered is the extent to which "judges" agree about a category to be labelled. The categories can be phonemes, lexical items, prosodic factors such as tone units, etc.

10.2.10.1 Judges

The problem that experienced judges might be able to read spectrograms has been mentioned in connection with choosing judges to perform the segmentation task. Another factor which commends using inexperienced judges for classification is as follows: Working through the file left to right might induce contextual effects: If the judge is a phonetician, he might well be influenced in boundary placement by the sound just heard. For example this sort of expert judge knows the effects of plosives on duration of the following vowel or those of pre-pausal lengthening and, consequently, this might influence his categorisation of events in a way atypical of the population of listeners from (in our ANN example) the EU country at large. (It is presumed that the recogniser is to be a model of representative listeners, not a model of listeners trained to hear things in ways that might be coloured by alternative theories.)

10.2.10.2 Procedural

Some of the biasses and measurement considerations that procedures should take into account will now be outlined. The following topics will be covered:

1. Limitations of category responses and ways of circumventing,
2. Range effects.

10.2.10.2.1 Limitations of category responses and ways of circumventing

To consider this topic, a hypothetical example will be considerd: Phone categories are embedded in a speech context and the phones have to be labelled. It is necessary to consider what sorts of factors are going to affect labelling performance in order to develop good procedures. Some of the phoneme categories which occur are differentiated along continuous acoustic measure, e.g. vowel duration signalling the distinction between /i/ and /I/ or frication duration signalling affricate or fricative. The exemplars from each continuum will range on a continuum from very clear cases of (say) /tS/ (short duration) to very clear cases of /S/ (long in duration). In this experiment, one group of subjects might indicate where each sound heard would be placed on a duration continuum using a 5-point scale. The duration of each sample is characterised by the modal value given by this group of judges. An additional group of judges then categorises each sample as /tSa/ or /Sa/. The proportion of /tSa/ category judgments (ordinate) is plotted against severity rating (abscissa) for each judge. The principle difference between groups is in the position of the /tSa/ – /Sa/ category boundary on the duration continuum (i.e. the estimated severity value where subjects gave 50% /tSa/ and 50% /Sa/ responses).

Intervals that are assigned to different categories (/tSa/ or /Sa/) by different judges leading to disagreement may be due either to (alternative 1) different judges using a common category boundary but judging the interval to be placed at different points on the duration continuum, or (alternative 2) the judges placing the interval at the same point on the duration continuum but with the judges employing different category boundaries. The first type of difference between judges are real disagreements and have a different status to the second type. The investigator might want to make a decision whether or not to include the instances in training.

To illustrate diagrammatically the situation where judges (J1 and J2) employ a common category boundary with respect to the duration continuum, intervals that judges disagree about would be depicted as follows:

Figure 10.1: *Judgement on category boundaries*

J1 J2
short extreme– – – – – – – – – – – – – – – – – – – | – – – – – – – – – – – – – – – – – – –long extreme
Common category boundary

Assuming judges place their category boundaries at different points along the duration continuum, intervals given the same duration value can be placed in different affricate/fricative categories as shown here (‘|’ indicates the category boundary):

J1
short extreme– – – – – – – – – – – | –long extreme

J2
short extreme– – – – – – – – – – – – – – – – | –long extreme

To differentiate these alternatives, it is necessary to obtain a measure of where items that judges “disagree” about lie on the continuum. The general recommendation that follows is that ratings of the physical dimension of each event being judged as well as the binary categorisation of that event should be collected in the procedures.

An additional advantage offered by collecting ratings is the improved basis of comparison between judges for establishing agreement. The measures that need developing concern how to take account of the information provided by ratings and the (related) information about where judges place their boundary to obtain an aggregate agreement measure about categories. Also, procedures for categorising synthetic phonemes have been developed where endpoints from physical continua (clear exemplars of the category) are trained and generalisation across more ambiguous instances trained (Damper) which show good comparability with human judgements. To apply these procedures to real speech, an indication where the phones lie relative to the physical dimensions on which the sound is judged needs to be obtained. The ratings provide this information.

10.2.10.2.2 Range effects

The incidence of category counts can be biassed by different acoustic properties in the surrounding recording (referred to as the *context*). To illustrate, the example of /tSa/s and /Sa/s embedded in speech that need labelling that are signalled by duration can be taken though the effects described extend to other acoustic parameters of speech. Speech variation will occur locally within an utterance and between utterances (affected by, for example, speech rate differences between speakers). Usually events are judged in context and the rate of the context, therefore, depends on which speaker is selected. Judgments about the attribute value an event has is affected by the range of the attributes in the contextual material presented for judgment at the same time (range effect, Parducci (1965)). So, here judgment about the temporal characteristics of an event to determine, for example, whether it is a /tSa/ or /Sa/ will be affected by the temporal properties of the rest of the material: A sound will have to be longer to be judged /Sa/ in a slow context than a fast one. Thus, judgments will be biassed by the contextual material. The changes in /tSa/ and /Sa/ counts are due to judges being biassed by the context not of the segment, they are not due to changes in the behaviour of the speaker. The net effect of the spurious decrease in /Sa/s when the rate is slow.

10.3 Assessing recognisers

10.3.1 Baseline performance

Many of the issues raised in connection with experimentation on segmentation (section 10.2.7) are relevant here as should be apparent. One point is the relationship between the training and test data for setting up a baseline which illustrates, once again the notions of confound and bias Say a group of speakers is available and the set includes a speaker who is markedly different from the rest. If the recognisers are trained on a subset of speakers and tested on the remainder, then when the atypical speaker is included in the test set, he will be excluded from the test set and consequently performance would be reasonably good even though the model trained is not particularly good because it includes the atypical speaker. Conversely, when the atypical speaker is not included in the training set, the model will be good. However, when the model is tested it will produce poorer performance than previously due to the inclusion of the atypical subject. Thus, it is possible to have better baseline performance for a poor model than a good model.

10.3.2 Progress

The prerequisite for assessing progress is an adequate measure of the errors the recogniser produces and how this reduces over time. Unfortunately, it is not a simple matter to derive a measure of error performance.

Some measures of recognition performance mix up errors of segmentation and classification. Thus, one list of the types of events that might occur when comparing a human judge's labels with a machine's are:

Correct: Phoneme A occurred at that point according to the transcription and an A was reported during recognition at that point too.

Mismatch: Phoneme A occurred at that point according to the transcription but something other than an A (B, C, D, ... etc.) was recognised.

Deleted: Phoneme A occurred at that point according to the transcription but nothing was reported (not an A nor anything else). This is a special subclass of mismatches.

Inserted: The transcription would lead one, say, to expect two phonemes to occur in some stretch of speech but three (including an A) were recognised. The remaining two phonemes can be aligned with the transcription so it appears that an A was inserted thing other than an A (B, C, D, ... etc.) was recognised as an A.

The thing to note here is that it is not possible to decide whether deleted and inserted phonemes are instances of segmentation or classification errors: A human judge might label a portion of speech as an affricate whereas the machine might indicate a plosive plus fricative. If the machine had used the same segment boundaries as the human, performance might have been equivalent.

The simplest type of error measure is the number of phonemes that the recogniser got correct compared with the number the human judges got correct. A basic (unresolved) question for this measure is that if humans cannot provide "perfect" classifications, the machine is being given noisy data. Specifically in connection with assessing, say accuracy of classification, the problem is what is the "correct" answer on phones that subjects are not agreed. This raises another issue specifically in connection with a particular technique that has been applied for assessing recognisers. The technique is signal detection theory and the technique will first be outlined before problems in applying it to assess recognisers (humans and machines) are discussed.

The basic idea behind signal detection theory is that errors convey information concerning how the system is operating (in this respect, it is an advance on simple error measures). In the signal detection theory model, it is assumed that there is a distribution of activity associated with the event to be detected (i.e., recognition of phoneme A). The recogniser is performing according to some criterion such that if activity is above the criterion, the recogniser reports that the phoneme is present and below the criterion, subjects report that the phoneme did not occur. Usually, the threshold is set so that most but not all activity associated with a signal leads to that phoneme being recognised. Activity associated with the signal distribution above the criterion threshold results in signals being detected (hit) and those below are "missed". Diagrammatically:

The abscissa is activity level and the distribution represents the probability distribution of events associated with the signal (phoneme A) at the various activity levels. The signals associated with other phones are "noise" in relation to the phoneme and they give rise to a distribution of noise activity which influences recognition. The noise distribution represents the probability distribution of activity levels and the criterion activity level is the same as that applied to the signal distribution. Most of the noise distribution on processes associated with good recognisers will be below the criterion but some will be above. When activity associated with the noise distribution below the criterion is encountered, subjects correctly reject this activity as being associated with phoneme A whilst when it is above this criterion, they incorrectly report a signal to have occurred — referred to in signal detection theory as a *false alarm*. Diagrammatically again:

Since the criterion is at the same activity level in each case, the two figures can be combined to give the complete model of the recognition process:

The error classes described earlier are associated with the categories needed for a signal detection analysis as follows:

Hits = correct

False alarms = False + insertions

Misses = Mismatch + deleted

Correct rejections = total phonemes – (correct + false + mismatch + deletions)

With the data available in the form of frequency counts of these categories, standard methods can be employed to ascertain (a) the separation between the mean of the noise and signal distributions and (b) the decision criterion that has been applied. These are referred to as *d'* and β respectively and *d'* is particularly important in the present context as it is a measure of the discriminability of the signal distribution fron the noise distribution which takes into account all the error information available. A work sheet of the calculations of *d'* and β and tables needed for the calculation are included as Appendix 1.

The relationship between the two classes of error — false alarms and misses — has been considered an important indicator of performance. Judges, be they machines or humans, can trade between these two classes of error. The trading relationship is referred to as a *Receiver Operating Characteristic (ROC)*.

The problem alluded to earlier in connection with SDT is what is signal and what is noise. In earlier work it has been assumed that human judges are capable of providing the "correct" answers. However, agreement between judges is notoriously low even for gross classifications (in the stuttering literature, inter-judge agreement about stutterings is as low as 60% for expert judges). The finer level of classification called for here would lead one to expect that agreement about phone classes would also be low. Possible ways out of this dilemna are (1) improvement in psychophysical procedures and (2) (related) normalisation procedures across judges to obtain some composite level of agreement.

10.3.3 Functional adequacy and user acceptance

Functional adequacy refers to the fact that recognisers have only got to perform a limited range of functions (they might not be required to deal with unrestricted speech for example). User acceptance refers to the fact that users might tolerate something that is not perfect. Each of these topics calls for metrics other than percent correct (10.3.2) and can involve subjective judgments on the part of subjects. So, for these topics, it is necessary to consider what is the best way to obtain information from users about the acceptability of a system.

The recommended way of obtaining the information is in the form of a summated rating scale (Likert (1932)). These scales are constructed by preparing sets of statements designed to measure an individual's attitude about a particular concept (here, for instance, recogniser acceptance). Typically scales are comprised of several different sub-scales (in assessing user acceptance of a recognition system, these sub-scales might include response time, format of feedback to user, etc). Respondents indicate the extent to which they agree with each statement by giving a rating (usually between 1 and 5) indicating the extent to which they agree with each statement. In order to counterbalance for response biasses, it is usual to phrase questions so that, here for example, questions indicating affirmative user acceptance would lead to low rating responses for some questions and high ratings for others. Examples of questions and response format which might be appropriate for assessing user acceptance might be:

1 2 3 4 5

I found the system easy to use. ... Sometimes I experienced difficulties in using the system.

These two questions would tend to lead users to use different poles of the rating scale. During analysis, the scale values need to be reversed.

The advantages of Likert scales are:

1. All questions constituting a concept or sub-concept can be summed to give a composite response.
2. A quantitative rather than qualitative measure is provided.
3. They have no 'right' answer, which makes it different from multiple choice format.
4. Likert scales are relatively cheap and easy to develop.
5. They are usually quick and easy for respondents to complete.

A limitation is that some level of expertise and statistical sophistication is needed to develop and validate a good scale.

The construction of questionnaires based on Likert's scale format for the items of any identified concept will involve the following steps: (i) Define the concept or set of sub-concepts to measure. (ii) Design the scale. (iii) Preliminary assessment of questionnaire. (iv) Administration and item analysis. (v) Validate and produce norms.

1. The literature needs to be reviewed to ratify the concepts identified and check whether others ought to be included. Care should be taken to ensure that these are clearly and precisely defined: A scale cannot be developed until it is clear exactly what is intended to be measured.
2. Scales will be defined for each item. This will be based on Likert's format where appropriate. The format is not appropriate for collecting information on some concepts (principally, demographic details). At this stage response choices will be specified and instructions formulated for the evaluative basis of each item. A pool of items will be generated at this step, which will be subjected to statistical analysis in later steps.
3. Preliminary assessment

Factor analysis may be used for two purposes for validating the scales:

1. Exploratory Factor Analysis. This is used for studying the mulitidimensionality of the Likert scales that underlie a concept. Two aspects of this are (a) to establish the number of factors that best represent the items, and (b) the interpretation of the factors.
2. Confirmatory Analysis Exploratory factor analysis provides an optimum statistical description of the data. However, scale construction is premised on certain assumptions about what the scale is intended to measure (e.g., response time). *Confirmatory Factor Analysis (CFA)*, may be used to verify the latter hypothesised factor structure. This can be performed using one of the available covariance structure modelling programs such as LISREL (Joreskog & Sorbom (1984)) and ESQ (Bentler (1985)).

10.3.4 Methodology

10.3.4.1 Application oriented

Assessment as part of the application will lead to speech being encountered that was not involved in setting up the material employed during training and testing.

10.3.4.2 Reference oriented

It has been considered that perfomance of a newly-developed recogniser be compared against a reference algorithm (eg Chollet and Gagnoulet). The procedures for comparing performance between the reference and newly-developed algorithms would be similar, and encounter the same problems, as those described in connection with human-human and human-algorithm performance.

10.3.4.3 Calibrated databases

The procedures for calibrating databases rely in part on checking that the sampling of the corpus is satisafactory (see section 10.2.3), or involve being able to compare performance against known answers (the problems involved in providing them has been described in section 10.2.7).

10.3.4.4 Manipulated, artificial and diagnostic databases

Specially constructed data may need to be constructed in order to test some specific ideas about why performance of the recogniser is poor. This may involve difficulties in dealing with breathing noises, hesitations, etc. or because difficulty is experienced in recognising particular phonemes or phoneme types. The construction of special data for these purposes needs to bear in mind the concerns discussed above in connection with providing adequate samples of speech (section 10.2.8).

10.3.5 Experimental design

There has been relatively little work on what psychophysical procedures are appropriate for assessing speech recognisers. Fortunately, there is a second area of investigation which involves similar questions — measuring intelligibility using different procedures in the hearing impaired. It is straightforward to see how these techniques could be applied to speech synthesis where two or more synthetic versions can be compared with the procedures. In applying the procedures to speech recognition, two or more recognisers would have to produce output which is then assessed with the procedures. The outputs of the recogniser could be in the form of transcriptions which subjects would judge or converted to speech form using a text to speech system, and the tests could then proceed in exactly the same way as with synthetic speech.

10.3.5.1 Magnitude estimation

In *magnitide estimation*, subjects choose a positive number to represent the subjective magnitude of the intelligibility of the output of different recognisers.

10.3.5.2 Rank order

The *rank order procedure*, as its name suggests, would require subjects to place the different recogniser outputs in order of magnitude of increasing intelligibility.

10.3.5.3 Pairwise comparison

In *paired comparison* as applied to recogniser assessment, subjects judge which of two recogniser outputs has more or less intelligibility.

10.4 Assessing speaker verification and recognition systems

Speaker verification needs to make use of many of the statistical and experimental procedures outlined in previous sections. The principal procedures that are required for assessing speaker verification and recogniton systems are those applied to speech recognition processes in the preceding section. The dependent variable is going to be whether speaker verified or recognised, that is a discrete variable. Discrete variables are somewhat simpler measures than those considered in connection with acoustic measures of speech recognition (note, I am not saying that the problems involved are necessarily simpler). The particular types of statistical tests called for will be non-parametric (see section 10.2.6.3).

As mentioned before, assessing speech recognisers is integrally related to speech corpora. (It is for this reason that an example was worked through from scratch addressing setting up a corpus and assessing recognition). The same applies here. So, a lot of what needs to be said has been covered. Here only the additional considerations that need dealing with are described — those to do with sampling speakers for recognition and verification and those to do with assessment.

10.4.1 Sampling rare events in speaker verification and recognition systems

In section 10.2.5.2 consideration of estimated sample size for equally frequent events was covered. In that example, p and q were the same at 0.5. If p is quite small as it may be in the case of impostors into speaker verification systems (p = 0.05, 0.01 or 0.001), it would be necessary to obtain very large samples (running in to tens of thousands). In these cases the Poisson distribution is frequently used.

10.4.2 Employing expert judgments to augment speaker verification and assessment for forensic aspects of speaker verification and recognition

One important area of speaker verification concerns forensic applications. The forensic data that might be used can be obtained in a number of ways: Experts might listen to the material and offer a subjective judgement. This judgment is basically whether the speaker is or is not a particular speaker (i.e. a discrete measure). There are lots of ways that expert judgments about speakers could be standardised which would enhance the judgments in the ways that, for example, doctors knowledge in formulating diagnosis has been done and incoporated into automatic disease diagnosis expert systems. It is beyond the scope of the current chapter to present this information. Nor is it clear whether presenting here would in fact reach the target audience. Another approach that has been taken to speaker recognition and verification has been through spectrographic measures (a continuous measure). To use all these measurements together calls for statistical techniques which can deal with mixes of continuous (parametric) and discrete (non-parametric) measures.

Generalised Linear Modelling Techniques (GLIM) allow models to be constructed that involve predicting dependent variables like those required here from mixture of binary, discrete and continuous measures (aitkin anderson francis hinde 1989)). These have not been applied to forensic applications of speaker verification/recognition but seem appropriate for the task. The technique as it would be applied here enables the experimenter to establish which acoustic and subjective factors differentiates one speaker from another. In contrast, Analysis of Variance can only deal with continuous measures.

10.5 Interactive Dialogue systems

Dialogue systems include aspects of speech recognition and synthesis. Therefore, once again, much of what has been said before is relevant to this topic. Similarly, decisions about trial sizes can be made on the basis of the statistical information provided previously.

10.5.1 WOZ

A desription of WOZ has been included in another chapter. Here it is intended to ask what this technique is useful for. The simple answer is setting up simulation of a dialogue system which will allow testing implications of a required system without committing the investigator to its actual implementation. Itself it is an experimental procedure. The advantage is its flexibility in comparison with a computational implementation. WOZ design brings the engineer and ergonomist together to realise a complex task. It should be borne in mind that WOZ is a **means to an end** not **an end**. Ultimately, if WOZ has done the job expected of it, it will be discarded and a system implemented. The expectations of the two parties to the bargain needs to be clearly specified at the outset.

The engineer expects the WOZ simulation to provide answers as to what the likely structure of the dialogue system should eventually look like. The engineers should be expected to provide certain information necessary to get the ergonomist started, in particular, a language model for the dialogue system and on-going advice about the advisability from an implementational point of view about changes the ergonomist is proposing to effect. A potential problem at the outset is if this initial specification of language is seen by the engineer as the core of the work. After initial experimentation the ergonomist will need to go back and request ways in which this and other aspects of the simulation be altered by the Language Engineer.

It is necessary to realise that for WOZ to be useful it has to be cheaper. It should also be more efficient due to the flexibility it offers. The engineer should make the group aware what time is allowed for development and testing, what call they can have on the engineer's time/expertise, etc. There seems to be a feeling that the engineer often presents problems that they know when they see a proposed answer whether it is what they want or not. This "searching in the dark" of the invesigators is an unsatisfactory situation. Obviously, it is not the role here to arbitrate which is more in the realms of politics.

The actual aim will be to attempt to come up with some experimental procedures that will allow the engineer and ergonomist to achieve their ultimate goals. The major practical requirement of WOZ is that it provides answers or proposals about what the ultimate system should be like. The major methodological question is somewhat different — how good is the simulation. Statistical and experimental procedures will be focussed on that. The comments will be organised in the same order as in the dialogue chapter (principally sub-sections 2 and 3, not the first stage nor iteration which involves the same procedures as those considered for sections 2 and 3.

Audio-alone simulations are described, followed by some brief comments on current developments of multi-modal systems.

10.5.1.1 Audio-only simulations

10.5.1.1.1 Requirements

The Language engineer should provide a description of the user dialogue which supports activities for the proposed task. A representation such as a state transition diagram, would be appropriate. A second requirement is a performance specification, i.e. a description of exactly what the device is supposed to do.

Once provided with these basics, the ergonomist needs to establish what sort of factors might limit subjects in their ability to keep up the pretext that the dialogue is with a machine and, correspondingly, what sorts of factors in the task (the job of the wizard) are likely to facilitate or prevent this pretext.

10.5.1.1.2 Subject variables

Some precautions which need to be taken are listed. First, take care to ensure that the observations are made under conditions which will elicit representative performance: the vocabulary, device users and operating environment should be as similar as possible to the device being simulated. This includes making errors similar to those which might occur in the actual device.

A typical set of instructions to the people trying to create the simulation might be as follows: You are required to transcribe all utterances of the user subject onto the computer using the keyboard. Speak aloud what is displayed by the computer, and your utterances will be transmitted back to a subject. Be careful to read out what appears on the screen, and not just repeat what the subject said. Although your speech will be distorted to make it sound like synthesised speech, try to minimise the inflections in your voice and to speak as consistently as possible, in order to enhance the “mechanical” effect.

10.5.1.1.3 Wizard variables

The requirements associated with the wizard are really those of a good experimental procedure (section 10.2.7). The output should be consistent in content, style and pace. Two examples are

- A given command should trigger the same response from the wizard,
- The response time must comply to the subject’s expectations.

Since the job of being a wizard is not easy, wizards may need to be trained to produce predefined replies or menus, etc.

Factors likely to determine variability in wizards are: the level of skill exhibited by the system subject which will include fatigue and individual differences in aptitude. The first factor may be controlled by recruiting wizards who are likely candidates for developing these skills, and by training. Individual differences will be eliminated if only one wizard is employed; however, this advantage may be reduced if the study is large in scope (see section on multimodal simulations).

Since cognitive load is high, two wizard configurations are used in recent studies — one performs the I/O (receives questions and generates the answers), the second performs “task level processing” (generates the answers to be formulated by I/O wizard). It is considered that the two-wizard setup is more likely to achieve consistency and not increase response time though these need verifying experimentally.

A final important recommendation is that there should be a permanent record of performance. To this end, questions and answers should be tape-recorded.

10.5.1.2 Multimodal

Future interactive systems may require input from more than one modality: Examples would be speech input to generate visual text or voice operated drawing programmes. When WOZ techniques are employed for these applications, the extra factors that need to be considered are:

Task complexity: The more modalities, the more functions need to be simulated.

Information bandwidth: There are many ways of providing input. In addition, the input may be too high for single wizard so that his behaviour becomes inconsistent and multiple wizard configurations would be essential.

Multiwizard configs: The need for multiwizard configurations results in issues about how to organise collaboration between them. Workload must ideally be spread equally. This is difficult since it relies on the subjects' behaviour and, thus, the roles of the wizards may need to change dynamically. For this reason another (supervisory) wizard may be needed.

Coutaz and associates have spent time developing recommendations for wizard collaboration. They structure a wizard's task in three steps:

1. Acquisition (analysis of message).
2. Interpretation what the subject's response means in connection with the task faced.
3. Formulation (the emission of an answer).

10.5.2 Dialogue metrics

The two major aspects of dialogue between humans and machines concern the content of the message and the sequence of conversational episodes between the participants. Techniques for analysis of the content of dialogue are being developed and a lot has been written about them. Space and time prevents a comprehensive treatment being given. The situation for turn-taking is more appropriate since less that is germane to Language Engineering has been written about them.

In dialogue communication between humans, conversation among speakers is characterised by turn taking: one participant, A, talks, stops; another, B, starts, talks, stops, and so we obtain an A-B-A-B-A-B distribution of talk across two participants. This transition from one speaker to another has been shown to be orderly, with precise timing and with less than 5% overlap (ervin-tripp 1979)). The mechanism responsible for regulating turn taking is also capable of operating in quite different circumstances: The number of parties may vary from two to 20 or more; persons may enter and exit the pool of participants; turns at speaking can vary from minimal utterances to many minutes of continuous talk; and if there are more than two parties then provision is made for all parties to speak without there being any specified order or 'queue' of speakers. The same system seems to operate equally well both in face-to-face interaction and in the absence of visual monitoring, as on the telephone. Concern here will be with two communicators. Though these comments have been directed at inter-human dialogue communication, the same considerations apply to human-computer interaction. The question is what sort of metrics are appropriate to characterise the interaction between two communicating parties?

10.5.2.1 Psycholinguistic metrics

sacks schegloff 1974) and sacks schegloff 1978) suggest that the mechanism that governs turn-taking, and accounts for the properties noted, is a set of rules with ordered options which operates on a turn-by-turn basis, and can thus be termed a 'local management system'. One way of looking at the rules is a sharing device operating over a scarce resource, namely control of the 'floor'. Such an allocational system will require minimum units over which it will operate. These units are, in this model, determined by various features of linguistic surface structure: they are syntactic units (sentences, clauses, noun phrases, and so on) identified as turn-units in part by prosodic means.

It is important to see that, although the phenomenon of turn-taking is obvious, the suggested mechanisms operating it are not. For a start, things could be quite otherwise: For example, it is reported of the African people, the Burundi, that turn-taking (presumably in rather special settings) is pre-allocated by the rank of the participants, so that if A is of higher social status than B, and B than C, then the order in which the parties will talk is A-B-C (albert 1972)). Of course in English-speaking cultures too there are special non-conversational turn-taking systems operative in, for example, classrooms, courtrooms and other 'institutional' settings, where turns are (at least in part) pre-allocated rather than determined on a turn-by-turn basis. Nevertheless, there is good reason to think that like may aspects of conversational organisation, the rules are valid for the most informal, ordinary kinds of talk across all the cultures of the world. There is even evidence of ethological roots for turn-taking and other related mechanisms, both from work on human neonates (trevarthen 1974) and trevarthen 1979), and research on primates (hinoff)).

Other psychologists working on conversation have suggested a quite different solution to how turn-taking works. According to this other view, turn-taking is regulated primarily by signals, and not by opportunity assignment rules at all. duncan 1972), for example, describes three basic signals for the turn-taking mechanism:

- Turn-yielding signals by the speaker,
- Attempt-suppressing signals by the speaker, and
- Back-channel signals by the auditor.

These signals are used and responded to in a relatively structured manner. On such a view a current speaker will signal when he intends to hand over the floor, and other participants may bid by recognised signals for rights to speak. One of the most plausible candidates for such signals is gaze: it seems roughly true, for example, that a speaker will break mutual gaze while speaking, returning gaze to the addressee upon turn completion (kendon 1967); argyle 1973)).

The problem here is that if such signals formed the basis of our turn-taking ability, there would be a clear prediction that in the absence of visual cues there would either be much more gap and overlap or that the absence would require compensation by special audible cues. However, work on telephone conversation shows that neither seems to be true — for example, there is actually less gap and shorter overlap on the telephone (butterworth hine 1977); ervin-tripp 1979)) and there is no evidence of special prosodic or intonational patterns at turn-boundaries on the telephone although there is evidence that such cues are utilised both in the absence and presence of visual contact to indicate the boundaries for turn-constructional units (Duncan Jr. & Fiske (1977)).

In any case it is not clear how Duncan's signal-based system could provide for the observed properties of turn-taking; for example, a system of intonational cues would not easily accomplish the observable lapses in conversation, or correctly predict the principled bases of overlaps where they occur, or account for how the particular next speaker is selected (goodwin 1979)). Therefore, the signalling view, plausible as it is, viewed as a complete account of turn-taking seems to be wrong: signals indicating the completion of turn-constructional units do indeed occur, but they are not the essential organisational basis for turn-taking in conversation.

When one speaker interrupts another, the two can be said to be disputing who has the turn. Interruptions can occur because one participant tries to dominate or disrupt the conversation, but it can also be the case that mistakes occur in the way these subtle turn-yielding signals are transmitted and received. One study has demonstrated (cf. beattie 1982)) that many interruptions in an interview with Mrs Margaret Thatcher, the ex-British Prime Minister, occur at points where (individual judges) independent judges agree that her turn appears to have finished. Beattie goes on to suggest that she is unconsciously displaying turn-yielding cues at certain inappropriate points. The analysis of the results from this study suggests that while the speaker is actually giving a number of cures to the end of her turn, the one which she considers paramount may be different from the cue which her interlocutor considers paramount. If she considered that the most decisive cue to the end of her turn was letting her voice drop to around 140Hz instead of keeping it no lower than 160Hz, whereas her interlocutor considered that the most decisive cue was a rapidly executed final fall rather than a slow fall, their respective decisions as to whether or not she had finished her turn would differ in precisely those cases which were disputed in the present sample of utterances. The transcription of the turn-final and turn-medial utterances were separately inspected, and the prosodic and vocal quality features occurring in three or more of these utterances listed. Five 'final' features were found: pitch downstep (rapid drop) before the main fall; double falling contour; allegro portion before the tonic; whispery voice; and creaky voice.

10.5.2.2 Acoustic-based measures

A disadvantage of dialogue-based metrics is that (like content analysis), they require time-consuming manual analyses. It would be better, therefore, if automatic, acoustic-based procedures could be developed. A potential problem for acoustic-based metrics of dialogue interaction is that often speakers are not acoustically isolated. This need not (and in some available recordings does not) apply over telephone connections and potentially, therefore, for many dialogue interaction systems. These allow, then, acoustic metrics of disruption to be developed which have the advantage that they are automatic.

Little work has been done on this topic (cf. howell 1990)) Clear speech and turn-taking cues in telephone dialogue, Report to BT). Since prosodic factors are the main source of turn taking cues, acoustic metrics associated with these factors (amplitude, pitch and duration) have been measured.

The terms used to describe the various components of an interruption are summarised in the next figure.

Figure 10.2: *Schematic illustration of terms used in connection with speaker interruption patterns*

Speaker A.

————————————————

————-/ Lead in /Interrupted/ Continuation \————-

Speaker B.

—————

———————-/ Interrupting \———————-

In this example, Speaker A is interrupted (unsuccessfully) by speaker B. The ordinate represents activity (which is happening when the speech is above the baseline).

Amplitude on each channel can be measured as described in CCITT Rec. P.56 (see also Carson (1984)). "Active speech level is measured by integrating a quantity proportional to instantaneous power over the aggregate of time during which the speech in question is present (called the *active time*), and then expressing the quotient, proportional to total energy divided by active time, in decibels relative to the appropriate reference." (CCITT (1988: 112)). The method works by exponentially averaging the rectified signal values. Once these are obtained, a set of thresholds spaced 6 dB apart are applied to the envelope and each value of the envelope is compared against each of the thresholds in turn. A hangover function is defined for each of the thresholds and these allow a certain length of time for the speech to expire before it is judged inactive at the threshold level in question.

The true active level occurs when the threshold is such that the difference between the active level estimate and the threshold is equal to 15.9 dB. Where (as is usually the case) the true active level falls between two test threshold values, the true threshold is found by linear interpolation between these two. At the end of processing, for each speaker, the programme gives the threshold level calculated from pass 1, the percentage of the time that the speaker is active, the activity level, and the margin activity level-threshold value. This latter value is included as a check as it should be close to 15.9.

The new additions that need to be made to apply this to telephone dialogue are that once the threshold for active speech on each channel has been determined, these values are used to determine when the speakers are active. The data can then be passed through the same process as before, except this time, each processed signal value is only compared with the threshold that was set for its channel in the first pass.

The times when a channel becomes active and inactive can be recorded, and active levels calculated for the individual active periods. When one speaker interrupts another, the level for the already active speaker from the start of this stretch of speech up to this point is calculated. The level from the point of interruption to the point where one speaker stops is measured. The level of speech of the speaker who continues from the termination of simultaneous speech to the point where this speaker becomes inactive or the other speaker interrupts again (whichever is sooner) is also calculated. Finally, group statistics are calculated (number of occasions the various sequences occur, mean and standard deviation of level before interruption, for both speakers during interruption and for the speaker who continues and duration of the interrupted speech) for the various speaker-interruption combinations.

The occurrence of pauses is calculated by determining whether there is activity at the previous time step but not at the present. A variable is set to say who the last speaker was (both if they became inactive simultaneously), and for every sample where both channels are inactive, a pause counter is incremented. When activity begins again on either or both channels, if the current speaker (where A and B become active simultaneously, they are both the current speaker) matches the previous speaker (either where both became inactive simultaneously), the silence is recorded in that person's speech. A program can give the start time and duration of pauses made by each speaker, and start time, duration and average level for each period of activity. The definition of a "pause" requires further work and no analyses are reported.

10.5.2.3 Interruption parameters

All sections of speech where a speaker is interrupted can be located by the programme described in Appendix 2. These data can be partitioned with respect to which speaker was originally speaking and which speaker continued after the interruption. These are designated AA, AB, BB and BA where the first letter refers to the speaker speaking before the interruption and the second letter for the speaker speaking after the interruption. Two factors differentiate these data — the initial speaker (factor 1) which differentiates the first and second from the third and fourth and whether the same or a different speaker continues (factor 2) which differentiates the first and third from the second and fourth.

Six measures are available for each of these combinations. These are:

1. Number of occurrences of each type.
2. Mean duration of the sections.
3. Level of the lead in for a speaker.
4. Level during interruption for that speaker.
5. Level during the interruption for the other speaker, and
6. Level of the continued speech.

The data are, necessarily incompletely crossed — the sections before and after interruption mean that no level can be measured for the other speaker on these occasions whilst levels for both speakers can be measured during an interruption. This calls for a somewhat more complicated ANOVA model to determine what factors lead to differences in interruption patterns (howell 1990)) which illustrates some of the flexibility of this analysis technique mentioned previously.

Pitch analysis. Pitch movements (fall, rise, rise-fall, fall-rise and flat, designated /, /\, \/,–) cue turn-taking behaviour under different conditions. However, there appears to be no objective method for determining what category a particular pitch movement falls into. Moreover, it is pitch movements in nuclear syllables that should be of particular interest to us, despite there being no apparent economical way of determining the location of nuclear syllables. In the absence of a better method, what happens to the fundamental frequency in the 200ms (roughly the length of an average syllable) period of voiced utterance before the time specified by the classification routine (see previous section) can be ascertained? This can be done as follows: Starting at the specified offset time, fundamental frequency is estimated in 25ms windows, moving backwards 5ms at a step. This is done by first filtering the signal at 1000Hz, autocorrelating, and then examining autocorrelation peaks for periods between 3 and 12.4ms (c. 80–333 Hz). Taking the highest peak as the fundamental frequency estimate often leads to wildly fluctuating estimates, presumably because more energy was often present in the first harmonic than in the fundamental. Thus we would see successive windows returning fundamental frequency estimates of 160Hz, 320Hz, 320Hz, 160Hz, 160Hz, etc. To iron out such fluctuations, if the peak found corresponded to a frequency value within twenty Hertz of the last estimate obtained, the frequency estimate for the current window was halved. This gives far more consistent results. If the peak found by autocorrelation is high enough — compared to a threshold which is higher if frequency is wildly fluctuating or the previous window is judged to be unvoiced — the estimated frequency is returned, otherwise the window is judged to be unvoiced. Utterances often ended with fricatives or exhalations of air with no voicing. To analyse the portion of the utterance 200ms before the end would often yield no useful information about voicing changes. The method used, therefore, was to move back in 5ms steps, but only start recording fundamental frequency values for analysis after 2 successive windows had been determined by autocorrelation to be voiced. If this has not occurred within 300ms back from the point specified by the classification routine, the pitch movement is deemed to be indeterminate. Otherwise, 36 successive fundamental frequency values are recorded (or zeros if they are deemed to be unvoiced). These are subjected to a smoothing process in the reverse order from which they were recorded (ie "forwards" in time rather than backwards) which once again substitutes half of an estimated fundamental frequency value for the estimate if the half value is within 10Hz of the previous value, and passed to the frequency movement classification routine. This routine crudely splits the 36 values into three groups of 12. If any group of 12 has all zero values, the movement type is classified indeterminate, otherwise the means of each section are compared. A frequency "change" is noted if the mean of a section is more than 10% altered from the mean of a previous section. This is a totally arbitrary method, but is no less valid than applying a t-test or any complicated and time consuming statistical analysis. The pitch movement type is classified as "rise-fall" if the mean frequency is at least 10% higher in the second section than the first, and at least 10% lower in the third section than the second. "fall-rise" is returned when the second section is lower than the first and the third section is higher than the second. A "fall" is registered if there are no "changes" noted other than in a downwards direction: either the second section is lower than the first and the third is lower than the second, or the second does not "change" from the first, but the third is lower than either the second or the first, or the second is lower than the first, but the third does not change from the second. Similarly, "rise" is returned if there is at least one upward change noted and no downward ones. "flat" is returned otherwise. In the light of this, explanation of the relationship between pitch movements classified by the program should be seen as being only loosely related to the pitch movements a phonetician might perceptually categorise the speech that is analysed as having. As well as movement classification, this part of the program returns the first, last, and mean fundamental frequency values recorded in the analysed section of speech.

10.6 Appendix 1

Calculation of d' and β for phoneme recognition

The total frequency of each of the classes 'hits', 'misses', 'correct rejections' and 'false alarms' is available. From these, we can calculate two statistics, d' and β. d' is a measure of sensitivity and β the criterion.

1. **Calculation of β**

   Let us deal with things concretely with a worked example. Here is the matrix with our four categories:

	Recognition of phoneme	
	Present	Not present
Signal	7 (hit)	3 (miss)
Noise	5 (false alarm)	5 (correct rejection)

   From this matrix work out the following probabilities:

$$p\ (Hits) = \frac{no.\ Hits}{no.\ Signals}$$

$$p\ (FAs) = \frac{no.\ FAs}{no.\ Noise\ Signals}$$

   Here,

$$p\ (Hits) = 0.70 \text{and}\ p\ (FAs) = 0.50$$

$$\beta = \frac{Zordinate\ for\ p\ (Hits)}{Z\ ordinate\ for\ p\ (FAs)}$$

   Here,

$$\beta = \frac{0.3485}{0.3989} = 0.87$$

2. **Calculation of d'**

   For d', Z values are needed for p (H) and p (FA). It is necessary to read a table of the standard normal distribution and get these values making sure the correct sign is obtained.

   (a) If the probability is greater than 0.50, it will be negative, that is, to the left of the mean of a normal curve.

   (b) If the probability is greater than 0.50, it will be positive, that is, to the right of the mean.

Since most tables of Zs are based on probabilities as area to the left of a point, either a conversion will be necessary or use the special table provided below.

The formula:

$$d' = Z < p\ (FA\text{'}s) > -Z < p\ (Hits) >$$

Here,

$$d' = 0 - (-.52) = .52$$

Prob Z	*Z*	*ord*	*Prob Z*	*Z*	*ord*
.00	indeterminate		.51	−.03	.3988
.01	+2.33	.0264	.52	−.05	.3984
.02	+2.05	.0488	.53	−.08	.3977
.03	+1.88	.0681	.54	−.10	.3970
.04	+1.75	.0863	.55	−.13	.3956
.05	+1.65	.1023	.56	−.15	.3945
.06	+1.56	.1182	.57	−.18	.3925
.07	+1.48	.1334	.58	−.20	.3910
.08	+1.41	.1476	.59	−.23	.3885
.09	+1.34	.1626	.60	−.25	.3867
.10	+1.28	.1758	.61	−.28	.3836
.11	+1.23	.1872	.62	−.31	.3802
.12	+1.18	.1989	.63	−.33	.3778
.13	+1.13	.2107	.64	−.36	.3739
.14	+1.08	.2227	.65	−.39	.3697
.15	+1.04	.2323	.66	−.41	.3668
.16	+.99	.2444	.67	−.44	.3621
.17	+.95	.2541	.68	−.47	.3572
.18	+.92	.2613	.69	−.50	.3521
.19	+.88	.2709	.70	−.52	.3485
.20	+.84	.2803	.71	−.55	.3429
.21	+.81	.2874	.72	−.58	.3372
.22	+.77	.2966	.73	−.61	.3312
.23	+.74	.3034	.74	−.64	.3251
.24	+.71	.3101	.75	−.67	.3187
.25	+.67	.3187	.76	−.71	.3101
.26	+.64	.3251	.77	−.74	.3034
.27	+.61	.3312	.78	−.77	.2966
.28	+.58	.3372	.79	−.81	.2874
.29	+.55	.3429	.80	−.84	.2803
.30	+.52	.3485	.81	−.88	.2709
.31	+.50	.3521	.82	−.92	.2613
.32	+.47	.3572	.83	−.95	.2541
.33	+.44	.3621	.84	−.99	.2444
.34	+.41	.3668	.85	−1.04	.2323
.35	+.39	.3697	.86	−1.08	.2227
.36	+.36	.3739	.87	−1.13	.2107
.37	+.33	.3778	.88	−1.18	.1989
.38	+.31	.3802	.89	−1.23	.1872
.39	+.28	.3836	.90	−1.28	.1758
.40	+.25	.3867	.91	−1.34	.1626
.41	+.23	.3885	.92	−1.41	.1476
.42	+.20	.3910	.93	−1.48	.1334
.43	+.18	.3925	.94	−1.56	.1182
.44	+.15	.3945	.95	−1.65	.1023
.45	+.13	.3956	.96	−1.75	.0863
.46	+.10	.3970	.97	−1.88	.0681
.47	+.08	.3977	.98	−2.05	.0488
.48	+.05	.3984	.99	−2.33	.0264
.49	+.03	.3988	1.00	indeterminate	
.50	0	.3989			

10.7 Appendix 2

This document describes the method we have used to analyse two-channel conversation data. The data has been digitised at 5KHz, and low-pass filtered at 2025Hz. This sampling rate was used rather than the more normal 10KHz rate for reasons of practicality: Some of the conversations are over 10 minutes long, and it proved difficult to reserve the requisite amount of contiguous disc space for a direct disc transfer of this length. Each conversation is stored in a single file with samples from each channel interlaced.

This analysis method may be broken up into two stages:

1. Locating sections of speech

   The first task is to locate the beginning and end points of connected utterances. This is done using an algorithm based on Rabiner and Schaffer's algorithm for the detection of the start and endpoints of isolated utterances. The signal on a channel is analysed in 10ms windows. The presence or absence of speech is determined not only by absolute amplitude of the signal in the window, but also by the number of zero crossings. This latter measure is included so that low energy fricatives at the beginning and end of utterances are not ignored. Pauses of under 50ms are ignored by the algorithm. The algorithm takes one pass through the data. Start and end times of utterances are written to a file. There are two such files, each containing the start and end time information for a particular channel. An utterance which is ongoing in the last window of the file is deemed to end there. This algorithm appears to work quite well, but it cannot tell the difference between "speech" and other assorted noises made by the person recorded on a particular channel (sneezes, coughs, etc). This does not appear to be a problem, as most of the "non speech" utterances ("mm's", "uh-huhs", etc.) appeared to be part of the dialogue. The files containing start and end times are used as input to the next phase of the analysis.

2. Determining the status of an utterance

   Using the start and end time data from the utterance detection module, a process of logic is applied to determine for each utterance whether it constitutes a turn taken (an utterance occurs on a channel after a pause, where the last utterance was on the other channel), or an interruption (an utterance on the other channel was ongoing when the speaker began). Interruptions are further subdivided into "successful" interruptions (the interrupter's utterance continues after the utterance that was interrupted), "unsuccessful" (where the interrupted utterance continues past the end of that of the interrupter) and interruptions where both the interrupting and interrupted speakers finish their utterances simultaneously. "Simultaneous", in this study, is only an approximate term due to the 10ms analysis window method used in computing start and end times. Cases where a speaker begins an utterance, and there has been no activity on the other channel since the end of their last utterance ("Speaker continues after pause") were not of interest in this study, as it is assumed that they were not cued by the other speaker's intonation. Where two speakers begin an utterance simultaneously, this counts as continuation after a pause for the speaker whose last utterance ended most recently ("the last person to speaker"), and as a turn take for the other speaker. If, in such a case, both speakers had also ended their last utterances simultaneously, a turn take would be made for both of them. If the beginning of a speaker's utterance is coincidant with the end of the other speaker's utterance, this is deemed to be an interruption rather than a turn take. The first utterance in a conversation is not counted.

   One pass is made through the start and end times data. When an interesting utterance (turn or interruption) is made, the frequency analysis subroutine is called to analyse part of the utterance on the other channel which immediately precede the turn/interruption. In the case of a turn take, the offset passed to the frequency analysis is that of the end of the last utterance on the other channel. In the case of an interruption the offset is the point where the interruption occurred.

The "latency" figure in the results has a different meaning depending on whether it applies to a turn take or an interruption. In the case of a turn take, it represents the perceived duration of the pause between the end of the last utterance on the other channel and the beginning of the turn taking utterance. In the case of an interruption, it represents the duration of the period where utterances are perceived to be occurring on both channels.

Chapter 11

Recognition assessment

Chapter 12

Assessment of speaker verification systems

12.1 Presentation

Until recently (say less than a century), there were hardly any circumstances when somebody at the reach of the voice was not also at the reach of the eyes, and face or silhouette were probably much more used than voice by human beings to identify each other[1]. With the development of telecommunications and recordings, the need for speaker recognition has become more pressing.

The ground for automatic speaker classification and recognition is that, beside the linguistic message, human voice conveys a lot of paralinguistic information relative to the speaker, i.e the "encoder". These factors of variability are well-known obstacles to speech recognition, as they influence the acoustic characteristics of the speech signal.

The main sources of a speaker's specificity are the physiological conformation of his speech production organs, his neuro-motor control of these organs, and his internal speech pattern prototypes. In practice, there may exist more or less systematical correlations between these factors and some of the speaker's characteristics, such as his sex, his age, his health conditions, his mood, his regional, cultural, educational background, his possible foreign accent, and the language he is speaking.

In this chapter, we address a class of pattern recognition problems where the goal is to classify a speech pattern according to some characteristics of the speaker who uttered it. We recommend the general term of *speaker classification* as the denomination of such problematics.

12.1.1 Speaker classification tasks

- If the goal is to decide whether a given speech utterance was uttered by a male speaker or a female speaker, this particular problem of speaker classification can be refered to as *sex identification*.
- When the goal is to classify a speaker within an age group, from a spoken utterance, the problem can be called *age identification*.
- Some health professionals are interested in detecting pathologies using voice samples (for instance, vocal cords disfunctionings). This problem of pathology identification is a particular case of *health state identification*[2].

[1] Except blind people.

[2] This concept could be extended to the characterization of voices modified by external temporary factors that affect speech production, such as alcohol for instance.

- Any task that would consist in determining whether a speaker is angry, sad, stressed, calm, happy, relaxed,... would resort to *mood identification*[3].
- We will gather under the term *accent identification* any process consisting in determining some aspects of the sociological background of the speaker. The most realistic one is certainly to try to identify his regional accent if he is a native speaker, or his linguistic origin if he is a non-native speaker.
- For some applications, it is necessary to classify a speaker into one out of several categories, the characteristics of which can not be expressed in objective terms[4]. This task can be covered under the general term of *speaker cluster selection*.

In the special case when the goal is to identify in which language a given speech utterance has been produced, we recommend to use the term *spoken language identification* instead of the usual expression of *language identification*, as the latter can be confused with *written language identification*.

Finally, if the task consists in finding information about the very identity of the speaker from a speech signal, it is classically designated as *speaker recognition*.

For speaker classification and recognition tasks, a general distinction must be made between *identification* and *verification*. While *identification* consists in finding to which class or speaker a speech utterance is most likely to belong, *verification* aims at validating or dismissing the hypothesis that the utterance pertains to a given class or speaker.

Examples of *speaker class identification* are given above. For what concerns *speaker class verification*, a typical problem of *age verification* would consist in checking whether a speaker is an adult or not, and *spoken language verification* would aim at making sure that such or such utterance was pronounced in a given language (the expected language of an application, for instance).

In the rest of this chapter, we will mainly focus on speaker identification and verification. However, most concepts are easy to generalise to other speaker classification problems.

12.1.2 General definitions

Directly inspired from Atal [Atal 76], here are some general definitions :

Speaker classification : any decision-making process that uses some features of the speech signal to determine some characteristics of the speaker of a given utterance.

Speaker recognition : any decision-making process that uses some features of the speech signal to determine some information on the identity of the speaker of a given utterance.

Speaker class identification : any decision-making process that uses some features of the speech signal to determine the class to which the speaker of a given utterance belongs.

Speaker class verification : any decision-making process that uses some features of the speech signal to determine whether the speaker of a given utterance belongs to a given class.

Speaker identification : any decision-making process that uses some features of the speech signal to determine who is the speaker of a given utterance.

Speaker verification : any decision-making process that uses some features of the speech signal to determine whether the speaker of a given utterance is a particular person, whose identity is specified.

[3] It is essential to underline here that this family of problem is not solved yet, and will probably never be. In particular, lie detection from the speech signal is not considered as a realistic research area.

[4] For instance, some speech recognition systems use models of speech units that have variants across several speaker clusters. These clusters may be obtained in an unsupervised manner, and it is usually impossible to find a posteriori an objective attribute that would qualify each cluster.

Spoken language identification : any decision-making process that uses some features of the speech signal to determine what language is spoken in a given utterance.

Spoken language verification : any decision-making process that uses some features of the speech signal to determine whether the language spoken in a given utterance is a particular language.

12.2 A taxonomy of speaker recognition systems

Any speaker classification or recognition system functions under 3 modes :

- A *training mode*, during which speaker or speaker class models are built and stored in the system, as reference patterns. Alternative terms : learning mode, registration, enrolment, subscription...
- A *test mode*, during which the system performs the recognition of an utterance to be identified (or verified). Alternative terms : recognition mode, trial mode, operating mode,...
- An *untraining mode*, during which a speaker or speaker class model is removed from the list of reference patterns. Alternative terms : unlearning mode, unregistration, subscription cancellation,...

Naturally, a given speaker can not be identified or verified in the test mode before he has been registered during a training phase. However, training is usually incremental in several ways. Firstly, a speaker recognition system is usually operational before all possible speakers are referenced, and new users can be added while the system is already functioning in test mode for those who are registered. Secondly, some system implementations use speech material obtained during a test phase to update training references (when the recognition decision is judged reliable enough).

Unlike training and test, untraining does usually not require an active participation of the speaker being unregistered.

12.2.1 Task typology

12.2.1.1 Speaker identification versus speaker verification

As we briefly mentioned it in the introduction, speaker recognition covers 2 different areas : on the one hand, speaker identification, on the other hand, speaker verification. As Doddington describes it [Doddington 1985], the goal of a *speaker identification* task is to classify an unlabeled voice token as belonging to one of a set of n reference speakers, whereas the *speaker verification* task is to decide whether or not the unlabeled voice belongs to a specific reference speaker.

Speaker identification is therefore a 1 out of n decision, in the case of *closed-set* identification, the result of which is an *identity assignment* to an *applicant speaker*. However, in practical applications, *open-set* speaker identification requires an additional outcome of *rejection*, corresponding to the possibility that the unlabeled speech token does not belong to any of the *registered speakers*. In such circumstances, the applicant speaker is called an *impostor*.

Speaker verification can be viewed as a particular case of open-set speaker identification, corresponding to $n = 1$. The speaker verification system takes a test voice sample and a *claimed identity* as input, and returns a binary decision : *acceptance* if the applicant speaker is considered to be the *genuine speaker* or *rejection* if he is considered to be an *impostor* (as regards the claimed identity)[5].

Conversely, open-set speaker identification can be understood as a step of closed-set speaker identification, followed by a step of speaker verification, the latter using the identity assigned by the former, as the claimed identity.

[5] Note here that the term *impostor* covers 2 slightly different concepts : a non-registered speaker in identification, and a speaker claiming a false identity in verification.

12.2.1.2 Related tasks

Beyond this major distinction between identification and verification, other related tasks can be mentioned.

- *Speaker matching*, that is to choose a speaker in a closed-set of references which is most similar to a current speaker, even though it is known in advance that the applicant speaker is not a registered speaker. This appears to be a particular case of speaker cluster selection, where each cluster consists of one speaker only.
- *Speaker labeling*, when the identity of speakers taking part to a conversation are registered, and the goal is to localise when do their successive interventions begin and end[6].
- *Speaker alignment*, when the identity and order of speakers taking part to a conversation are known, and the goal is to localise when each of their interventions begins and ends.
- *Speaker change detection*, when the goal is to detect a change of speaker along a speech stream.

12.2.1.3 Types of errors

For closed-set speaker identification, a *misclassification* error occurs when a registered speaker is mistaken for an other registered speaker (the *mistaken speaker*).

For speaker verification, two types of errors must be distinguished : *false rejection* when a genuine speaker is rejected and *false acceptance* when an impostor is accepted as the speaker he claimed he was (the *violated speaker*).

For open-set identification, the three types of errors can occur. Usually, misclassifications and false acceptances are considered as equally harmful, and therefore merged together. However, these 2 types of errors may not have the same consequences in some practical applications.

12.2.2 Levels of text-dependence

Another feature which is classically used to specify a speaker recognition system is its level of *text-dependence*, i.e the constraints on the linguistic material imposed to a test utterance. A main distinction is conventionally set between *text-dependent* systems and *text-independent* systems. Though this basic distinction is not accurate enough to cover the range of practical possibilities, we give below a definition of these 2 terms, according to the usage met in the literature. To simplify, in text-dependent systems, the linguistic content of the training and test material are totally identical, while in text-independent systems test utterances vary across trials (at least in terms of word order).

However, a deeper study of the various strategies used in practice shows that at least 5 levels of text-dependence should be distinguished. Two of them resort to text-dependent approaches, but can be opposed on the use of either a *personal password*[7] or a *common password*[8]. The other 3 can be viewed as several variants of text-independent approaches, using either fixed words in a random order (*fixed-vocabulary* systems[9]), a specific linguistic event, wherever it occurs (*event-dependent* systems[10]), or a completely unrestricted text (*unrestricted text-independent* systems[11]).

[6] Including a possible outcome of *none of the registered speaker*, in case of open-set labeling.
[7] For instance, its own name, or a personal identification number.
[8] For instance, the voice request for a given protected service.
[9] For instance, having the applicant speaker pronounce a new sequence of digits for each trial session.
[10] For instance the vowel [i], a nasal sound, the word /dog/,...
[11] However, such systems may be language dependent.

12.2.3 Interaction mode with the user

For text-dependent systems, the speech material that the user must pronounce in front of the system is a priori determined during the registration. While common-password systems have absolutely no flexibility in the choice of the linguistic material, personal-password systems can enable some text-customization, in particular the possibility for the registered user to change his voice password.

For text-independent systems, at least 3 modes of interaction with the user can be distinguished : *text-* and *voice-prompted* systems impose the (unpredictable) linguistic material to the user, whereas *unprompted systems* use totally spontaneous speech. In parallel with speaker recognition, prompted systems do explicitely or implicitely some kind of speech recognition, in order to check that the applicant speaker has really uttered what he was asked to say.

12.2.4 Definitions

Applicant speaker : The speaker using a speaker recognition system at a given instant. Alternative terms : current speaker, test speaker, unknown speaker[12], customer, user,...

Registered speaker : A speaker who belongs to the list of registered users for a given speaker recognition system[13]. For speaker classification systems, we propose the term *conform speaker* to qualify a speaker who belongs to one of the classes of speakers for a given speaker classification system[14]. Alternative terms : reference speaker, valid speaker, authorised speaker, subscriber, client,...

Genuine speaker : A speaker whose real identity is in accordance with the claimed identity. By extension : a speaker whose actual character and claimed class are in accordance[15]. Alternative terms : authentic speaker, true speaker, correct speaker,...

Impostor (speaker) : In the context of speaker identification, an impostor is an applicant speaker who does not belong to the set of registered speakers. In the context of speaker verification, an impostor is a speaker whose real identity is different from his claimed identity. Alternative terms : impersonator, usurpator[16],... For speaker classification tasks, this concept is better rendered by the term : *discordant speaker*[17].

Identity assignment : Decision outcome which consists in attributing an identity to an applicant speaker, in the context of speaker identification. For speaker classification, the term *class assignment* should be used instead.

Acceptance : Decision outcome which consists in responding positively to a speaker (or speaker class) verification task.

Rejection : Decision outcome which consists in refusing to assign a registered identity (or class) in the context of open-set speaker identification (resp. classification), or which consists in responding negatively to a speaker (class) verification trial.

(Speaker) misclassification : Erroneous identity assignementto a registered speaker in speaker identification.

[12] This term can be ambiguous in certain contexts, as it may also be understood as a speaker who is unknown to the system. Though it is frequently found in the literature, we do not recommend to use it.

[13] Usually, a speaker who is entitled to use the facilities, the access of which is restricted by the system.

[14] For instance, for a spoken language identification system that discriminates between languages spoken in Switzerland, a *conform speaker* is a speaker who speaks either German, French, Italian or Romansch, but not any other language that the system does not expect.

[15] For instance, a female speaker claiming that she is a female speaker, in sex verification.

[16] Both terms are very rarely used.

[17] For instance, a child claiming that he is an adult, in age verification.

False (speaker) rejection[18] : Erroneous rejection of a registered speaker or of a genuine speaker in open-set speaker identification or speaker verification respectively.

False (speaker) acceptance[19] : Erroneous acceptance of an impostor in open-set speaker identification or in speaker verification.

Mistaken speaker : The registered speaker owning the identity assigned erroneously to an other registered speaker by a speaker identification system.

Violated speaker : The registered speaker owning the identity assigned erroneously to an impostor in open-set speaker identification system. The registered speaker owning the identity claimed by a successful impostor, in speaker verification.

Text-dependent speaker recognition system : A speaker recognition system for which the training and test speech utterances are composed of the exact same linguistic material, in the same order (typically, a password).

Text-independent speaker recognition system : A speaker recognition system for which the linguistic content of test speech utterances varies across trials.

Personal-password speaker recognition system : A text-dependent speaker recognition system for which each registered speaker has his own voice password. *Common-password* speaker recognition system : A text-dependent speaker recognition system for which all registered speakers have the same voice password.

Fixed-vocabulary speaker recognition system : A text-independent speaker recognition system for which test utterances are composed of words, the order of which varies across speakers and sessions, but for which all the words were pronounced at least once by the speaker when he registered to the system.

Event-dependent speaker recognition system : A text-independent speaker recognition system for which test utterances must contain a certain linguistic event (or class of events) while the rest of the acoustic material is discarded. This approach requires a preliminary step for spotting and localising the relevant event.

Unrestricted text-independent speaker recognition system : A text-independent speaker recognition system for which no constraints apply to the linguistic content of the speech material.

Text-prompted speaker recognition system : a speaker recognition system for which, during the test phase, a written text is prompted (through an appropriate device) to the user, who has to read it aloud.

Voice-prompted speaker recognition systems : a speaker recognition system for which, during the test phase, the user has to repeat a speech utterance, which he listens to through an audio device.

Unprompted speaker recognition systems : a speaker recognition system using totally spontaneous speech, i.e for which the user is totally free to utter what he wants[20], or for which the system has no control over the speaker[21].

12.2.5 Examples

In this section, we give examples of well-known speaker recognition systems which can be found in the literature, in order to illustrate the taxonomy described above.

[18] Sometimes called type-I error.
[19] Sometimes called type-II error.
[20] Here, a further distinction could be made between *language dependent* and *language independent* systems.
[21] For instance, in forensic applications, the speaker may not be physically present, or may not be willing to cooperate.

12.2.5.1 Text-dependent systems

Among examples of text-dependent systems, the *Bell Labs system*, reported by Rosenberg [Rosenberg 1976] and improved by Furui [Furui 1981] is tested by the latter under the following protocol :

> *Several [six] kinds of utterance sets were used to evaluate [the] system. (). Two all-voiced sentences were used in the recordings. The males used the sentence, "We were away a year ago" and the females used the sentence, "I know when my lawyer is due."*

The first 5 utterance sets are composed of male speakers, while the last one is composed of female speakers. Performances are reported for speaker verification experiments on each set. Following our terminology, these experiments simulate a *common-password text-dependent speaker verification system*. Here, the password is an entire sentence. As Rosenberg notes it, to justify the use of a text-dependent system in practical applications :

> *For many applications, the speakers are expected to be cooperative so that a prescribed text is perfectly feasible.*

The use of a prescribed text has also the advantage that it does not need any prompting, but the drawback that it may be forgotten by the user. As discussed in a next example, a second drawback of text-dependent systems is the possibility for impostors to use pre-recorded speech.

As an example of a *personal-password text-dependent speaker verification*, one can mention a new service offered by the American telephone operator *SPRINT*. For this service, the user must speak his telephone card number through the phone, in order to have his home bill charged directly for the call he is willing to make. The system identifies the claimed customer by recognising the sequence of digits, and then verifies, on the very sequence of digits, the match between the actual user and the assumed customer. Here, the sequence of digits has a double function : a mean for customer identification and a personal voice password for speaker verification.

12.2.5.2 Fixed-vocabulary systems

An other very popular speaker verification systems was developped by Doddington, at Texas Instruments, in the early 70s. Here follow excerpts of the description given by the author [Doddington 1985] :

> *To use the system an entrant first opens the door to the entry booth and walks in, then he identifies himself by entering a user ID into a keypad, and then he repeats the verification phrase(s) that the system prompts him to say. If he is verified, the system [...] unlocks the inside door of the booth so that he may enter into the computer center. If he is not verified, the system notifies him by saying "not verified, call for assistance".*
>
> *Verification utterances are constructed randomly to avoid the possibility of being able to defeat the system with a tape recording of a valid user. An innocuous four-word fixed phrase structure is used, with one of sixteen word alternatives filling each of the four word positions (see Table 12.1).*
>
> *An example verification utterance might be "Proud Ben served hard". These utterances are prompted by voice. This is thought to improve verification performance by stabilizing the pronunciation of the user's utterance.*

GOOD	BEN	SWAM	NEAR
PROUD	BRUCE	CALLED	HARD
STRONG	JEAN	SERVED	HIGH
YOUNG	JOYCE	CAME	NORTH

Table 12.1: *Verification Phrase Construction for the TI Operational Voice Verification System (after Doddington)*

Therefore, the TI system turns out to be a *voice-prompted fixed-vocabulary speaker verification system*, the claimed identity being input as a personal identification number on a keypad. Doddington's excerpt illustrates well the motivations behind the voice-prompted fixed-vocabulary approach : the relative randomness of the verification utterances protects against impostors using pre-recorded speech, while the use of voice prompts tends to control the reproductibility of the user's pronunciation. However, it must be noted that voice-prompting may also neutralise some of the speaker specifity (in particular prosodic factors), owing to an inconscious mimicry of the prompt. At the same time, text-prompting has the drawback of requiring a specific device, such as a screen, which is not always possible to implement[22].

The experiments reported by Soong and Rosenberg [Soong 1987] where sequences of digits are used for speaker verification is another example of a fixed-vocabulary system.

12.2.5.3 Unrestricted text-independent systems

Unrestricted text-independent speaker recognition is usually considered as desirable for several reasons. Even if the user does not have the initiative of the text to utter, prompted systems are less likely to be defeated by a recorded voice, as the linguistic material is virtually unpredictable. For unprompted systems, identification or verification can take place unobstrusively, during a telephone transaction, for instance. Moreover, unprompted approaches do not require the speaker to be actively cooperative.

Here is the general structure of a *text-* (or voice-) *prompted unrestricted text-independent system*, as described by Furui [Furui 1994 !] :

> *The recognition system prompts each user with a new key sentence every time the system is used, and accepts the input utterance only when it decides that the registered speaker has uttered the prompted sentence. [...]. This method not only can accurately recognize speakers but also can reject utterances whose text differs from the prompted text, even if it is uttered by the registered speaker.*
>
> *[During registration], since the text of training utterances is known, these utterances can be modeled as the concatenation of [speaker-independent] phoneme models, and these models can be automatically adapted [to the new registered speaker]. In the recognition stage, the system concatenates phoneme models according to the prompted text [i.e a speaker-specific model and a speaker-independent model]. If the likelihood*[23] *of both speaker and text is high enough, the speaker is accepted as the claimed speaker.*

Note here that the fundamental difference between the system described above and a fixed-vocabulary system is the use of subword speech units (here, phonemes) which allow to build

[22] On the telephone, for instance

[23] The author refers to a probabilistic model, namely a Hidden Markov Model.

speaker-specific models of test words (or sentences) which were not pronounced during the registration phase. Note also the use of an explicit step of *speech* recognition.

In opposition to prompted systems, here is the example of an experiment in unprompted speaker recognition, as reported by Gish [Gish 1986], concerning the ISIS system from BBN :

> *[...] We wish to identify an unknown speaker, from an utterance, [...] knowing that the utterance was made by one of a set of M possible speakers. We have available training data for each of the M speakers that consists of speech from one or more telephone calls, all distinct from the test telephone call. The text of all utterances is assumed to be unknown.*

Here, the protocol described is *unprompted unrestricted text-independent closed-set speaker identification.* Note also the *multi-session* character of the experiment, i.e that the training and test material have been recorded through different channels, probably on different days.

12.3 Influencing factors

Many factors influence the performance of a speaker recognition system. In this section, we list the most obvious ones and stress their role on the general performance of speaker recognition systems. In particular, we discuss which are the most relevant aspects which should be explicitely taken into account and reported on when defining and describing an evaluation protocol, and how they should be handled and expressed.

12.3.1 Speech quality

A major factor influencing the performance in speaker recognition is the *speech quality*, in particular the bandwidth of the recording, the location of the speech acquisition system (anechoic chamber, studio, office, street,...), the type of microphone used, and the transmission channel (direct recording, telephone transmission, radio transmission, type of signal coding,...).

In practice, it is essential that the description of speech quality reports accurately the general characteristics, but also underlines what are the variable factors and those that remain constant, between several trials of a same user. For instance, speaker recognition through the telephone is generally more difficult than speaker recognition using studio recordings. But, two evaluations using telephone speech may give different results if for the first one, each speaker always calls from the same place using the same telephone, while for the second one he is asked to call from different locations. Naturally, the latter is likely to give better results than the former[24].

12.3.2 Temporal drift

The characteristics of a voice vary over time, depending on how tired the speaker is along the day, how stressed he is, what his mood is, whether he has a or not, etc... Moreover, it was often noted that the behaviour of users modifies while they are getting accustomed to a system. These trends can be gathered under the term *temporal drift.*

Temporal drift usually affects significantly the performance of a speaker recognition system. Intra-speaker variability within a single recording *session* is usually much smaller than *inter-session* variability. In practice, performance levels deteriorate significantly a few days, or even a few hours after registration, as compared to those obtained with *contemporaneous speech*, i.e when test utterances are pronounced immediately after the training phase is terminated. A partial solution

[24] In fact, the system may be more efficient in recognizing the handset rather than the speaker...

to temporal drift consists in using training material which is gathered over several sessions : as the collected data are more representative of the intra-speaker variability over time, more robust speaker models can be built. However, this approach makes the registration process heavier.

When the targeted application is intended to operate along time, it is necessary to design an evaluation experiment for which test material was recorded in several sessions, separated from each other by at least one day, and covering a reasonable time-span (at least a month)[25]. When multi-session recordings are available, the training material should be chosen so that it corresponds to the first recording session (or sessions, for multi-session training). Conversely, the material of a given session should never be split between training and test[26], as this would lead to an unrealistic protocol.

When these constraints are fulfiled, the number of training and test sessions, and the time-span covered by both phases should be explicited. Note that the number and time-span of training sessions has an influence on performance levels and on the user acceptability, whereas the number and time-span of test sessions have only an impact on the statistical validity of the evaluation results.

12.3.3 Speech quantity and variety

An other fundamental factor having an effect on the performance level is the *speech quantity*, i.e the amount of speech material used during the registration phase, and during each user trial phase. Usually, the level of performance increases with the speech quantity, but stabilises after a certain amount[27]. As the user's acceptability generally drops when the training phase (and a fortiori the test phase) lasts too long, a compromise has to be found when the performance improvement is judged marginal.

When registration consists of several training sessions, it is the *total training speech quantity*, i.e the *speech quantity per training sessions* multiplied by the number of training sessions, which has the most clear impact on performance. When tests are carried out with speech material from several sessions, the relevant factor is the *speech quantity per test session*.

It may also be important to distinguish the quantity of speech *uttered* by the speaker from the quantity of speech *actually used* for training and testing. This distinction is mostly appropriate for event-dependent systems and mainly for diagnostic analysis. In this case, the (average) *proportion* of speech actually used can be given for information, for the training mode and for the test mode. However, the *absolute* quantity of used speech should *not* be reported, in order to avoid ambiguities.

A related factor to speech quantity is *speech variety* : for a given speech quantity, it is usually more efficient to cover a wider range of linguistic phenomena[28]. In the lack of a universal quantitative measure of linguistic coverage, a qualitative description of the linguistic material is the only way of specifying this aspect.

12.3.4 Speaker population size and typology

In this section, we indicate in what way, and to what extent the population composition, in terms of size and typology, can affect the performance of a speaker recognition system and how it should be taken into account when designing an evaluation experiment.

When the goal is closed-set speaker identification, it is clear that the complexity of the task increases with n, the *registered speaker population size*. However the *proportion of men and women* in the population also has a direct influence, as same-sex confusions are usually much more likely

[25] One year would be ideal, for a good representativity of weather influence...

[26] Except, of course, when the evaluation is carried out on contemporaneous speech.

[27] See for instance [Soong 1987] for an illustration

[28] See again [Soong 1987] for an illustration.

than cross-sex errors[29]. If additional geographical, physiological, and even psychological and sociological information seems particularly relevant or clearly specific of the tested population, the experimentator should be aware of it and explicit it.

For what concerns *speaker verification*, the level of performance does *not* depend on the registered speaker population size, as for each trial, the complexity of the task corresponds to an open-set speaker identification with $n = 1$. A large representative population of registered speakers will only guarantee a higher statistical validity of evaluation results, whereas general conclusions will be less reliable with a small specific population.

However, a relevant issue for speaker verification (and open-set identification) is the number and typology of *pseudo-impostors*[30], i.e speakers used to model impostors during the registration phase. With more pseudo-impostors, the modeling of imposture is usually more accurate. The way pseudo-impostors are selected, and in what they differ from authorised users is also essential.

In general, each registered speaker has a corresponding *impostor model*, which represents real impostors who could claim his identity. The impostor model can be *common* to all registered speakers, or *specific* to each authorised user, if the pseudo-impostor population varies across subscribers. Pseudo-impostors can be chosen within the population of registered speakers, or originate from an external population. We will use the term *pseudo-impostor bundle*[31]to refer to the group of speakers who have been used to build the impostor model of a given registered speaker.

From a practical point of view, when impostor models are built from other registered speakers, the recording burden is lighter, but the impostor models may be less representative of imposture in general. If an additional population of external speakers is used, the *number of additional pseudo-impostors*, their population typology, as well as the speech quantity and number of sessions required from each of them should be specified.

Incidentally, for the evaluation of a speaker verification system, a test impostor should not be part of the pseudo-impostor bundle of the speaker he is claiming to be, as the real rejection abilities of the system may be over-estimated otherwise. On the opposite, there is no objection in having a registered speaker belonging to his own pseudo-impostor bundle, as is the case when the whole registered population is used to build a common impostor model.

12.3.5 Speaker purpose and other human factors

The motive for which a speaker in using a system also influences considerably its performance profile. We first describe a possible typology of applicant speakers as regards their objective. Then we mention other relevant human factors.

When the user's goal is conform to the system's purpose, a *cooperative (registered) speaker* can be defined as an authorised applicant who is willing to be identified or as a genuine speaker who intends to be verified positively. Their counterparts, in the impostor population would be a *well-intentioned impostor*, i.e an impostor having the goal of being rejected[32].

When the user's goal and the system's purpose are inverse, an *uncooperative (registered) speaker* knows that he is being verified but wants the system to reject him[33]. For instance, an uncooperative speaker is likely to use natural or artificial *voice masking* in order to remain anonymous. On the opposite, an *intentional impostor* has the clear goal of being identified or verified though he is not registered (*violation*), or to be identified as somebody else (*usurpation*).

[29]In other words, it is intrinsically easier to identify 1 speaker among 10 than to identify 1 among 1000, and it is intrinsically easier to identify 1 male speaker among 1000 adult speakers rather than among 1000 male speakers.

[30]Sometimes called *background speakers*

[31]New term, that generalises the concept of *cohort*.

[32]This concept is mostly academic !

[33]This can be the case for forensic applications.

Here, a distinction must be made within intentional impostors depending on whether they previously have or have not been in contact with the voice of the authentic user whose identity they are claiming. We propose the term *acquainted impostor* to qualify an intentional impostor who has some knowledge on the voice of the authorised speaker, by opposition to *unacquainted impostors*, when the impostor has never been in contact with the genuine or authentic user. The degree of success of an acquainted intentional impostor will ultimately depend on his *imitation* skills.

The term *casual impostor* is often used to qualify speakers who are used as impostors in an evaluation, but who were not recorded with the explicit instruction of trying to defeat the system. In the same way, the term *casual registered speakers* can be used to refer to a population of registered speakers who have not received an explicit instruction of succeeding in being identified or verified positively[34].

Here again, variants appear depending on the way the experimentator chooses the claimed identity of a casual impostor, in a verification experiment. A casual impostor can be tested against all registered users systematically[35], against all other registered speakers of the same-sex, against all other registered speakers of the opposite sex, against k registered speaker chosen at random, against the k nearest neighbours in the registered population,...

Whereas, in a first approximation, a population of casual registered speakers may be relatively representative of a population of cooperative registered speaker, no test protocol using casual impostors can accurately approximate the behaviour of intentional impostors. In practice, a real impostor could try to vary his voice characteristics for a fixed identity along successive trials, until he succeeds to defeat the system, gives up, or until the system blacklists the genuine user. Or he may try as many registered identities as he can with his natural voice or a disguised voice, until he succeeds, gives up, or until the police arrives !

However, most laboratory evaluations use speech databases which have usually not been recorded in a real-world situation. Therefore, they do not model accurately, neither cooperativeness, nor intentional imposture and the impostor speakers are casual impostors. A frequent practice is to use an *exhaustive attempt* test configuration, for which each impostor is successively tested against each registered speaker. We suggest to adopt a slightly different approach. Two distinct experiments should in fact be carried out. One for which each casual impostor utterance is tested against all registered identities of the same sex, and a second one for which each casual impostor utterance is tested against all registered identities of the *opposite* sex. The first experiment allows to estimate the rejection ability of a system towards unacquainted intentional impostors how would know the sex of the genuine speaker, even though casual impostors are almost well-intentioned impostors. The second experiment allows to check that the system is really robust to cross-sex imposture[36]. We will refer to these configurations as a *selective attempt against all same-sex speakers* and *selective attempt against all cross-sex speakers* respectively. In a first approximation, the proportion of successful violations does not depend on the number of registered speakers.

In addition, testing each impostor utterance against its nearest neighbour in the registered population can give a *hint* of the system's robustness against intentional imposture. However, the result will be directly influenced by the registered speaker population size. Therefore, this approach is only meaningful in the framework of a comparative evaluation, on a common database. This approach can be qualified as a *selective attempt towards the nearest registered neighbour*. Other selective attempts are possible, such as towards the same age class speakers, for instance.

To summarize, registered speakers should be qualified as cooperative, casual or uncooperative, whereas a distinction should be made between well-intentioned, casual or (acquainted / unacquainted) intentional impostors. Only field data can provide realistic instances of user behaviour.

[34] Or who are not even aware that they are being recorded.

[35] Except himself, if he is also a registered speaker

[36] This step is really necessary. It can happen that a system is more robust to same-sex imposture than to cross-sex imposture, in particular if the pseudo-impostor bundle of a given speaker is only composed of speakers of the same sex [Reynolds 95].

Additionally, the general motivation and behaviour of the users can have an impact on the performance of a system : for instance, what are the stakes of a successful identification or verification, the benefits of an imposture, the feeling of users towards voice technology in general,... In evaluation, all these aspects influence the motivation of the user, and therefore the interpretation of the results.

12.3.6 Recommendations

The description of an evaluation experiment or an assessment protocol, concerning a speaker classification or recognition system should explicitely report on the following items :

- *Speech quality* :
  - the *general* speech acquisition and transmission characteristics, in particular the signal bandwidth, the nature of the noise, the SNR when measurable, the characteristics of the transmission line,...
  - the speech quality factors that *remain constant* across training sessions and test sessions, in particular concerning the environment, the microphone, the channel,...
- *Temporal drift*
  - the way the speech material is *split* between training and test material, in correlation with the chronological order of its recording; for each speaker, test material should always be posterior to the latest training material.
  - the average *number of training sessions* that are necessary to register one new speaker; in practice, the number of distinct convocations he has to answer.
  - the average *registration time-span* elapsed between the first training session and the last training session, for one new speaker.
  - the average *number of test sessions* per registered user, taken into account in the evaluation.
  - the average *operation time-span* elapsed between the first test session and the last test session, per registered user, during the evaluation.
- *Speech quantity and variety* :
  - the average *speech quantity per training session* which is used per speaker for one training session, and if relevant, the average *percentage of effective training speech quantity*, i.e the proportion of training speech which is actually used to build the registered speaker models.
  - the average *speech quantity per test session* which is used per speaker for one test session, and if relevant, the average *percentage of effective speech quantity per test session*, i.e the proportion of test speech which is actually used to identify or verify the speaker, in test mode.
  - the qualitative description or characterization of training and test *linguistic content*.
- *Speaker population size and typology*
  - for speaker identification, the *registered speaker population size*, i.e the number of registered users; this figure can also be reported for speaker verification experiments, but only as an indication of the statistical validity of the results.
  - the *proportion of male and female registered speakers* as well as any other relevant characteristics of the *typology of registered speakers*, when known; in particular, concerning the age, the dialectal origin, whether they are native or non-native speakers, etc... In parallel, any geographical, physiological, psychological or sociological *feature* that would be *common* to the registered population members (or to a majority of them) should be identified and reported.

- for speaker verification (and open-set verification), the *origin of pseudo-impostors*, i.e whether they are chosen among the registered speakers or among an external pseudo-impostor population; in the latter case, the number of external pseudo-impostors, the proportion of male and female speaker, the population typology, the speech quantity per session (per pseudo-impostor) and the number of training sessions (per pseudo-impostor) should be reported.
- for speaker verification (and open-set identification), the *origin of test impostors*, i.e whether they are chosen among the registered speakers (but claiming a false identity), among the pseudo-impostors, or among an external test impostor population. The last approach is by far the most realistic. However, when it is not feasible, an impostor utterance should never be tested against a registered speaker whose bundle of pseudo-impostor contains the test impostor.
- when an external test impostor population is used, the *number of external test impostors*, the proportion of male and female speaker, the population typology, and in particular in what their profile differs from the registered population and from the pseudo-impostor population should be described. The speech quantity per session (per test impostor) should be reported, as an indication of the statistical confidence of the evaluation results.

- Speaker purpose and other human factors
  - the *general purpose* of the system, and the adequation of the evaluation data to this purpose.
  - the *intention of registered speakers*, i.e whether, in the test mode, they are cooperative speakers, uncooperative speakers, or if they behave as casual registered speakers.
  - the *intention of impostors*, i.e whether they are well-intentioned impostors, casual impostors or intentional impostors. In the case of intentional impostors, the amount of knowledge that they have on the true speaker should be specified, in particular whether they are *acquainted by voice* to the genuine speaker, and whether they are or not provided with the *password*, for text-dependent systems.
  - the *impostor test configuration*, i.e what is the simulated (or real-life) strategy of an impostor in choosing to claim such or such identity. For instance an *exhaustive attempt*, if each impostor tries each registered identity, or a *selective attempt*, if a certain criterion guides the impostor's choice, this criterion being stipulated. With laboratory recordings of casual speakers, we recommend the same-sex selective attempt configuration and the cross-sex selective attempt configuration, to which can be added the selective attempt towards the nearest registered speaker, especially for comparative evaluation of two systems on a same database.
  - the *stakes* of the system, i.e. what are the sources of motivation for registered speakers to be recognised (or not recognised), and those of an impostor to be accepted (or rejected).

12.3.7 Example

Here is what could be the description of an evaluation experimental protocol[37].

> *The following protocol was designed to estimate the performance of a speaker verification system for the protection of personal portable telephones. The principles of the targeted security system is a personal-password text-dependent speaker verification system. Before a user can place a call on his portable phone, he is asked to utter his identity. i.e his name and surname. The compatibility between the speaker and the authorised owner is checked locally, and in case of acceptance, the speaker is allowed to*

[37] The experiment reported here is *hypothetical*.

dial his number.

To simulate this application, the following experimental protocol was set up. A group of portable phone owners were provided with a (slightly modified) miniature tape-recorder (the size of a dictaphone), and were asked to record their name and surname before they placed a call on their phone, except if they had already done so during the previous 3 hours. To make sure that some users would not record all their utterances consecutively, a temporisation was implemented in the tape-recorder, so that a time-interval of 3 hours had to be respected between 2 activations of the record function. In counterpart to a user's participation, his subscription to the portable phone service was paid for, for 6 months. In practice, the 6 month subscription was refunded to a user when he brought back a recorded tape containing 100 recordings. This number corresponds approximately to 1 session every other day over 6 months. In reality, the average time after which a tape was returned was 4.2 months.

Once a tape and a tape-recorder were returned, the tape's content was digitized at a sampling frequency of 16 kHz, and the data were segmented automatically (a beep had been internally recorded on the tape each time the "stop" button was pressed). The speech material was not verified exhaustively, but a speech activity detector was used to discard utterances that were composed of silence only. In the average, 97 % utterances were kept. Silent signal portions lasting longer than 0.2 seconds were removed automatically. The typical bandwidth of the tape-recorder's microphone is 150-6000 Hz, which is within the tape's bandwidth. All tapes were of the same trade mark, and their noise level was judged negligible. Despite the fact that, for a given speaker, the microphone and the tape characteristics remained constant for all recording sessions, the data collection protocol can be considered as realistic for the targeted application.

The first 5 recordings for each speaker were used as training material, whereas the remaining ones were used as test material (92, in the average). The average registration time-span was estimated to be (5 / 97) $\times$ *4.2 months* $\approx$ *6.5 days, which may be an over-estimate of the actual time-span, as users were probably more bound to record often their voice at the beginning of the experiment. In the same way, the average operation time-span was considered to cover approximately 4 months.*

An initial population of 188 persons accepted to take part to the experiment, but 19 of them never returned the recording device, either because they lost it, or because they lost interest in the experiment. Additionally, 7 tape-recorders and 3 tapes got deteriorated during the 6 month time-span. Altogether, only 159 different speakers were thus taken under consideration as registered speakers, among which 92 were male speakers (i.e 58 %). All of them were adults over 18. Nothing else about their profile was studied, but they are likely to correspond to a relatively rich fraction of the population, since they can afford a portable phone.

In this database, a speaker utters his name and surname in 0.8 seconds in the average, but this figure varies significantly from one person to the other. The linguistic content of the speech material can not be specified other than exhaustively.

For impostor modeling, we used all recordings corresponding to the registration phase for all speakers, which we pooled together to form a speaker-independent text-independent model. We then derived an impostor model for each registered speaker as the representation of the user's training pronunciations according to the speaker-independent model. In other words, all registered speakers were part of the pseudo-impostor bundle of a given speaker, including this very speaker.

A set of 6 professional imitators (4 male, 2 female) where then asked to simulate acquainted intentional test impostors. For each registered speakers of the same sex, they were provided with the tape-recorder of the genuine user, and could listen as much as they wanted to the training material of this user. Then, they were asked to produce 5 imitated utterances of the speaker saying his name and surname. These imitations were recorded on the user's tape-recorder, at the end of the user's tape. Given the experimental protocol, it was not possible to provide the imitators with any feed-back concerning their success or failure to break the system. Altogether, each male imitator recorded approximately 5 $\times$ 92 = 460 impostor tests against registered male speakers, and each female speaker produced about 5 $\times$ 67 = 335 impostor tests against registered female speakers. All imitators were paid for their work. The imitated speech followed the same processing as the genuine one.

For the evaluation of the system performance, each authentic test utterance was tried with the genuine identity (159 $\times$ 92 = 14628 authentic trials) and each imitated utterance was tried against the targeted identity (4 $\times$ 460 + 2 $\times$ 335 = 2510 impostor trials).

We leave to the reader the pleasure of tracking the unavoidable experimental biases that remain in this imaginary experimental protocol, and how they could be circumvented.

12.4 Scoring Procedures

The design and implementation of an experimental evaluation protocol is usually intended to provide an estimation of a system performance. Beside the overall figures of error rates, other relevant quantities can be derived from the analysis of the test results, for diagnostic purposes.

In this section, we detail what pieces of information can be derived from evaluation results, and how to score them.

12.4.1 Notation

12.4.1.1 Registered speaker population

We will denote as $\mathbf{X}$ a population of m registered speakers :

$$\mathbf{X} = \{X_1 \ldots X_m\} = \{X_i\}_{1 \le i \le m}$$

For scoring purposes, we will consider a set of test utterances. We will use the term *genuine test utterances* for those which correspond to registered speakers, and the term *impostor test utterances* for those which are treated as belonging to impostor speakers. Note that the same speech segment can be used as a genuine utterance and as an impostor utterance, in some test configurations. Therefore, different notations can correspond to the same speech utterance.

Each registered speaker X_i is supposed to have produced c_i genuine test utterances, the set of which will be denoted as $\mathbf{x}_i$:

$$\mathbf{x}_i = \{x_i^1 \ldots x_i^{c_i}\} = \{x_i^k\}_{1 \le k \le c_i}$$

where superscript k denotes the k^{th} test utterance of speaker X_i.

In the rest of this chapter, we will denote as c the total number of genuine test utterances, and as p_i the proportion of utterances belonging to speaker X_i in the test set, i.e :

$$c = \sum_{i=1}^{m} c_i \quad \text{and} \quad p_i = \frac{c_i}{c}$$

With the convention :

$$u_i^* = 1 \quad \text{if} \quad c_i \neq 0 \quad \text{and} \quad u_i^* = 0 \quad \text{otherwise}$$

we will denote :

$$m^* = \sum_{i=1}^{m} u_i^*$$

Integer m^* is the number of registered speakers for which there is at least one genuine test utterance.

Finally, we will denote as M the set of *male* registered speakers, as F the set of *female* registered speakers and as m_M^* and m_F^* the respective number of male and female registered speakers for which there is at least one genuine test utterance.

12.4.1.2 Test impostor population

In the most general case, the whole set of impostor test utterances can be divided in subsets corresponding to one among n impostors $\{Y_j\}_{1 \leq j \leq n}$ using the system with a claimed identity X_i. Hence the (heavy) general notation :

$$\mathbf{y}_{j\bar{\imath}} = \{y_{j\bar{\imath}}^1 \dots y_{j\bar{\imath}}^{d_{j\bar{\imath}}}\} = \{y_{j\bar{\imath}}^k\}_{1 \leq k \leq d_{j\bar{\imath}}}$$

to denote the set of $d_{j\bar{\imath}}$ impostor test utterances produced by impostor Y_j claiming he is X_i.

Similarily to genuine test utterances, we will denote as d the total number of impostor utterances, and as $q_{j\bar{\imath}}$ the proportion of impostor tests by impostor Y_j against registered speaker X_i, i.e :

$$d = \sum_{j=1}^{n} \sum_{i=1}^{m} d_{j\bar{\imath}} \quad \text{and} \quad q_{j\bar{\imath}} = \frac{d_{j\bar{\imath}}}{d}$$

We will also denote :

$$d_{\bar{\imath}} = \sum_{j=1}^{n} d_{j\bar{\imath}} \;, \quad \tilde{d}_j = \sum_{i=1}^{m} d_{j\bar{\imath}} \qquad \text{and} \qquad q_{\bar{\imath}} = \frac{d_{\bar{\imath}}}{d} \;, \quad \tilde{q}_j = \frac{\tilde{d}_j}{d}$$

When the identity of impostors is not a relevant factor (or is unknown), subscript j can be dropped, and $y_{\bar{\imath}}^k$ denotes the k^{th} impostor attempt *against* registered speaker X_i ($1 \leq k \leq d_{\bar{\imath}}$). Conversely, in open-set speaker identification, impostors do not claim a particular identity; they just try to be identified as one of the registered speakers, who ever this speaker may be. In this case, subscript $\bar{\imath}$ can be dropped, and y_j^k denotes the k^{th} impostor attempt *from* impostor Y_j ($1 \leq k \leq \tilde{d}_j$). If, moreover, the impostor's identity does not matter, subscript j can also be dropped, and y^k simply denotes the k^{th} impostor attempt ($1 \leq k \leq d$).

With the conventions :

$$\begin{array}{llll} v_{j\bar{\imath}}^* = 1 & \text{if} \quad d_{j\bar{\imath}} \neq 0 & \text{and} \quad v_{j\bar{\imath}}^* = 0 & \text{otherwise} \\ v_{\bar{\imath}}^* = 1 & \text{if} \quad d_{\bar{\imath}} \neq 0 & \text{and} \quad v_{\bar{\imath}}^* = 0 & \text{otherwise} \\ \tilde{v}_j^* = 1 & \text{if} \quad \tilde{d}_j \neq 0 & \text{and} \quad \tilde{v}_j^* = 0 & \text{otherwise} \end{array}$$

we will denote :

$$n_{\tilde{\imath}}^{*} = \sum_{j=1}^{n} v_{j\tilde{\imath}}^{*} \qquad \tilde{m}_{j}^{*} = \sum_{i=1}^{m} v_{j\tilde{\imath}}^{*}$$

$$n^{*} = \sum_{j=1}^{n} \tilde{v}_{j}^{*} \qquad \tilde{m}^{*} = \sum_{i=1}^{m} v_{i}^{*}$$

and

$$(mn)^{*} = \sum_{i=1}^{m} \sum_{j=1}^{n} v_{j\tilde{\imath}}^{*}$$

Integer $n_{\tilde{\imath}}^{*}$ is the number of impostors for which there is at least one test utterance against registered speaker X_i, integer $\tilde{m}_j^{*}$ represents the number of registered speakers against which there is at least one test utterance from impostor Y_j, n^* is the number of impostors from which there is at least one impostor test utterance, $\tilde{m}^*$ is the number of registered speakers against which there is at least one impostor test utterance and $(mn)^*$ denotes the total number of couples (X_i, Y_j) for which there is at least one impostor test utterance (from Y_j against X_i).

Finally, we will denote as $\tilde{M}$ the set of *male* impostor speakers, as $\tilde{F}$ the set of *female* impostors speakers, and

$\tilde{m}_M^{*}$: the number of male registered speakers against which there is at least one impostor test utterance

$\tilde{m}_F^{*}$: the number of female registered speakers against which there is at least one impostor test utterance

$(m_M n)^*$: the number of couples $(X_i \in M, Y_j)$ for which there is at least one impostor test utterance (from Y_j against X_i, X_i being a male speaker)

$(m_F n)^*$: the number of couples $(X_i \in F, Y_j)$ for which there is at least one impostor test utterance (from Y_j against X_i, X_i being a female speaker)

12.4.2 Closed-set identification

A closed-set identification system can be viewed as a function which assigns, to any test utterance z, an estimated speaker index $\hat{\imath}\,(z)$, corresponding to the identified speaker $X_{\hat{\imath}\,(z)}$ in the set of registered speakers.

In closed-set identification, all test utterances belong to one of the registered speakers. Therefore, a *misclassification error* occurs for test utterance number k produced by speaker X_i when :

$$\hat{\imath}\,(x_i^k) \neq i \quad \text{or equivalently} \quad \delta\,[\hat{\imath}(x_i^k), i] = 0$$

where δ denotes the Kronecker function.

12.4.2.1 Misclassification rates

The most natural figure that indicates the performance of a speaker identification system is the relative number of times the system fails in identifying correctly an applicant speaker; in other words, how often a test utterance will be assigned an erroneous identity. Whereas it is straightforward to calculate a performance figure on a speaker-by-speaker basis, care should be taken when deriving a global score.

With our notation, and assuming that $c_i \neq 0$, we define the *misclassification rate* for speaker X_i as :

$$\gamma_i \;=\; 1 \,-\, \frac{1}{c_i} \sum_{k=1}^{c_i} \delta\,[\hat{\imath}(x_i^k), i] \tag{12.1}$$

If we denote as $\mathcal{P}\,(\hat{X}_{j\neq i}\,|\,X_i)$ the probability that the system under test identifies speaker X_i when he is *not* the actual speaker, the quantity γ_i provides an estimate of this probability, whereas $1-\gamma_i$ provides an estimate of $\mathcal{P}\,(\hat{X}_i\,|\,X_i)$. However, it is preferable to report error scores rather than success scores, and performance improvements should be measured as relative error rate reduction[38]. If $c_i = 0$, γ_i is undefined but $p_i\,\gamma_i = 0$.

We suggest to use the term *dependable speaker* to qualify a registered speaker with a low misclassification rate[39], and the term *unreliable speaker* for a speaker with a h misclassification rate[40].

From speaker-by-speaker figures, the *average misclassification rate* can be derived as :

$$\bar{\gamma} \;=\; \frac{1}{m^*} \sum_{\substack{i=1 \\ c_i \neq 0}}^{m} \gamma_i \tag{12.2}$$

and by computing separately :

$$\bar{\gamma}_M \;=\; \frac{1}{m^*_M} \sum_{\substack{i=1 \\ X_i \in M \\ c_i \neq 0}}^{m} \gamma_i \quad \text{and} \quad \bar{\gamma}_F \;=\; \frac{1}{m^*_F} \sum_{\substack{i=1 \\ X_i \in F \\ c_i \neq 0}}^{m} \gamma_i \tag{12.3}$$

the *gender-balanced misclassification rate* can be obtained as :

$$\bar{\gamma}_{MF} \;=\; \frac{1}{2}\,(\bar{\gamma}_M + \bar{\gamma}_F) \tag{12.4}$$

The previous scores are different from the *test set misclassification rate*, calculated as :

$$\gamma \;=\; 1 \,-\, \frac{1}{c} \sum_{i=1}^{m} \sum_{k=1}^{c_i} \delta\,[\hat{\imath}(x_i^k), i] \;=\; \sum_{i=1}^{m} p_i\,\gamma_i \tag{12.5}$$

Scores $\bar{\gamma}$ and γ are formally identical if and only if c_i does not depend on i, i.e when the test set contains an identical number $c_0 = c/m$ test utterances for each speaker. As it is usually observed that speaker recognition performances may vary with the speaker's gender, the comparison of $\bar{\gamma}$ and $\bar{\gamma}_{MF}$ can show significant differences, if the registered population is not gender-balanced. Therefore, we believe that an accurate description of the identification performance requires the 3 numbers $\bar{\gamma}_{MF}$, $\bar{\gamma}$, and γ to be provided.

12.4.2.2 Mistrust rates

Taking an other point of view, performance scores can be designed to measure how reliable is the decision of the system when it has assigned a given identity; in other words, to provide an estimate of the probability $\mathcal{P}\,(X_{j\neq i}\,|\,\hat{X}_i)$, i.e the probability that the speaker is *not* really X_i when the system under test has output X_i as the most likely identity.

[38] In fact, $log_{10}\,\gamma_i$ would be the most relevant figure, especially for small error rates, but the intuitive meaning of the logarithm of an error rate is less immediate. This remark also stands for mistrust rates.

[39] Sometimes called familiarily *a sheep*.

[40] Sometimes called familiarily *a goat*.

To define the *mistrust rate*, we have to introduce the following notation :

$$\hat{c}_i = \sum_{k=1}^{c} \delta\left[\hat{\imath}(x^k), i\right] \quad \text{and} \quad \hat{p}_i = \frac{\hat{c}_i}{c}$$

$$\hat{m}^* = \sum_{i=1}^{m} \hat{m}_i^* \quad \text{where} \quad \hat{m}_i^* = 1 \ \text{ if } \ \hat{c}_i \neq 0 \ \text{ and } \ \hat{m}_i^* = 0 \ \text{ otherwise}$$

By definition, $\hat{c}_i$ and $\hat{p}_i$ are respectively the number and proportion of test utterances identified as X_i over the whole test set, while $\hat{m}^*$ is the number of registered speakers whose identity was assigned at least once to a test utterance.

The *mistrust rate* for speaker X_i can then be computed (for $\hat{c}_i \neq 0$) as :

$$\hat{\gamma}_i \;=\; 1 - \frac{1}{\hat{c}_i} \sum_{k=1}^{c_i} \delta\left[\hat{\imath}(x_i^k), i\right] \;=\; 1 - (1-\gamma_i)\,\frac{c_i}{\hat{c}_i} \tag{12.6}$$

Here again, if $\hat{c}_i = 0$, $\hat{\gamma}_i$ is undefined, but $\hat{p}_i\,\hat{\gamma}_i = 0$.

We suggest that the term *resistant speaker* could be used to qualify a registered speaker with a low mistrust rate[41], and the term *vulnerable speaker* for a speaker with a high mistrust rate[42].

From this speaker-by-speaker score, the *average mistrust rate* can be derived as :

$$\bar{\hat{\gamma}} \;=\; \frac{1}{\hat{m}^*} \sum_{\substack{i=1 \\ \hat{c}_i \neq 0}}^{m} \hat{\gamma}_i \tag{12.7}$$

and by computing separately[43] :

$$\bar{\hat{\gamma}}_M \;=\; \frac{1}{\hat{m}_M^*} \sum_{\substack{i=1 \\ X_i \in M \\ \hat{c}_i \neq 0}}^{m} \hat{\gamma}_i \quad \text{and} \quad \bar{\hat{\gamma}}_F \;=\; \frac{1}{\hat{m}_F^*} \sum_{\substack{i=1 \\ X_i \in F \\ \hat{c}_i \neq 0}}^{m} \hat{\gamma}_i \tag{12.8}$$

the *gender-balanced mistrust rate* is defined as :

$$\bar{\hat{\gamma}}_{MF} \;=\; \frac{1}{2}\,\left(\bar{\hat{\gamma}}_M + \bar{\hat{\gamma}}_F\right) \tag{12.9}$$

By noticing now that :

$$\sum_{i=1}^{m} \hat{p}_i\,\hat{\gamma}_i \;=\; \sum_{i=1}^{m} p_i\,\gamma_i$$

there appears to be no need to define a test set mistrust rate $\hat{\gamma}$. In other words : *the test set mistrust rate is equal to the test set misclassification rate*[44].

From a practical point of view, misclassification rates and mistrust rates can be obtained by the exact same scoring programs, operating successively on the confusion matrix and on its transpose.

[41] The familiar term of *ram* could be used to extend the ovine analogy...

[42] The familiar term of *lamb* seems appropriate here...

[43] With $\hat{m}_M^*$ and $\hat{m}_F^*$ being respectively the number of male and female speakers whose identity was assigned at least one to a test utterance.

[44] This property can be interesting to debug a scoring program.

12.4.2.3 Confidence ranks

Most speaker identification systems use a similarity measure between a test utterance and all training patterns to decide, by a nearest neighbour decision rule, which is the identity of the test speaker. In this case, for a test utterance x, an ordered list of registered speaker can be produced :

$$\hat{L}\,(x) = [X_{l_1(x)} \dots X_{l_m(x)}] = [X_{l_j(x)}]_{1 \le l_j(x) \le m}$$

where, for all index J, $X_{l_j(x)}$ is judged closer to the test utterance than $X_{l_{j+1}(x)}$ is.

The *identification rank* of the genuine speaker of utterance x_i^k can then be expressed as :

$$\rho\,(x_i^k) \;=\; \sum_{j=1}^{m} j \cdot \delta\,[l_j(x_i^k), i] \quad \text{or equivalently} \quad \rho\,(x_i^k) \;=\; \underset{j}{\operatorname{argmax}}\;\; \delta\,[l_j(x_i^k), i] \tag{12.10}$$

In other words, ρ is the position at which appears the correct speaker in the ordered list of neighbours of the test utterance. Note that a correct identification of x_i^k corresponds to $\rho\,(x_i^k) = 1$.

Under the assumption that $c_i \neq 0$, let us now denote :

$$\eta_{(r)}\,(x_i^k) = \sum_{j=1}^{r} \delta\,[l_j(x_i^k), i] \quad \text{and} \quad \eta_{i\,(r)} = \frac{1}{c_i} \sum_{k=1}^{c_i} \eta_{(r)}\,(x_i^k) \tag{12.11}$$

$$\text{i.e :} \qquad \eta_{(r)}\,(x_i^k) = 1 \;\;\text{ if }\;\; \rho\,(x_i^k) \le r \qquad \text{and} \qquad 0 \;\;\text{otherwise}$$

The $\alpha\,\%$ *confidence rank* for speaker X_i, which we will denote here as $\rho_{i\,(\alpha\,\%)}$ can then be defined as the smallest integer number for which $\alpha\,\%$ of the test utterances belonging to speaker X_i are part of the ρ_i nearest neighbours in the ordered list of candidates. Hence the formulation :

$$\rho_{i\,(\alpha\,\%)} \;=\; \min_{r} \left\{ r \;\middle|\; \eta_{i\,(r)} \;\ge \alpha\,\% \right\} \tag{12.12}$$

Then, the *average* $\alpha\,\%$ *confidence rank* can be computed as the average of $\rho_{i\,(\alpha\,\%)}$ over all registered speakers (for which $c_i \neq 0$) :

$$\bar{\rho}_{(\alpha\,\%)} \;=\; \frac{1}{m^*} \sum_{\substack{i\,=\,1 \\ c_i\,\neq\,0}}^{m} \rho_{i\,(\alpha\,\%)} \tag{12.13}$$

Though a gender-balanced confidence rank could be defined analogously to gender-balanced misclassification and mistrust rates, the relevance of this figure is not clear.

If finally we denote :

$$\eta_{(r)} \;=\; \sum_{i=1}^{m} \sum_{k=1}^{c_i} \eta_{(r)}\,(x_i^k) \;=\; \sum_{i=1}^{m} p_i\,\eta_{i\,(r)} \tag{12.14}$$

the *test set* $\alpha\,\%$ *confidence rank* is defined as :

$$\rho_{(\alpha\,\%)} \;=\; \min_{r} \left\{ r \;\middle|\; \eta_{(r)} \ge \alpha\,\% \right\} \tag{12.15}$$

12.4.2.4 Comments

Average scores, gender-balanced scores and test set scores all fall under a same formalism. If we denote as π_i a certain quantity which we will call the *relative representativity* of speaker X_i and which satisfies $\sum \pi_i = 1$, and if we now consider the linear combination :

$$\gamma_\pi \;=\; \sum_{i=1}^{m} \pi_i\,\gamma_i \tag{12.16}$$

it is clear that :

$$\gamma_\pi = \bar{\gamma} \qquad \text{for } \pi_i = \frac{1}{m^*}$$

$$\gamma_\pi = \bar{\gamma}_{MF} \qquad \text{for } \pi_i = \frac{1}{2m^*_M} \text{ if } X_i \in M \quad \text{and} \quad \pi_i = \frac{1}{2m^*_F} \text{ if } X_i \in F$$

$$\gamma_\pi = \gamma \qquad \text{for } \pi_i = p_i$$

Therefore $\bar{\gamma}$, $\bar{\gamma}_{MF}$ and γ correspond to different estimates of a global score, under various assumptions on the relative representativity of each test speaker[45]. For average scores, it is assumed that each speaker is equally representative, irrespectively of its sex group, but if the test population is strongly unbalanced this hypothesis may not be relevant (unless there is a reason for it). For gender-balanced scores, each test speaker is supposed to be representative of its sex group, and each sex group is supposed to be equiprobable. Test set scores make the assumption that each test speaker has a representativity which is proportional to its number of test utterances, which is certainly not always a meaningful hypothesis.

Test set scores can therefore be used as an overall performance measure if the test data represent accurately the profile and behaviour of the user population, i.e both in terms of population composition and individual frequency of use. If only the composition of the test set population is representative of the general user population, average scores allow to neutralise the possible discrepancies in number of utterances per speaker. If finally the composition of the test set spekaer population is not representative, gender-balanced score provide a general purpose estimate.

If there is a way to estimate separately the relative representativity for each test registered speaker, a *representative misclassification rate* can be computed as in equation (12.16). Conversely, some techniques such as those used in public opinion polls can be resorted to in order to select a representative test population, when setting up an evaluation experiment.

12.4.2.5 Example

Figures 12.1 and 12.2 give examples of misclassification rates, mistrust rates and confidence ranks. However, it must be kept in mind that the number of tests used to design these examples is too small to guarantee any statistical validity of the figures.

12.4.3 Verification

A verification system can be viewed as a function which assigns, to a test utterance z and a claimed identity $\tilde{\imath}$, a boolean value $\hat{v}_{\tilde{\imath}}(z)$, which is equal to 1 if the utterance is accepted, and 0 if it is rejected.

Two type of errors can then occur. Either a genuine speaker is rejected, or an impostor is accepted. Hence, a *false rejection* corresponds to :

$$\hat{v}_{\tilde{\imath}}(x_i^k) = 0$$

and a *false acceptance* happens if :

$$\hat{v}_{\tilde{\imath}}(y_{j\tilde{\imath}}^k) = 1$$

In the rest of this section, we will denote as follows the events :

- $\hat{V}$: the system accepts the applicant speaker

[45]This remark is also true concerning mistrust rates, with the estimates of π_i being obtained from the output of the identification system.

$m=5$	X_1	X_2	X_3	X_4	X_5	$\hat{c}_i$	$\hat{\gamma}_i$	
	Male speakers			Female speakers				
$\hat{X}_1$	4	-	-	-	-	4	0 %	$\hat{m}^* = 5$
$\hat{X}_2$	-	0	$\boxed{1}$	-	$\boxed{1}$	2	100 %	$\boxed{\bar{\hat{\gamma}} = 30\,\%}$
$\hat{X}_3$	-	-	1	-	-	1	0 %	$\hat{m}^*_M = 3$ $\hat{m}^*_F = 2$
$\hat{X}_4$	$\boxed{4}$	$\boxed{1}$	-	5	-	10	50 %	$\bar{\hat{\gamma}}_M \approx 33\%$ $\bar{\hat{\gamma}}_F \approx 25\%$
$\hat{X}_5$	-	-	-	-	3	3	0 %	$\boxed{\bar{\hat{\gamma}}_{MF} \approx 29\,\%}$
c_i	8	1	2	5	4	$c = 20$		
γ_i	50 %	100 %	50 %	0 %	25 %			

$m^* = 5$

$\boxed{\bar{\gamma} = 45\,\%}$

$\hat{m}^*_M = 3$ $\qquad$ $\hat{m}^*_F = 2$

$\bar{\gamma}_M \approx 67\,\%$ $\qquad$ $\bar{\gamma}_F \approx 13\,\%$

$\boxed{\bar{\gamma}_{MF} \approx 40\,\%}$

$\boxed{\gamma \equiv \hat{\gamma} = \frac{7}{20} = 35\,\%}$

Figure 12.1: An example illustrating how to score misclassification rates and mistrust rates from a confusion matrix, in speaker identification. Here X_4 would be a *dependable* speaker, X_2 an *unreliable* speaker, X_1, X_3 and X_5 would appear as *resistant* speakers and X_2 (him again !) would seem to be a *vulnerable* speaker.

- $\hat{\bar{V}}$: the system rejects the applicant speaker
- X : the applicant speaker is a genuine speaker
- Y : the applicant speaker is an impostor ($Y \equiv \bar{X}$)
- i or j : the identity of the applicant speaker is X_i or Y_j
- $\tilde{\imath}$: the claimed identity is i

We first address aspects of *static evaluation*, that is what meaningful figures can be computed to measure the performance of a system over which the experimentator has absolutely no control. Then, after discussing the role of *decision thresholds*, we review several approaches that allow to obtain a *dynamic evaluation* of the system, i.e in a relatively threshold-independent manner.

12.4.3.1 False rejection rates

If $c_i \neq 0$, the *false rejection rate* for speaker X_i is defined quite naturally as :

$$\alpha_i \;=\; 1 \;-\; \frac{1}{c_i} \sum_{k=1}^{c_i} \hat{v}_{\bar{i}}\,(x_i^k) \tag{12.17}$$

Rate α_i provides an estimate of $\mathcal{P}(\hat{\bar{V}} \mid X, i)$, i.e the probability that the system makes a diagnostic of rejection, given that the applicant speaker is the authorised speaker X_i (claiming his own identity). If $c_i = 0$, α_i is undefined but $p_i\,\alpha_i = 0$.

As for closed-set identification, the terms *dependable* speaker and *unreliable* speaker can be used to qualify speakers with a low or (respectively) high false rejection rate.

From speaker based figures, the *average false rejection rate* can be obtained as :

$$\bar{\alpha} \;=\; \frac{1}{m^*} \sum_{\substack{i\,=\,1 \\ c_i\,\neq\,0}}^{m} \alpha_i \tag{12.18}$$

while the *gender-balanced false rejection rate* is :

$$\bar{\alpha}_{MF} \;=\; \frac{1}{2}\,(\bar{\alpha}_M + \bar{\alpha}_F) \tag{12.19}$$

where :

$$\bar{\alpha}_M \;=\; \frac{1}{m^*_M} \sum_{\substack{i\,=\,1 \\ X_i\,\in\,M \\ c_i\,\neq\,0}}^{m} \alpha_i \quad \text{and} \quad \bar{\alpha}_F \;=\; \frac{1}{m^*_F} \sum_{\substack{i\,=\,1 \\ X_i\,\in\,F \\ c_i\,\neq\,0}}^{m} \alpha_i \tag{12.20}$$

The *test set false rejection rate* is calculated as :

$$\alpha \;=\; 1 \;-\; \frac{1}{c} \sum_{i=1}^{m} \sum_{k=1}^{c_i} \hat{v}_{\bar{i}}\,(x_i^k) \;=\; \sum_{i=1}^{m} p_i\,\alpha_i \tag{12.21}$$

Rates $\bar{\alpha}$, $\bar{\alpha}_{MF}$ and α provide 3 different estimates of $\mathcal{P}\,(\hat{\bar{V}} \mid X)$. Rate α is influenced by the test set distribution of genuine attempts which may only be artefactual. The issue of "representativity" of a test speaker, raised in section 12.4.2.1, is also relevant here, and for false acceptance rates as well.

12.4.3.2 False acceptance rates and imposture rates

As opposed to false rejection, there are several ways to score false acceptance, depending on whether it is the vulnerability of registered speaker which is considered or the skills of impostors. Moreover, the way to evaluate false acceptance rates and imposture rates depends on whether the identity of each impostor is known or not.

Known impostors

If the impostor identities are known, the false acceptance rate <u>in favour of</u> impostor Y_j <u>against</u> registered speaker X_i can be defined, for $d_{j\bar{i}} \neq 0$, as :

$$\beta_{j\bar{i}} \;=\; \frac{1}{d_{j\bar{i}}} \sum_{k=1}^{d_{j\bar{i}}} \hat{v}_{\bar{i}}\,(y_{j\bar{i}}^k) \tag{12.22}$$

Here, $\beta_{j\tilde{\imath}}$ can be viewed as an estimate of $\mathcal{P}(\hat{V} \mid Y, j, \tilde{\imath})$, i.e the probability that the system makes a diagnostic of acceptance, given that the applicant speaker is the impostor Y_j claiming identity X_i.

Then, the *average false acceptance rate* against speaker X_i can be obtained (if $n^*_{\tilde{\imath}} \neq 0$) by averaging the false acceptances over all impostors :

$$\beta_{\tilde{\imath}} \;=\; \frac{1}{n^*_{\tilde{\imath}}} \sum_{\substack{j=1 \\ d_{j\tilde{\imath}} \neq 0}}^{n} \beta_{j\tilde{\imath}} \tag{12.23}$$

and similarily the *average imposture rate* in favour of impostor Y_j can be calculated (for $\tilde{m}^*_j \neq 0$) as :

$$\tilde{\beta}_j \;=\; \frac{1}{\tilde{m}^*_j} \sum_{\substack{i=1 \\ d_{j\tilde{\imath}} \neq 0}}^{m} \beta_{j\tilde{\imath}} \tag{12.24}$$

Rates $\beta_{\tilde{\imath}}$ and $\tilde{\beta}_j$ provide (respectively) estimates of $\mathcal{P}(\hat{V} \mid Y, \tilde{\imath})$ and $\mathcal{P}(\hat{V} \mid Y, j)$ under the assumption that all impostors and all claimed identities are equiprobable. The number $\beta_{\tilde{\imath}}$ indicates the false acceptance rate obtained in average by each impostor in claiming identity X_i, while $\tilde{\beta}_j$ indicates the success rate of impostor Y_j in claiming an identity averaged over each claimed identity. A registered speaker can be more or less *resistant* (low $\beta_{\tilde{\imath}}$) or *vulnerable* (high $\beta_{\tilde{\imath}}$), whereas impostors with a high $\tilde{\beta}_j$ can be viewed as *skilled impostors*[46], as opposed to *poor impostors*[47] for those with a low $\tilde{\beta}_j$.

The *average false acceptance rate* which is equal to the *average imposture rate* is obtained as :

$$\bar{\beta} \quad = \quad \frac{1}{(mn)^*} \sum\sum_{\substack{(i,j)=(1,1) \\ d_{j\tilde{\imath}} \neq 0}}^{(m,n)} \beta_{j\tilde{\imath}} \quad = \quad \bar{\tilde{\beta}} \tag{12.25}$$

i.e as the average of the false acceptances over all couples (X_i, Y_j)[48], which provides an estimate of $\mathcal{P}(\hat{V} \mid Y)$ under the assumption that all couples (X_i, Y_j) are equally likely.

Here, separate estimates of the average false acceptance rate on the male and female registered populations can be obtained as :

$$\bar{\beta}_M \quad = \quad \frac{1}{(m_M n)^*} \sum\sum_{\substack{(i,j)=(1,1) \\ X_i \in M \\ d_{j\tilde{\imath}} \neq 0}}^{(m,n)} \beta_{j\tilde{\imath}} \quad \text{and} \quad \bar{\beta}_F \quad = \quad \frac{1}{(m_F n)^*} \sum\sum_{\substack{(i,j)=(1,1) \\ X_i \in F \\ d_{j\tilde{\imath}} \neq 0}}^{(m,n)} \beta_{j\tilde{\imath}} \tag{12.26}$$

and a *gender-balanced false acceptance rate* is provided by :

$$\bar{\beta}_{MF} = \frac{1}{2}\,(\bar{\beta}_M + \bar{\beta}_F) \tag{12.27}$$

The question could be raised of whether it is desirable to compute a score which would provide an estimation of the false acceptance rate for a gender-balanced impostor population. We propose not to go that far, as it would clearly lead to duplication of scoring figures, but the influence of impostors' gender could be partly neutralised by the experimental design :

[46] *Wolves*, in our animal analogy.

[47] *Badgers* seems particularly appropriate !

[48] Another way of estimating the average false acceptance rate and the average imposture rate could consist in computing :

$$\bar{\beta}^{(1)} \;=\; \frac{1}{\tilde{m}^*} \sum_{\substack{i=1 \\ d_{\tilde{\imath}} \neq 0}}^{m} \beta_{\tilde{\imath}} \qquad \text{and} \qquad \bar{\tilde{\beta}}^{(1)} \;=\; \frac{1}{n^*} \sum_{\substack{j=1 \\ \tilde{d}_j \neq 0}}^{n} \tilde{\beta}_j$$

but these estimates have the drawback that, in general, $\bar{\beta}^{(1)} \neq \bar{\tilde{\beta}}^{(1)}$.

- If the impostor population is composed of *acquainted intentional* impostors, the issue of impostor's gender balancing can be considered as relatively marginal, even though the impostor of a given sex may be more skilled in imitating a same-sex person than somebody of the opposite sex.

- If the impostor population is composed of *casual* impostors, we propose to restrict systematically the impostor utterance test set to same-sex trials. However, as we mentioned it earlier, it is safer to check in an independent experiment, that the system is really robust to cross-sex casual impostors.

It may also be interesting to calculate imposture rates regardless of the claimed identities. In this case, we define the *imposture rate* in favour of impostor Y_j *regardless of the claimed identity* as :

$$\tilde{\beta}'_j \;=\; \frac{1}{\tilde{d}_j} \sum_{k=1}^{\tilde{d}_j} \hat{v}_{\bar{\imath}}\,(y_j^k) \tag{12.28}$$

and the *average imposture rate regardless of the claimed identity* as :

$$\bar{\tilde{\beta}}' \;=\; \frac{1}{n^*} \sum_{\substack{j\,=\,1 \\ \tilde{d}_j\,\neq\,0}}^{n} \tilde{\beta}'_j \tag{12.29}$$

However, $\beta_{j\bar{\imath}}$, $\beta_{\bar{\imath}}$, $\tilde{\beta}_j$, $\tilde{\beta}'_j$, $\bar{\beta}$, $\bar{\beta}_{MF}$ and $\bar{\tilde{\beta}}'$ can not be evaluated when the identities of impostors are not known. In this case false acceptance rates and imposture rates can be calculated under the assumption that all impostor test utterances are produced by distinct impostors.

Unknown impostors

The *false acceptance rate* against speaker X_i *assuming distinct impostors* can be obtained (if $d^*_{\bar{\imath}} \neq 0$) as :

$$\beta'_{\bar{\imath}} \;=\; \frac{1}{d_{\bar{\imath}}} \sum_{k=1}^{d_{\bar{\imath}}} \hat{v}_{\bar{\imath}}\,(y_{\bar{\imath}}^k) \tag{12.30}$$

and the *average false acceptance rate assuming distinct impostors* is defined as :

$$\bar{\beta}' \;=\; \frac{1}{\tilde{m}^*} \sum_{\substack{i\,=\,1 \\ d_{\bar{\imath}}\,\neq\,0}}^{m} \beta'_{\bar{\imath}} \tag{12.31}$$

Here again, separate estimates of the average false acceptance rate assuming distinct impostors, on the male and female registered populations can be obtained as :

$$\bar{\beta}'_M \;=\; \frac{1}{\tilde{m}^*_M} \sum_{\substack{i\,=\,1 \\ X_i\,\in\,M \\ d_{\bar{\imath}}\,\neq\,0}}^{m} \beta'_{\bar{\imath}} \quad \text{and} \quad \bar{\beta}'_F \;=\; \frac{1}{\tilde{m}^*_F} \sum_{\substack{i\,=\,1 \\ X_i\,\in\,F \\ d_{\bar{\imath}}\,\neq\,0}}^{m} \beta'_{\bar{\imath}} \tag{12.32}$$

with the *gender-balanced false acceptance rate assuming distinct impostors* being :

$$\bar{\beta}'_{MF} = \frac{1}{2}\,(\bar{\beta}'_M + \bar{\beta}'_F) \tag{12.33}$$

Rate $\beta'_{\bar{\imath}}$ provides a speaker-dependent estimate of $\mathcal{P}\,(\hat{V}\,|\,Y, \bar{\imath})$ assuming distinct impostors. Rate $\bar{\beta}'$ can be viewed as an estimate of $\mathcal{P}\,(\hat{V}\,|\,Y)$ under the assumptions of distinct impostors and that all claimed identities are equally likely while $\bar{\beta}'_{MF}$ can be understood as another estimate of $\mathcal{P}\,(\hat{V}\,|\,Y)$

under the assumptions of distinct impostors, that attempts against male speakers and against female speakers are equiprobable, and that within a gender class all claimed identities are equally likely.

Test set scores

If finally false acceptances are scored globally, regardless of the impostor identity nor of the claimed identity, we obtain the *test set false acceptance rate* which is identical to the *test set imposture rate* :

$$\beta \; = \; \frac{1}{d} \sum_{k=1}^{d} \hat{v}\,(y^k) \; = \; \frac{1}{d} \sum_{i=1}^{m} \sum_{k=1}^{d_{\tilde{\imath}}} \hat{v}_{\tilde{\imath}}\,(y_{\tilde{\imath}}^k) \; = \; \tilde{\beta} \tag{12.34}$$

Here, $\beta \equiv \tilde{\beta}$ provides a test set estimate of $\mathcal{P}\,(\hat{V}\,|\,Y)$ which is biased by the composition of the registered population and a possible uneveness of the number of impostor trials for each speaker[49].

Summary

For scoring false acceptance rates, we believe that, beside β, it is necessary to report on $\bar{\beta}$ and $\bar{\beta}_{MF}$ (when impostors are known) or $\bar{\beta}'$ and $\bar{\beta}'_{MF}$ (when they are not known), as the score β may be significantly influenced by the test data distribution. The other scores described in this section are mainly useful for diagnostic analysis.

12.4.3.3 Relative unreliability, vulnerability and imitation ability

It can also be of major interest to estimate the contribution of a given registered speaker X_i to the overall false rejection rate, which can be denoted as $\mathcal{P}\,(i\,|\,\hat{\bar{V}}, X)$, i.e the probability that the identity of the speaker was i given that a (false) rejection diagnostic was made on a genuine speaker (claiming his own identity).

We can thus define the *average relative unreliability* for speaker X_i as :

$$\bar{\mu}_i \; = \; \frac{\alpha_i}{\sum_i \, \alpha_i} = \frac{1}{m^*} \, \frac{\alpha_i}{\bar{\alpha}} \tag{12.35}$$

or his *test set relative unreliability* :

$$\mu_i \; = \; \frac{p_i \, \alpha_i}{\sum_i \, p_i \, \alpha_i} = p_i \, \frac{\alpha_i}{\alpha} \tag{12.36}$$

By construction :

$$\sum_i \bar{\mu}_i \; = \; \sum_i \mu_i \; = \; 1$$

From a different angle, the relative vulnerability for a given registered speaker X_i (i.e $\mathcal{P}\,(\tilde{\imath}\,|\,\hat{V}, Y)$) can be measured as his contribution to the false acceptance rate.

Thus, the *average relative vulnerability* for speaker X_i can be defined as :

$$\bar{\nu}_{\tilde{\imath}} \; = \; \frac{\sum\limits_j \beta_{j\tilde{\imath}}}{\sum\limits_i \sum\limits_j \beta_{j\tilde{\imath}}} \; = \; \frac{n_{\tilde{\imath}}^*}{(mn)^*} \, \frac{\beta_{\tilde{\imath}}}{\bar{\beta}} \tag{12.37}$$

[49] Note the relations :

$$\beta'_{\tilde{\imath}} \; = \; \sum_{j=1}^{n} \frac{q_{j\tilde{\imath}}}{q_{\tilde{\imath}}} \, \beta_{j\tilde{\imath}} \;\; , \quad \tilde{\beta}'_j \; = \; \sum_{i=1}^{m} \frac{q_{j\tilde{\imath}}}{\tilde{q}_j} \, \beta_{j\tilde{\imath}} \quad \text{and} \quad \beta \; = \; \sum_{i=1}^{m} \sum_{j=1}^{n} q_{j\tilde{\imath}} \, \beta_{j\tilde{\imath}} = \sum_{i=1}^{m} q_{\tilde{\imath}} \, \beta'_{\tilde{\imath}} = \sum_{j=1}^{n} q_j \, \tilde{\beta}'_j$$

his *relative vulnerability assuming distinct impostors*, as :

$$\bar{\nu}'_{\bar{\imath}} = \frac{\beta'_{\bar{\imath}}}{\sum_i \beta'_{\bar{\imath}}} = \frac{1}{\tilde{m}^*} \frac{\beta'_{\bar{\imath}}}{\bar{\beta}'} \tag{12.38}$$

and his *test set relative vulnerability* as :

$$\nu_{\bar{\imath}} = \frac{\sum_j q_{j\bar{\imath}} \beta_{j\bar{\imath}}}{\sum_i \sum_j q_{j\bar{\imath}} \beta_{j\bar{\imath}}} = q_{\bar{\imath}} \frac{\beta'_{\bar{\imath}}}{\beta} \tag{12.39}$$

Here :

$$\sum_i \bar{\nu}_{\bar{\imath}} = \sum_i \bar{\nu}'_{\bar{\imath}} = \sum_i \nu_{\bar{\imath}} = 1$$

Finally, by considering the relative success of impostor Y_j, i.e $\mathcal{P}\,(j \,|\, \hat{V}, Y)$, we define in a dual way as above, the *average imitation ability* of impostor Y_j :

$$\bar{\tilde{\nu}}_j = \frac{\sum_i \beta_{j\bar{\imath}}}{\sum_j \sum_i \beta_{j\bar{\imath}}} = \frac{\tilde{m}^*_j}{(mn)^*} \frac{\tilde{\beta}_j}{\bar{\beta}} \tag{12.40}$$

his *imitation ability regardless of the claimed identity* :

$$\bar{\tilde{\nu}}'_j = \frac{\tilde{\beta}'_j}{\sum_j \tilde{\beta}'_j} = \frac{1}{n^*} \frac{\tilde{\beta}'_j}{\bar{\tilde{\beta}}'} \tag{12.41}$$

and his *test set relative imitation ability* :

$$\tilde{\nu}_j = \frac{\sum_i q_{j\bar{\imath}} \beta_{j\bar{\imath}}}{\sum_j \sum_i q_{j\bar{\imath}} \beta_{j\bar{\imath}}} = \tilde{q}_j \frac{\tilde{\beta}'_j}{\beta} \tag{12.42}$$

Naturally :

$$\sum_j \bar{\tilde{\nu}}_j = \sum_j \bar{\tilde{\nu}}'_j = \sum_j \tilde{\nu}_j = 1$$

The relative unreliability and vulnerability can also be calculated relatively to the male/female population.

12.4.3.4 Comments

As for misclassification rates, the gender-balanced, average and test set false rejection rates as well as the gender-balanced and average false acceptance rates assuming distinct impostors and the test set false acceptance rate correspond to different estimates of a global score, under various assumptions on the relative representativity of each genuine test speaker. The discussion of section 12.4.2.4 can be readily generalised.

For what concerns gender-balanced and average false acceptance rates with known impostors, a relative representativity $\kappa_{j\bar{\imath}}$ can be defined for each couple of registered speaker and impostor (X_i, Y_j) (with $\sum \sum \kappa_{j\bar{\imath}} = 1$), and if we denote :

$$\beta_\kappa = \sum_{i=1}^{m} \sum_{j=1}^{n} \kappa_{j\bar{\imath}} \beta_{j\bar{\imath}}$$

we have :

$$\beta_\kappa = \bar{\beta} \qquad \text{for } \kappa_{j\tilde{\imath}} = \frac{1}{(mn)^*}$$

$$\beta_\kappa = \bar{\beta}_{MF} \quad \text{for } \kappa_{j\tilde{\imath}} = \frac{1}{2(m_M n)^*} \text{ if } X_i \in M \quad \text{and} \quad \kappa_{j\tilde{\imath}} = \frac{1}{2(m_F n)^*} \text{ if } X_i \in F$$

In the case of casual impostors, choosing a selective attempt configuration towards same-sex speakers is equivalent to the assumption :

$$\left[X_i \in M \text{ and } Y_j \in \tilde{F}\right] \quad \text{or} \quad \left[X_i \in F \text{ and } Y_j \in \tilde{M}\right] \quad \Longrightarrow \quad \kappa_{j\tilde{\imath}} = 0$$

i.e that the representativity of a cross-sex attempt is zero.

Studies allowing to better define the representativity of impostor attempts against registered speaker would be of great help to increase the relevance of evaluation scores.

12.4.3.5 Example

Figure 12.3, 12.4 and 12.5 give examples of false acceptance rates, false rejection rates, and imposture rates, as well as unreliability, vulnerability and imitation ability. As for the closed-set identification examples, the number of tests used to design these examples is too small to guarantee any statistical validity.

12.4.3.6 Expected benefit

From now on, we will denote as α and β the false rejection and rejection rates, whichever exact estimate is really chosen.

If it is possible to get estimates of the following quantities :

- p, the probability that an applicant speaker is a genuine speaker,
- $B_X \geq 0$, the benefit of a true acceptance,
- $B_Y \geq 0$, the benefit of a true rejection,
- $C_X \geq 0$, the cost of a false rejection,
- $C_Y \geq 0$, the cost of a false acceptance,

the *expected benefit* $\mathcal{B}$ of a verification system with false rejection rate α and false acceptance rate β can be computed as :

$$\mathcal{B} = p \cdot ((1-\alpha)\, B_X \; - \; \alpha\, C_X) \; + \; (1-p) \cdot ((1-\beta)\, B_Y \; - \; \beta\, C_Y) \qquad (12.43)$$

In particular, when $p = 1 - p = \frac{1}{2}$ and $B_X = B_Y = C_X = C_Y = K$, the *equal-risk equal-cost expected benefit* is :

$$\mathcal{B}_{eq} = K\,(1 \; - \; (\alpha + \beta)) \qquad (12.44)$$

The expected benefit is usually a meaningful static evaluation figure for the potential clients of the technology. It must however be understood only as the average expected benefit for each user attempt. It does not take into account external factors such as the psychological impact of the system, its maintenance costs, etc...

12.4.3.7 Threshold setting

Speaker verification systems usually proceed in two steps. First, a matching score $s(z, X_i)$ is computed between the test utterance z and the reference model X_i corresponding to the claimed identity. Then, the value of the matching score is compared to a threshold θ_i, and a decision is taken as follows :

$$\text{if} \quad s(z, X_i) \geq \theta_i \quad \hat{v}_{\bar{i}}(z) = 1 \qquad \text{else} \quad \hat{v}_{\bar{i}}(z) = 0$$

In other words, verification is positive only if the match between the test utterance and the reference model (for the claimed identity) is close enough.

A distinction can be made depending on whether each registered speaker has his individual threshold or whether a single threshold is used which is common to all speakers. In other words, if θ_i depends on i, the system uses *speaker-dependent thresholds*, whereas if θ_i does not depend on i, the system uses a *speaker-independent threshold*[50]. We will denote as Θ the threshold vector $[\theta_i]_{1 \leq i \leq m}$, and as $\alpha(\Theta)$ and $\beta(\Theta)$ the false rejection and acceptance rates corresponding to Θ.

The values of Θ have an inverse impact on the false rejection rate and on the false acceptance rate. Thus, with a low θ_i, fewer genuine attempts from speaker X_i will be rejected but more impostors will be erroneously accepted as X_i. Conversely, if θ_i is increased, β_i will generally decrease, at the expense of an increasing α_i. The goal of *dynamic evaluation* is to provide a description of the system performance which is as independent as possible of the threshold values.

The setting of thresholds is conditioned to the specification of an operating constraint which expresses the compromise that has to be reached between the two types of errors. Among many possibilities, the most popular ones are :

- A specified false rejection rate α_0. If speaker-dependent thresholds are used, the thresholds are tuned so that the false rejection rate for each speaker is equal to α_0 whereas, with speaker-independent thresholds, the constraint is only satisfied in the average.
- A specified false acceptance rate β_0. Here also, the constraint can be satisfied for each speaker with speaker-dependent thresholds, or in the average for speaker-independent thresholds.
- The maximisation of the expected benefit. Once again, the corresponding α and β can be obtained by a speaker-by-speaker optimisation or on an average basis.
- An *equal error rate (or EER)* $\alpha = \beta = \epsilon$. In fact, this is the most popular constraint, as the equal error rate ε is felt as a simple way to summarise the overall performance of a system into a single figure. Moreover, for any threshold Θ :

  $$\begin{array}{llll} \text{if} & \alpha(\Theta) \geq \epsilon & \text{then} & \beta(\Theta) \leq \epsilon \\ \text{if} & \alpha(\Theta) \leq \epsilon & \text{then} & \beta(\Theta) \geq \epsilon \end{array}$$

  In most practical applications however, the equal error rate does not correspond to an interesting operating constraint.

Two procedures are classically used to set the thresholds : the *a priori threshold setting* procedure and the *a posteriori threshold setting* procedure.

When the *a priori* threshold setting procedure is implemented, the threshold vector $\Theta = \Theta_{prio}$ is estimated from a set of *tuning data* which can be either the training data themselves, or a new set of unseen data. Then, the false rejection and acceptance rates $\alpha(\Theta_{prio})$ and $\beta(\Theta_{prio})$ are estimated on a disjoint test set. Naturally, there must be no intersection between the tuning data set and the

[50] In theory, a system using speaker-dependent thresholds should never perform worse than the same system using a speaker-independent threshold, as the latter is finally a particular case of the former. However, difficulties may arise in obtaining reliable estimates of each threshold, in the speaker-dependent approach.

test data set. Not only must the speech material of genuine attempts and impostor attempts be different between these two sets, but also the bundle of pseudo-impostors used to tune the threshold for a registered speaker should not contain any of the impostors which will be tested against this very speaker within the test set. Of course, the volume of additional speech data used for threshold setting must be counted as training material, when reporting on the *training speech quantity.*

When the *a posteriori* threshold setting procedure is adopted, $\Theta = \Theta_{post}$ is set on the test data themselves. In this case, the false rejection and acceptance rates $\alpha\,(\Theta_{post})$ and $\beta\,(\Theta_{post})$ must be understood as the performance of the system *with ideal thresholds.* Though this procedure does not lead to a fair measure of the system performance, it can be interesting, for diagnostic evaluation, to compare $\alpha\,(\Theta_{prio})$ and $\beta\,(\Theta_{prio})$ with $\alpha\,(\Theta_{post})$ and $\beta\,(\Theta_{post})$.

12.4.3.8 System operating characteristic

Whichever operating constraint is chosen to tune the thresholds is only one of the infinite number of possible trade-off, and it is generally not possible to predict, from the false rejection and false acceptance rates obtained for a particular functioning point, what would be the error rates for an other functioning point. In order to be able to estimate the performance of the system under any conditions, its behaviour has to be modeled so that its performance can be characterised independently from any threshold settings.

Speaker-independent threshold

In the case of a speaker-independent threshold, the false rejection and the false acceptance rates can be written as functions of a single parameter θ, namely $\alpha = \alpha\,(\theta)$ and $\beta = \beta\,(\theta)$. Then, a more compact way of summarising the system's behaviour consists in expressing directly β as a function of α (or the opposite), i.e :

$$\beta = f\,(\alpha)$$

Using a terminology derived from Communication Theory, function f is sometimes called the *Receiver Operating Characteristic* and the corresponding curve $\mathcal{F}$ the *ROC curve.* Generally, function f is monotonically decreasing, and satisfies the limit conditions $f\,(0) = 1$ and $f\,(1) = 0$. Figure **??** depicts a typical ROC curve.

The point-by-boint knowledge of function f provides a threshold-independent description of all possible functioning conditions of the system. In particular :

- If a false rejection rate α_0 is specified, the corresponding false acceptance rate is obtained as $\beta_0 = f\,(\alpha_0)$. Graphically, this corresponds to the intersection of the ROC curve $\mathcal{F}$ with the vertical straight line of equation $\alpha = \alpha_0$.

- If a false acceptance rate β_0 is specified, the corresponding false rejection rate is obtained as $\alpha_0 = f^{-1}\,(\beta_0)$. Graphically, this corresponds to the intersection of the ROC curve $\mathcal{F}$ with the horizontal straight line of equation $\beta = \beta_0$.

- If the expected benefit is to be maximized, the derivation of equation (12.43) shows that :

$$\frac{\partial \mathcal{B}}{\partial \theta} = 0 \quad \Longrightarrow \quad \frac{\partial \beta}{\partial \alpha} = -\frac{p}{1-p}\,\frac{B_X + C_X}{B_Y + C_Y} = \lambda \qquad (12.45)$$

  Graphically, the corresponding functioning point is obtained by sliding, from the origin, a straight line with slope λ, until it becomes tangent to the ROC curve. The point of contact then indicates the corresponding α and β.

- To obtain the equal error rate ϵ, the equation $\epsilon = f\,(\epsilon)$ has to be solved. This functioning point corresponds to the intersection of the ROC curve with the straight line of equation $\beta = \alpha$.

In practice, there are several ROC curves, depending on what type of false rejection and acceptance scores are used :

a *gender-balanced ROC* : $\bar{\beta}_{MF} = f(\bar{\alpha}_{MF})$ (or $\bar{\beta}'_{MF} = f(\bar{\alpha}_{MF})$ if impostors are unknown),

an *average ROC* $\bar{\beta} = f(\bar{\alpha})$ (or $\bar{\beta}' = f(\bar{\alpha})$ if impostors are unknown),

a *test set ROC* $\beta = f(\alpha)$.

Any other combination would not make much sense. However, keeping exhaustively a whole ROC curve lacks conciseness, and it is classically felt desirable to condense the system performance in a single figure. Traditionally, the EER is chosen for this purpose, In this case, there is a distinct equal error rate for each ROC curve, which can be denoted respectively $\bar{\varepsilon}_{MF}$, $\bar{\varepsilon}$ and ε.

Speaker-dependent thresholds

In the case of speaker-dependent thresholds, the false rejection and the false acceptance rates for each speaker X_i depend on a different parameter θ_i. Therefore, each speaker has his own ROC curve :

$$\beta_i = f_i(\alpha_i)$$

In this case, there is no trivial way for deriving an "average" ROC curve that would represent the general behaviour of the system. Current practice consists in characterising each individual ROC curve by its equal error rate ξ_i, and in summarising the performance of the system by the *average equal error rate* $\bar{\xi}$ computed as :

$$\bar{\xi} = \frac{1}{m} \sum_{i=1}^{m} \xi_i \qquad (12.46)$$

Note here that a *gender-balanced equal error rate* $\bar{\xi}_{MF}$ can be defined as :

$$\bar{\xi}_{MF} = \frac{1}{2}\left(\bar{\xi}_M + \bar{\xi}_F\right) = \frac{1}{2}\left(\frac{1}{m_M}\sum_{i \in M}^{m} \xi_i + \frac{1}{m_F}\sum_{i \in F}^{m} \xi_i\right) \qquad (12.47)$$

and a *test set equal error rate* as :

$$\xi = \sum_{i=1}^{m} p_i\, \xi_i \qquad (12.48)$$

Though we used the same terminology for denominating equal error rates with speaker-dependent and speaker-independent thresholds, it must be stressed that the scores are not comparable. Therefore, it should always be specified in which framework they are computed.

12.4.3.9 System characteristic modeling

Equal error rates can be interpreted as a very local property of the ROC curve. In fact, as the ROC curve has usually its concavity turned in the direction of the axis $\alpha = \beta$, the EER gives an idea of how close the ROC curve is to the axes. However, this is a very incomplete picture of the general system performance level, as it is virtually impossible to predict the performance of the system under a significantly different operating condition.

Recent work by Oglesby [Oglesby 95] has addressed the question of how to *encapsulate* the entire system characteristic into a single number. Oglesby's suggestions, which we will develop now, consist in finding a simple 1-parameter model which describes as accurately as possible the ROC curve over most of it definition domain. If the approximation is good enough, reasonable error rate estimates for any functioning point can be derived. As in the last section, we will first discuss the case of a system with a speaker-independent threshold, and then extend the approach to speaker-dependent thresholds.

For modeling the relation between α and β, the simplest approach is to assume a linear operating characteristic, i.e a relation between α and β of the kind :

$$\alpha + \beta \approx 2\hat{\epsilon}_{lin} \quad \Longleftrightarrow \quad \beta \approx \hat{f}\,(\alpha) = 2\hat{\epsilon}_{lin} - \alpha \tag{12.49}$$

where $\hat{\epsilon}_{lin}$ is a constant which can be understood as the *linear-model EER*[51]. However, typical ROC curves do not have at all a linear shape, and this model is too poor to be effective on a large domain.

A second possibility is to assume that the ROC curve has the approximate shape of the positive branch of a hyperbola, which supposes the relation :

$$\alpha \,.\, \beta \approx \hat{\epsilon}^2_{hyp} \quad \Longleftrightarrow \quad \beta \approx \hat{f}\,(\alpha) = \frac{\hat{\epsilon}^2_{hyp}}{\alpha} \tag{12.50}$$

Here $\hat{\epsilon}_{hyp}$ is another constant which can be interpreted as the *hyperbolic-model EER.* The hyperbolic model is equivalent to a linear model in the log-error domain. It usually fits much better the ROC curve[52]. However, it has the drawback of not fulfilling the limit conditions, as $\hat{f}\,(0) = +\infty$ and $\hat{f}\,(1) \neq 0$.

A third possibility, proposed by Oglesby, is to use the following model :

$$\alpha^{-\frac{1}{\log_2 \hat{\epsilon}_{ogl}}} + \beta^{-\frac{1}{\log_2 \hat{\epsilon}_{ogl}}} \approx 1 \quad \Longleftrightarrow \quad \beta \approx \hat{f}\,(\alpha) = \left(1 - \alpha^{-\frac{1}{\log_2 \hat{\epsilon}_{ogl}}}\right)^{-\log_2 \hat{\epsilon}_{ogl}} \tag{12.51}$$

where $\hat{\epsilon}_{ogl}$ will be refered to as *Oglesby's model EER.* Oglesby reports a good fit of the model with experimental data, and underlines the fact that $\hat{f}\,(0) = 1$ and $\hat{f}\,(1) = 0$.

The parametric approach is certainly a very relevant way to give a broader system characterisation. Nevertheless, several issues remain questionable.

First, it is clear that none of the models proposed above account for a possible skewness of the ROC curve. As Oglesby notes it, to address skewed characteristics would require introducing an additional variable, which would give rise to a second non-intuitive figure.

A second question is what criterion $\mathcal{C}\,(\hat{f})$ should be minimised to fit the model curve $\hat{\mathcal{F}}$ to the true ROC curve $\mathcal{F}$. If we denote as $[\alpha_1,\, \alpha_2]$ the *optimisation domain* on which the best fit is to be found, the most natural criterion would be to minimise the mean square error between $\hat{\mathcal{F}}$ and $\mathcal{F}$ over the interval $[\alpha_1,\, \alpha_2]$. However, an absolute error difference does not have the same meaning when α changes order of magnitude, and an alternative could be to minimise the mean square error between the curves in a log-log representation.

A third and most crucial question is how should the unavoidable deviations between the model and the actual ROC curve be quantified and reported.

Here is a possible answer to these questions. Though the approach that we are going to present has not been extensively tested so far, we believe that it is worth exploring it in the near future, as it may prove useful to summarise concisely the performance of a speaker verification system, in a relatively meaningful and exploitable manner.

The solution proposed starts by fixing an *accuracy* ω % for the ROC curve modeling, say for instance 5 %. Then, if we define :

[51] This approach is to be linked to a common practice which consists in summarising the performance of a speaker verification system as the arithmetic mean of the false rejection rate and the false acceptance rate obtained for a particular threshold.

[52] Hence, the geometric mean is usually a better way to average a false rejection rate and a false acceptance rate for a given threshold.

$$\phi^+_\omega(\alpha) = (1+\omega)\, f(\alpha) \qquad \phi^-_\omega(\alpha) = (1-\omega)\, f(\alpha)$$
$$\psi^+_\omega(\alpha) = f\,(\frac{\alpha}{1+\omega}) \qquad \psi^-_\omega(\alpha) = f\,(\frac{\alpha}{1-\omega})$$

the following properties are obvious :

$$\text{if} \quad \phi^-_\omega(\alpha) \le \hat{f}(\alpha) \le \phi^+_\omega(\alpha) \quad \text{then} \quad \frac{\left|\hat{f}(\alpha) - f(\alpha)\right|}{f(\alpha)} = \frac{\left|\hat{f}(\alpha)) - \beta\right|}{\beta} \le \omega \tag{12.52}$$

$$\text{if} \quad \psi^-_\omega(\alpha) \le \hat{f}(\alpha) \le \psi^+_\omega(\alpha) \quad \text{then} \quad \frac{\left|\hat{f}^{-1}(\beta) - f^{-1}(\beta)\right|}{f^{-1}(\beta)} = \frac{\left|\hat{f}^{-1}(\beta) - \alpha\right|}{\alpha} \le \omega \tag{12.53}$$

Hence, when both constraints are satisfied, both relative differences between the modeled and exact false rejection and acceptance rates are below $\omega\,\%$.

Then, a model of the ROC curve must be chosen, for instance Oglesby's model. However, if another model fits better the curve, it can alternatively be used, but it preferably should depend on a single parameter, and the link between the value of this parameter and the model equal error rate should be specified.

For a given parameter $\hat{\epsilon}$, the lower and upper bound of the *ω%-accuracy false rejection rate validity domain*, $\alpha^{inf}_\omega(\hat{\epsilon})$ and $\alpha^{sup}_\omega(\hat{\epsilon})$ are obtained by decreasing (respectively increasing) α, starting from the initial value $\alpha_{init} = \epsilon$, until one of the two constraints of equations (12.52) and (12.53) are no more satisfied. This process can be repeated for several values of $\hat{\epsilon}$ vaying for instance in small steps within the interval $[\epsilon\,(1-\omega),\, \epsilon\,(1+\omega)]$. Finally, the value of $\hat{\epsilon}$ corresponding to the wider validity domain can be chosen as the system performance measure, in the validity domain of the approximation. Note that $\hat{\epsilon}$ does *not* need to be inside the validity domain, for its value to be meaningful.

If the validity domain turns out to be too small, then the process could be repeated after having set the accuracy ω to a higher value. Another possibility could be to give several model equal error rates, corresponding to several adjacent validity domains (with a same accuracy ω), i.e a piecewise representation of the ROC curve.

A first advantage of the parametric description is that it allows to predict the behaviour of a speaker verification system for a more or less extended set of operating conditions. It could then be possible to give clear answers to a potential client of the technology, if this client is able to specify his constraints. The second advantage is that the model EER is a number which relates well to the conventional EER. Therefore, the new description would not require that the scientific community changes totally its point of view in apprehending the performance of a speaker verification system[53]. The main drawback of the proposed approach is that it lacks experimental validation for the time being. Therefore, we suggest to adopt it as an experimental evaluation methodology, until it will have proven efficient.

If dealing with a system using speaker-dependent thresholds, we are brought back to the difficulty of averaging ROC curve models across speakers. The ROC curve for each speaker X_i can be summarized by a model equal error rate $\hat{\xi}_i$ and a ω%-accuracy false acceptance rate validity domain $[\alpha^{inf}_{i\,\omega},\, \alpha^{sup}_{i\,\omega}]$. In the lack of a more relevant solution, we suggest to characterise the average system performance by averaging across speakers the model EER, and the bounds of the validity interval. Thus the global system performance could be given as an *average model EER* :

$$\bar{\hat{\xi}} = \frac{1}{m} \sum_{i=1}^{m} \hat{\xi}_i \tag{12.54}$$

and an *average ω%-accuracy false acceptance rate validity domain* :

$$[\bar{\alpha}^{inf}_\omega,\, \bar{\alpha}^{sup}_\omega]$$

[53] In fact, a conventional EER ε is the model EER $\hat{\varepsilon}$ for any model, but with a 0 %-validity domain of $[\varepsilon,\, \varepsilon]$!

where :

$$\bar{\alpha}_{\omega}^{inf} = \frac{1}{m} \sum_{i=1}^{m} \alpha_{i\,\omega}^{inf} \quad \text{and} \quad \bar{\alpha}_{\omega}^{sup} = \frac{1}{m} \sum_{i=1}^{m} \alpha_{i\,\omega}^{sup} \tag{12.55}$$

The same approach can be implemented, with different weights, to compute a *gender-balanced model EER* $\hat{\bar{\xi}}_{MF}$ and a *test set model EER* $\hat{\xi}$, and the corresponding validity domains.

Another possibility would be to fix a speaker-independent validity domain $[\alpha^{inf}, \alpha^{sup}]$ for each ROC curve, and then compute the individual accuracy ω_i. Then, to obtain a global[54] score, all ω_i could be averaged (using weights depending on the type of estimate), and the performance would be a global model equal error rate together with a false acceptance rate domain $[\alpha^{inf}, \alpha^{sup}]$ common to all speakers, but at an average accuracy.

12.4.3.10 Example

For example, consider a verification system *with a speaker-independent threshold* that has a *gender-balanced Oglesby's equal error rate of* 4.7 % *with a* 5%*-accuracy false rejection rate validity domain of* [0.5 %, 25.0 %]. Here, the ROC curve under consideration is $\bar{\beta}_{MN} = f(\bar{\alpha}_{MN})$. We will denote now $\bar{\alpha}_{MN} = a$ and $\bar{\beta}_{MN} = b$, for simplicity reasons.

For any false rejection rate a satisfying $0.005 \leq a \leq 0.250$, the difference between the actual false acceptance rate b and the estimated false acceptance rate $\hat{f}(a)$ predicted by Oglesby's model with parameter $\hat{\bar{\varepsilon}}_{MF} = 0.047$ satisfies $|\, b - \hat{f}(a)\,| \leq 0.05\, b$. It can then be computed (using equation (12.51)) that the 5%-accuracy false *acceptance* rate validity domain is [0.3 %, 20.6 %], and it is guaranteed that, for any value of b in this interval, the difference between the actual false rejection rate a and the estimated false rejection rate $\hat{f}^{-1}(b)$ (predicted by Oglesby's model with EER 0.047) satisfies $|\, a - \hat{f}^{-1}(b)\,| \leq 0.05\, a$. In particular, the exact (gender-balanced) EER of the system, $\bar{\varepsilon}_{MF}$, is equal to 0.047, at a 5 % relative accuracy.

12.4.4 Open-set identification

An open-set identification system can be viewed as a function which assigns, to any test utterance z, an estimated speaker index $\hat{w}(z) \in \{1...m\}$, corresponding to the identified speaker $X_{\hat{w}(z)}$ in the set of registered speakers, or outputs 0 if the applicant speaker is considered as an impostor.

In open-set identification, three types of error can be distinguished :

- a *misclassification error* occurs for a genuine test utterance x_i^k when :

$$\hat{w}(x_i^k) \neq 0 \quad \text{and} \quad \hat{w}(x_i^k) \neq i$$

- a *false rejection* occurs for a genuine test utterance x_i^k when :

$$\hat{w}(x_i^k) = 0$$

- and a *false acceptance* occurs if, for an impostor test utterance y_j^k :

$$\hat{w}(y_j^k) \neq 0$$

Here, two points of view can be adopted.

Either a misclassification error is considered as a false acceptance (while a correct identification is treated as a true acceptance). In this case, open-set identification can be scored in the same way as verification, namely by evaluating a false rejection rate α and a false acceptance rate β. The

[54] That is gender-balanced, average or test set.

concept of ROC curve can be extended to this family of systems, and in particular, an equal error rate ϵ can be computed. However, the false acceptance rate β is now bounded by a value $\gamma^c \leq 1$ when the threshold θ tends to 0, γ^c being the closed set misclassification rate of the system, i.e the performance that the open-set identification system would provide if it was functioning in a closed-set mode. Therefore, a parametric approach for dynamic evaluation would require a specific class of ROC curve models (at least with 2 parameters). Moreover, merging classification errors with false acceptances may not be appropriate if the two types of errors are not equally harmful.

An alternative solution is to keep distinct the three types of errors, and measure them by three rates α, β and γ. The ROC curve is now a curve in a 3-dimensional space, with equation $\varphi(\alpha, \beta, \gamma) = 0$. The two extremities of this curve are the points with coordinates $(\alpha = 1, \beta = 0, \gamma = 0)$ and $(\alpha = 0, \beta = 1, \gamma = \gamma^c)$. The ROC curve can be projected as $\beta = f(\alpha)$ and $\gamma = g(\beta)$. The first projection is a monotonically decreasing curve such as $f(0) = 1$ and $f(1) = 0$, whereas the second projection is also monotonically decreasing, and satisfies $g(0) = 0$ and $g(1) = \gamma^c$. A minimal description of the curve of φ could then be the equal error rate ϵ of function f and the closed-set identification score γ^c from function g. Parametric models of φ with 2 degrees of freedom could be thought of, but to our knowledge, this remains an unexplored research topic.

Among both possibilities, we believe that the second one is to be prefered, though it is slightly more complex.

12.4.5 Recommendations

These recommendations indicate how the performance of a speaker recognition system should be scored.

- For closed-set identification
  - Beside the test-set misclassification rate (γ), report on average misclassification and mistrust rates ($\bar{\gamma}$ and $\bar{\bar{\gamma}}$), and provide also gender-balanced rates ($\bar{\gamma}_{MF}$ and $\bar{\bar{\gamma}}_{MF}$) if the test population is composed of male and female speakers.
  - As the number of registered speakers is a crucial factor of performance, it is essential to indicate the number of speakers in the registered speaker population. Mention also the proportion of male and female speakers for information.
  - For statistical validity information, indicate the number and male/female proportion of speakers in the test population and the average number of test utterances per test speaker.
- For verification
  - For static evaluation, beside the test set false rejection rate (α) and the test set false acceptance rate (β), provide the average false rejection rate $(\bar{\alpha})$ and the average false acceptance rate ($\bar{\beta}$ or $\bar{\beta}'$ depending on whether the impostors' identities are known or not). Gender-balanced rates ($\bar{\alpha}_{MF}$ and $\bar{\beta}_{MF}$ or $\bar{\beta}'_{MF}$) should also be reported.
  - For dynamic evaluation and a speaker-independent threshold, the system ROC curve should be obtained as :

    either $\bar{\beta}_{MF} = f(\bar{\alpha}_{MF})$ (or $\bar{\beta}'_{MF} = f(\bar{\alpha}_{MF})$ if impostors are unknown),
    or $\bar{\beta} = f(\bar{\alpha})$ (or $\bar{\beta}' = f(\bar{\alpha})$ if impostors are unknown),
    or $\beta = f(\alpha)$

    Summarise a ROC curve by its traditional equal error rate (respectively $\bar{\varepsilon}_{MF}$, $\bar{\varepsilon}$ and ε). Investigate on the possibility of finding a ROC curve model, and report the model equal error rate ($\hat{\bar{\varepsilon}}_{MF}$, $\hat{\bar{\varepsilon}}$ and $\hat{\varepsilon}$) and the ω%-accuracy false rejection rate validity domain $[\alpha^{inf}, \alpha^{sup}]$. Find a reasonable compromise between ω and $\alpha^{sup} - \alpha^{inf}$.

- For dynamic evaluation with speaker-dependent thresholds, compute the individual equal error rate (ξ_i) of each ROC curve $\beta_i = f(\alpha_i)$ and give the gender-balanced equal error rate ($\bar{\xi}_{MF}$), the average equal error rate ($\bar{\xi}$) and the test set equal error rate (ξ). Investigate the possibility of fitting a common ROC curve model by adjusting individually a model equal error rate ($\hat{\xi}_i$) for each curve. Here, either an accuracy ω is fixed and speaker-dependent validity domains are computed, or the validity domain is fixed in a speaker-independent manner and the individual accuracy ω_i is computed. Compute anyway the global model equal error rates ($\bar{\hat{\xi}}_{MF}$, $\bar{\hat{\xi}}$ and $\hat{\xi}$). Then give accordingly either the average validity domain for a speaker-independent accuracy or the average accuracy for a speaker-independent validity domain.
- For statistical validity information, indicate the number of registered speakers, the proportion of male and female registered speakers, the number of genuine test speakers, the proportion of male and female genuine test speakers and the average number of genuine test utterance per genuine test speaker. Give also a relevant description of the test impostor configuration and population.

- For open-set identification
  - For static evaluation, score separately the false rejections (α, $\bar{\alpha}$ and $\bar{\alpha}_{MF}$) , the false acceptances (β, $\bar{\beta}$ or $\bar{\beta}'$ and $\bar{\beta}_{MF}$ or $\bar{\beta}'_{MF}$) and the misclassifications (γ, $\bar{\gamma}$ and $\bar{\gamma}_{MF}$).
  - For dynamic evaluation and a speaker-independent threshold, project the 3-dimensional ROC curve $\varphi(\alpha, \beta, \gamma)$ into 2 curves $\beta = f(\alpha)$ and $\gamma = g(\beta)$. Summarise the first one as its equal error rate ε and the second one as its extremity γ^c. Investigate the possibility of using a parametric approach.
  - For dynamic evaluation with speaker-dependent thresholds, average individual ξ_i and individual γ_i^c. Investigate the possibility of using a parametric approach.
  - As for closed-set identification and for verification, give all relevant information concerning the registered population, the genuine test population and the impostor population and test configuration.

In practice, gender-balanced, average and test set scores are obtained very easily as various linear combinations of individual speaker scores.

12.5 Complementary assessment tools

In the previous paragraphs, the focus was centered on performance evaluation of the algorithms with emphasis on scoring procedures. In many cases the scores are meaningless if they are not compared with the scores of other systems on the same database. A reference is necessary to give its full value to a measure. This is particularly true when the measure is based on training and test data which may not be publicly available. In such a case, it is impossible to reproduce the test independently. Databases could be designed to be representative of a particular application or could be more generic. The sampling of speakers may be biased toward similar voices, similar accents, ... The similarities in the voice characteristics of speakers should be quantified anyway. A reference system is a tool to evaluate the complexity of a database.

This section proposes techniques to evaluate databases, provide reference measures for comparison and tools to test the limits of speaker verification systems.

12.5.1 Standard reference systems

A complementary approach to system evaluation consists in defining a verification system, to which other systems are to be compared. For any new system embedded in a given application, the reference voice recognition system could be substituted to the new one, and the differences in the overall performance of the application would give an indirect figure of merit of the new system compared to the reference one, for this given application.

(draw a figure)

Desirable properties of such reference systems are their relative efficiency and robustness and their easy implementation and reproductibility (from the algorithmical point of view). They should not require sophisticated training procedures. In particular, they should be able to operate with very limited training data (like one utterance) per speaker. Such a system allows to measure indirectly the complexity of a task, by giving a reference figure on this task.

Among the features of a reference system, the decision strategy has a considerable impact on the results of the system. Besides the raw performance, the way doubtful decisions are handled is likely to improve considerably the efficiency in an operating mode. This fact should be taken into account when defining a standard system.

Let us examine a few possible candidates.

For text-dependent and text-prompted applications, a baseline DTW system offers a number of advantages:

- training could be done with one utterance per speaker
- multi-references can be used if more training data are offered
- a distance measure between reference and test is returned
- the ROC curve can be obtained by varying a threshold

but DTW is very sensitive to end-point detection. Such a reference DTW system was proposed in the context of the SAM-A project [Homayoun 93].

For text-independent applications, a second order statistical (sphericity) measure was proposed in SAM-A [Bimbot 93]. It captures the correlations of a time-frequency representation of the signal. The correlation matrix of a speaker is better estimated on a long period of speech. Nevertheless some encouraging results have been obtained on segments as short as 3 sec.

Both systems were investigated because they are based on totally reproductible algorithmic methods; in particular, they do not require any iterative learning of models, while as Vector Quantization, Hidden Markov Models or Neural Networks performances can be affected by learning strategies. The issue of defining one or two reference systems is certainly crucial for evaluation methodology.

Commercial products could also be evaluated in parallel on the same database. This is not an easy task. On the shelf systems are usually targeted for a specific application. The type of input they expect could be restricted. They usually have specific training protocols. They may not provide a digital speech input. A specific 'harness' must be developed to input samples from the database.

12.5.2 Reference human test

In speech assessment, the human performance in achieving a well defined task is one of the most popular methods to evaluate the limits of automatic methods, to investigate the human approach to the solution of the problem (so that the automatic system may take advantage of this knowledge) and to score the human performance in comparison to "system" performance. For instance the classical HENR method to assess speech recognisers, and the well defined and established set of listening tests (DRT, MOS, etc.) used in assessment of speech output systems are well known techniques in speech assessment.

There is a wide literature about speaker recognition by human listeners, and only a small part of these experiments report on specific data dealing with speaker verification. Unfortunately for speaker recognition, no formalism has been established and the experiments reported in the literature usually do not have the same experimental conditions, so that results can not be easily compared.

In the field of speaker identification, the large amount of possible distinctive factors that should be managed in a listening session, such as number of speakers, duration of voice material per talker, voice familiarity, phonetic content of the speech material, delay between sessions, etc..., make the definition of a "standard" listening test a very difficult goal. Nevertheless, for speaker verification the problem becomes simpler, and reference may be found in [Rosenberg 73], or in [Federico 89].

12.5.3 Automatic / human tests

It is a fact that, in the past, effort in this field have been spent to answer the question of how accurate are the automatic methods for talker recognition compared to human listeners, i.e. to realise if the performance of an automatic system are acceptable, supposing that when system performance are similar to listener performance, they are acceptable. Further it has been set that automatic speaker identification is one area of speech processing where machines can exceed human performance. This is true when short utterances are used, or a great number of speakers, or when the voices are "unfamiliar". On the contrary it has been reported that humans appear to handle mimics better than automatic speaker verification systems. In fact [Reich 81], listeners seem able to associate the mimic voice with the intended one, but do not confuse them.

The comparison between automatic methods and human listeners usually ends up in the selection of a task reasonable for the automatic speaker identification and unfair to listener capability. For these reasons, we believe that results reported in the literature should be interpreted with considerable caution. In the meantime, more knowledge and experience about relationships between listener performance and automatic system performance would surely be very helpful both in the design and development of speaker recognition and in the assessment of these systems : a compelling step is now to define and test procedures for listening test in automatic speaker recognition, so that effort in this field will not vanish owing to a lack of reproductibility, or to multiple test conditions.

Given that listening tests are very time consuming and cost expensive research activities, it is not realistic to envisage such a procedure on every existing database. A good compromise would be to dedicate some effort to human calibration of standard databases. Nevertheless further research and experiments in the human test field is necessary, in order to fix possible standards and recommendations supported by both theoretical models and experimental results, as the listening methods would surely be helpful both in development and assessment of speaker recognition systems.

Preliminary experiments [Homayounpour 93] have been conducted in the framework of the SAM-A project, on the comparative evaluation of speaker verification by human listeners and by an automatic system (based on DTW). Tests were conducted on emotional and imitated speech. Both databases present difficulties to human and to machine. On these tasks, under similar conditions, human listeners still perform better than our reference system, but differences are not large.

12.5.4 Transformation of speech databases

Speaker recognition systems are sensitive to intra-speaker variability and to environmental factors, such as noise, transducer and channel characteristics, etc... In order to quantify this sensitivity, these factors should be modeled and varied independently. For each dimension, the limits of acceptable variation can be measured, providing a sensitivity profile for the system under test.

Goldman et al. [Goldman 93] have investigated the feasibility of using speech transformation techniques to find the limits of speaker verification systems. Voluntary modifications directed to mimic several emotional states or to mimic an other speaker are first tested, then artificial transformations are investigated. These preliminary tests are limited to the modification of some prosodic factors (namely duration and fundamental frequency), and to simulated vocal tract length modifications. Two prosodic features (duration and fundamental frequency) and an articulatory feature (simulated vocal tract length) are varied independently, and the consequences of such artificial modifications are evaluated in the framework of text-dependent speaker verification. Techniques to simulate Lombard effect and environmental factors were also investigated in the context of ASR assessment.

Accurate models of speech variability need to be investigated further. It must be based on data from large databases specifically designed to capture intra- and inter-speaker variability.

12.6 Typology and assessment of applications

The speech signal conveys a message (the linguistic information) originating from a source (a human speaker or a synthesizer) and transmitted through a channel to be understood by the receiver. These four aspects should be kept in mind during the evaluation of a speech application. The paralinguistic information conveyed by the speech signal is and will be exploited in many practical applications. In some cases, it is used to improve an other aspect of speech communication (as in speech recognition and speech synthesis). In other cases (speaker verification/identification, speaker change detection, pathology detection, age/sex/language identification, ...), it is the essence of the application.

Let us list a number of application areas:

- Speech recognition (accent (regional, foreign), speaking habit, speaker classification, adaptation, intent of the speaker, speaking style, change in speaking rate
- Speech synthesis: increase naturalness, offer the possibility of changing the speaker characteristics, convey emotion
- identification/verification of identity
- change of speakers
- speech pathology detection and evaluation, physical and psychological state
- pronounciation training and reeducation
- ...

It is unrealistic to review all possible applications, many of which being unknown at this time. What is attempted here is a classification of potential applications with a main focuss on speaker verification.

The scientific interest here is to assess how reliable the paralinguistic information is for the intended application.

Let us illustrate our argument with the simple example of the 'lie detector', a system which could be capable of detecting when, on purpose, a speaker is not saying the truth. Such systems have been implemented and sold. It is a commercial application. It is doubtful that any objective evaluation of such a system has been conducted. It could even be argued that such a system should obtain approval from an independent (standardisation) institute prior to commercial exploitation. Such administrative and social issues can only be mentioned in this handbook. The responsibility of the scientific community should be limited to providing the tools and methods to assess the performance of systems. But, guidelines and recommandations should also be made about the feasibility of applications given our current knowledge.

These issues will be revisited in section XXX dedicated to forensic applications (legal and police investigations)

We are convinced that several approaches to the problem of speaker recognition are now efficient enough to bring advantages to some real-world applications. It is a matter of putting together a reasonably efficient algorithm, a robust decision strategy and a carefully designed ergonomy, into a well-engineered system. As an additional layer of security, voice verification can be effective in several applications, even if not 100 % reliable (personal identification numbers are not 100 % sure either). Beside its dissuasive effect, it is clear that it can reduce some significant part of unauthorized access or transactions, without being perceived as an hindrance by the conventional customer. Multi-modality will extend progressively its action field in the everyday world, and voice recognition will undoubtedly have a role to play in this development.

It is often stressed that the advantages of text-independence come from 2 sources: for machine-driven text-independent systems, if the speaker can not predict what he has to say, an impostor can not use any pre-recording of the true speaker's voice to cheat the system ; for user-driven text-independence, speaker recognition remains possible even if only recordings of the speaker are avalaible, or in case the speaker is non-cooperative. Note incidentally that, in case of security applications, the first argument is only valid if some kind of speech verification is coupled with the speaker recognition technique, to check that the speaker is indeed saying what he is supposed to say.

Recommendations

12.6.1 Applications and products in speaker recognition

In this chapter, we list several potential or existing applications of speaker identification and verification technology, and we give an overview of the existing products that we are aware of.

We distinguish remote and local applications. Remote applications are typically performed over the telephone. Telephone sets and channel paths present quite diverse characteristics. The microphone and the environment is much better controlled with local applications.

needs of the user, alternatives

advantages to the user/to the service provider

cost

the length of the speech material necessary for training and testing phases, the computation complexity, the number of speakers used for evaluation, etc... However, these factors are not structural characteristics of the system.

Telecommunication (remote) applications and products

One of the most obvious use of speaker recognition techniques is caller authentication over the telephone network. In the framework of such applications, the main task is therefore speaker verification. The user claims his identity, most of the time by dialing (or saying) a personal code number. Then, either a code word is required to authentify the speaker, or his utterance of the code number is used for verification purposes.

Typical application areas are:

- banking (checking balance, transfer of funds),
- credit cards. Remote payment by credit card is usually accepted by providing the name of the owner, the card number and the expiration date (all this information is printed on the card)
- teleshoping (grant autorisation to tranfer money)

- stock exchange (purchasing, selling stocks)
- home incarceration
- military applications
- calling cards (SPRINT), billing your own account
- access to/modification of information on remote servers (restrict access to authorized users).
- access to computing facilities from a remote terminal (coupled with encryption). Passwords are currently used. Alternatives necessitate non-standard equipment (card reader, scanner, electronic pen, encrypted key, ...)
- ...

The main applications for speaker verification over the telephone network are of two kinds : the first kind is banking and remote transactions, the second kind is access to licensed databases. It is obvious that both fields do not require the same level of security : it is usually less costly to let someone unauthorised have access to a database, than to allow somebody to operate some transaction that can involve large amounts. However, in many countries, it is possible to pay by credit card over the phone, without any other verification of the customer than the consistence between the customer's name, his credit card number and its expiry date (all of them being on the credit card itself !). Some kind of basic speaker verification (for instance user-specific text-dependent verification), even with tolerant thresholds, could certainly bring more security to this type of transactions. Note that, in this case, the speaker characteristics should be centralised in a place that delivers the transaction authorisations, and would require a special equipment for each supplier, but the verification could take place off-line because the confirmation of the transaction does not usually need to be immediate.

Some telephone operators provide a toll free service allowing customers to place a phone call from any terminal. The charge for the call will appear on the phone bill of the user. For such an application, the user must be identified. A personal identification number (PIN) could be requested. This information could be entered via the dial pad. Such a protocol lacks security as some people are chasing valid PINs to use or sell them. The SPRINT operator in the USA proposes a successful alternative. They use speaker verification. Upon a call to the service, the user is requested to say the sequence of digits of his account. Digit recognition and speaker verification is performed on this sequence. In case of doubt, he is then prompted a random sequence of digits for further validation (getting some confidence he is not playing back some prerecorded speech). He is then allowed to dial the phone number of the person he wants to reach.

¿¿¿¿¿(je n'ai pas souvenir des details precis de ce protocol! Could you help?)¡¡¡¡¡¡¡¡¡¡

The SPRINT operator has registered 1.5 milion customers for this service. In such an application, a rather high false acceptance rate is tolerable. Impostors are aware that their voice is being recorded. The introduction of speaker verification proved to be dissuasive against fraud. With a rather small increase in the complexity of the access protocol, a satisfactory level of security was achieved. Of course, a higher degree of security could be reached by letting the user dial by voice. The risk of recognition error (although rather small as recognition could be performed in a speaker dependent mode) would necessitate confirmation and a longer dialogue (paid by the service provider!).

Speaker verification over the telephone line in Germany is mainly done at SIEMENS Munich. At the end of 1990, the necessary database has been evaluated. The speech was collected from calls over analog and ISDN lines [Joachim90].

In the UK, the BT home banking trial with the Royal Bank of Scotland used speaker verification technology to provide an additional layer of security to PINs (Personal Identification Number). The concern of offending as few genuine customers as possible was very relevant to the bank.

VOCALIS Ltd. (at that time part of LOGICA) developed a real-time telephone-based text-dependent system in 1988. The system's decision making logic has 4 layers of authentication prior to making the final decision on the claimed identity of the speaker. Individual password and PIN number are used as the first 2 layers of authentication. A number of words from pre-defined vocabularies are also chosen randomly at the later stages of verification. The enrolment process follows a robust strategy, in order to guarantee that the reference template is relevant. Unlike many similar products, VOCALIS' speaker verification system is developped for UNIX platform and makes use of embedded hardware. The application is fully integrated into the company's communications infrastructure. An extensive field trial of the system was organised in conjunction with a telephone service company in the US, where over 400 people from 4 different geographical groups registered with the system. The system allowed users to access US Government databases out of office hours. All registered participants were encouraged to make at least 10 calls over a 3 month period, using as many handsets as possible. The registered speakers were also asked to make 2 impostor calls, and were purposedly provided with information to break the first 2 layers of authentication. At the end of the trial, more than 4000 calls were logged. A detailed analysis of the results showed a 5 % equal error rate purely for the verification algorithm. None of the participants managed to break into the system as impostors when all layers of authentication were active. A similar speaker verification system is available also on the Callserver platform still from VOCALIS.

In the UK also, the Barclays Bank uses a speaker verification system.

Voice Control System have put speaker verification in their speech recognition system over the telephone line. It works on several platforms including PC and VME. By August 1990 [Hunt91] the performance was of 1% of false acceptance, and 2% of false rejection.

A speaker verification facility is also part of the AT&T HuMaNet teleconferencing system [Berkley90] that works via ISDN and uses an AT&T digital signal processor DSP32C board on a Personal Computer.

The CNAPS of Adaptive Solution Inc. works over the telephone line and gives both speaker verification and speaker identification capabilities [Skinner92] ; averaging the true acceptance and true rejection rates the system reaches 95

ENSIGMA Ltd. produces a speaker verification system named VERIFIER [Moody 91], that runs on the Loughborough AT&T DSP32C telephone board. It works both over the telephone line and audio line. Ensigma claims full portability of the speaker verification software on other boards. The system uses a new approach to the speaker verification problem. In fact, instead of simply matching a speaker's voice against the user's template, ENSIGMA's system compares the spoken word also against a "general world-model". The decision is made taking into account the results of these two comparisons. This method has been evaluated over the telephone line and gives a 1% of false rejection and 1% of false acceptance. The method seems also quite robust to background noise. The enrolment consists of the new user repeating a series of digits ; the duration of this phase is about one minute. The access consists of the pronunciation by the user of a list of (random) digits asked by the system. System response takes about 5 seconds. The VERIFIER has many properties that make this system a good candidate as a possible reference-hardware system for (text-dependent) speaker verification. The multi-lingual aspect, or generally speaking the vocabulary problem, may be solved only if procedures to build the "reference world model" used to score utterances similarity is released by the producer.

On-site applications

Typical applications are:

- control of / access to an equipment
- access to a secured area (nuclear plant, military)
- voice key (home, car)

- mobile telephone, personal assistant (only respond to the voice of his owner)
- Automatic Teller Machine (ATM)
- ...

With on-site applications, the verified person must be physically present in one particular place. It covers typically access control applications and fixed-place banking services (Automatic Teller Machine). These applications are the equivalent of database access and remote transactions over the telephone. However, the differences with telecommunication applications come from 4 facts : the environment factors and the signal bandwidth can be more easily controled, the automatic verification can send an alarm in case of doubt, the customer can bear his voice characteristics on him (on an intelligent card, for instance), and the voice verification technique can be associated more easily to additional identity verification (multimodal) techniques.

access control: open-set speaker id

alternatives:

key, badge, code (could be stolen)

finger print, face recognition, hand shape, signature, retinal scan

make sure that only one user gets access (2 doors, scale)

defeat prerecorded messages (text prompted)

A possible implementation of a voice verification system for money distributors can simply consist of a voice verifier which refuses the transaction (or limitates its amount) if the voice characteristics do not match sufficiently the identity corresponding to the PIN number. Note that, even if a thief steals a credit card and that the PIN number is attached to it, he may still not know the voice of the user, and may have difficulties to get information about it. If he knows the voice of the owner, he may still not be able to imitate it. Even if some false acceptances take place, the amount of fraudulous transactions is necessarily reduced. However, if too many false rejections occur, the bank may loose some of their clients. The probability of an impostor to have a voice similar to the user being small, the system can be quite tolerant and still reduce the number of fraudulous withdrawals, be somehow dissuasive to impostors, without offending a significant part of the regular users. If additional procedures are put into action, such as taking a picture of the user in case of doubt on his identity, or performing some kind of face recognition, dissuasion is reinforced, since it requires a more elaborate and less casual strategy to be resorted to by a possible impostor. For this kind of applications, the voice characteristics of the speaker can be stored on the magnetic tape or the chip of his card, which does not require a centralised storage. The verification must take place in nearly real-time to avoid undesirable delays in the transaction.

For what concerns access control, a reasonable system would be based on a rather strict verification, with a "call for assistance" alternative in case of rejection.

The system of Doddington [Doddington74], located at Texas Instruments, is the first example of access control via speaker verification. It is a vocabulary-dependent system, and speech is parameterised by 14 filter-bank coefficients. Test utterances are composed of a sequence of 4 monosyllabic words. During the training phase, a key-point corresponding to the maximum of energy in each word, is detected, and the spectral vector corresponding to this key-point is stored as the reference of the whole word. It is easy to understand that this key-point is nothing else than the center of the vowel in the monosyllabic word. During the test phase, a simplified version of dynamic programming is implemented to simultaneously locate the 4 key-points in the input utterance and to evaluate the minimum distance corresponding to the best match. The elementary distance measure for the DTW is the Euclidean distance. In real-world conditions, with an appropriate multiple verification strategy, Doddington [Doddington 85] reports in 1985, for the latest version of the system, less than 1 % for both false rejection and false acceptance rates. The operational system is located in a reasonably quiet environment, but the signal is band-passed between 300 Hz and 3000Hz. It is described in details in the previous chapter.

MSS

access to nuclear plants

Until 1990, in France, the available technology was provided by the LIMSI research laboratory [Sorin 90]. An industrial prototype of the SESAME system was expected by the VECSYS company. The SESAME experimental system has been used for about three years at the entrance of the LIMSI building in ORSAY.

Other applications

Among other possible applications of speaker recognition techniques in general, speaker change detection can be used for automatic speaker labeling in recordings or focussing on the actual speaker in videoconferences. For some multi-media applications, it may be very useful to access, speaker by speaker, the recording of a conversation or of a radio or television program. Here, the speech to be processed is reasonably contemporaneous, which makes the algorithms usually more efficient, but it may happen that several speakers talk at the same time.

Finally, we repeat here that research in speaker recognition can be used for speaker selection or adaptation of multi-speaker or speaker-independent speech recognition systems, and can bring improvements to speech processing techniques in general, including synthesis and coding.

Alternative techniques

Security may be achieved in several ways, and voice is only one of the possibilities. We give here a brief overview of the alternative techniques.

The physical supports to protect information or access to unauthorised persons are usually classified in three main families : physical-object elements (key, badge, etc...), information elements (password, combination, etc...) and physical-personal elements (finger-print, voice, DNA, etc...). The ensemble of systems based on this last family are called "biometric verificator" as they are directly tied to some biological characteristics of the (authorised) person. It is a trivial fact that high security systems must use a method based on the union of these three main families. The biometric family is divided into two sub-families: one based on physical characteristics of the person that do not change (unless crude events) in the range of few years ; the other based on behaviour characteristics that may change in relation to humour, ambient, physical state, and so forth. Because of the high rate of variability of the last sub-family, corresponding systems are quite difficult to achieve, especially in automatic mode, without an expert supervision. Just to clarify these two sub-families of biometrics verificators let consider the following list of biometrics verificators:

PHYSICAL	BEHAVIOUR
finger print	voice
retinal scan	signature
face conform	writing
genetic analysis	cardiac rhythm
hand structure	

There is not a unique strategy to select the most appropriate biometric verificator in a given task. Nevertheless, there is a list of properties that must always be considered in designing a security system. Among the most important, we find :

error(s) rates	execution swiftness	access facility
network capability	number of potential users	rate of test per time unit
acceptability	integration facility	cost

We now review some commercial systems that use biometric verificators, with a short description of how they work and their performance.

The IDENTIX of Sunnyvale (California) has developed an automatic system based on finger print acquisition and recognition. After the insertion of a badge, the user had to press his finger in a concave hall, and the finger print is scanned using infra-red rays. Response is given in about 5 seconds. False acceptance error is negligible (zero on 200 trials), while false rejection is quite high, about 10%.

The AUTOSIG SYSTEM of Irving (Texas) produces a system named SIGN/ON. It is a special tablet that records 13 different parameters (pressure, speed, etc.) while the user is writing his signature on a simple sheet of paper positioned on the tablet (of course the user had to use a special pen). Response is given in about 15 seconds depending on the length of the signature. False acceptance on a single trial is 9%, on two trials (or more) error is about 2%.

The ID3D made by RECOGNITION SYSTEM of San Jose (California) is a hit in recognition systems, as it shares many of the properties (say advantages) described above ; it is in fact a low cost system, with a high degree of integration, fast to use, with a high rate of correct recognition and also with a high user acceptability. After the input of the PIN (using a badge, or typing it on a keyboard) user is requested to place his hand on a tablet containing four rungs. When the hand is in the right position, a light is turned on signalling to the user that the infra-red scan is started. No information about the algorithm used in this system is available to our knowledge : we only know that the information is coded on a 9 bytes value, and that the equal error rate has been evaluated about 0.2%, a very impressive value.

The EYEIDENTIFY Co. of Portland (Oregon) has produced several systems based on the retinal scan using infra-red rays. The user is requested to look inside a special device until he focuses a specific figure. Trained persons take about 7 seconds to perform the measure. Glasses do not affect the measure, while contact lenses can not be used. Error rate is negligible, but user acceptability is very low.

Some comparisons between voice based systems and alternative systems have been carried out. Together with the previous systems the SANDIA laboratories of Albuquerque (New Mexico) James [James90] has evaluated two speaker verification systems : the VOXTRON (or Ver-A-Tel) by ALPHA MICROSYSTEM of Santa Clara (California), and the VOICE by ECCO INDUSTRIES of Danvers (Massachusetts). Both systems have almost the same rate of correct recognition (worse than the other systems). From the SANDIA report the voice systems have a high rate of (social) acceptability so as the ID3D, but this last system is preferred as more easy to use.

Generally speaking if we are interested only in performance, a broad table resuming the value for different biometrics verificators may be the one reported by Peckham [Peckham 90]:

modality	false rejection	false acceptance
voice (high quality)	1.0 %	0.1 %
voice (telephone quality)	3.3 %	0.4 %
retinal scan	2.8 %	0.0 %
signature	0.2 %	0.6 %
veincheck	¡ 1.0 %	¡ 1.0 %
handshape	0.1 %	0.1 %

But as noted in the same report, field tests carried out for the US government on speech, signature and fingerprint, resulted in a recommendation to use speech. Note also that the alternative techniques listed above are more appropriate for on-site applications, and that speech remains a unique biometrical identification feature over telecommunication channels, at least until videophone and ISDN become wide-spread. Even so, some of the alternative techniques could only be implemented through individual sensors, which limits their applicability.

Conclusions

Several commercial systems using speaker recognition techniques, are now avalaible on the market, mostly for speaker verification purposes. Similarly to existing algorithms, it is quite difficult to compare them objectively. This fact stresses the need of standard evaluation procedures for these systems : one possible approach is the definition of a standard application. Such a procedure should specify very accurately the nature of the application, and integrate the speaker verification layer into a predefined structure of authentication layers. It is indeed clear from several examples, that the error rate on the pure voice verification algorithm can have very different overall consequences over the global score of an application, depending on the way the voice verification is integrated with other verification layers. A drawback of such an approach is that the application to which the system is targeted may be very different from the reference application, and the evaluation figures may become somehow meaningless.

From the examples above, it is quite evident that voice verification can represent an advantageous additional layer for user authentication. It is also clear that the success of such an implementation relies on an adequate design and engineering of the application, including multiple layer decisions, error recovery strategies and ergonomical considerations. At this stage, additionnal alternative techniques can also be integrated, for increased security. The future outcome of voice verification techniques from a commercial point of view is certainly strongly connected to a reflexion in application design.

To conclude, voice verification technology is certainly at a stage where its future use is partly conditioned to the ability of the suppliers to convince their clients and the customers of their clients of the advantages that can be gained from such techniques. This conviction will only be reinforced by well-targeted application engineering.

Recommendations

12.7 Forensic applications

voice disguise

bomb threat, kidnapping

denonciation

tapping telephone lines

One important field in which speaker recognition is usually involved, is the field of forensic applications. This concerns cases for which victims or witnesses have an auditory memory of the criminal voice and cases for which recordings are available. In the case of recordings, the content of the tapes may be important (who is speaking? what is being said? which sounds could be identified? ...). The tapes should usually be analysed to check whether or not they have been manipulated, erased, recorded over, ... The signal may require enhancement to be comprehensible. In some cases, the voice of a suspect should be compared with a recorded voice. This is a very special case of speaker verification with many specific difficulties:

- the recordings are non contemporaneous
- the original recording could be noisy, distorted, ...
- the original recording could be of short duration
- the voices could be disguised
- the texts are usually different

- the speakers are not cooperative
- ...

Various methodologies for approaching this problem have been proposed ; they may substantially be classified into the following three categories according to the criterion adopted for the analysis and the characterisation of the voice signal : the listener method, the spectrographic method and the semi-automatic method.

12.7.1 Listener method

In the listener method, the decision of similarity is taken by human experts after audition of speech samples. One of the methods is the repeated listening of the available samples by a group of experts looking for similarities in linguistic, phonetic and acoustic features. The voices listening test can be done either sequentially or alternatively or finally, by simultaneous presentation of the voices, if allowed by the instrumentation [Ibba 86]. The acoustic material is carefully prepared, selecting from the database suitable utterances of the unknown and of the reference speaker, eventually augmented with voice samples coming from speakers with the same acoustic-phonetic characteristics. These utterances are then compared two by two, inverting couples of voices from the same speaker. Studies have been carried out to investigate how well listeners can recognise speakers [Stevens 68]. Listener performance is a function of acoustic variables such as the signal to noise ratio, the speech bandwidth, the amount of speech material, complex distortions of speech signal introduced by speech coding transmission systems, etc... Several experiments using different texts for comparing and evaluating the performance under different degraded conditions showed that human listeners are robust speaker recognisers when presented with degraded speech. This is owed to the fact that there are many sources of knowledge that contribute in various ways to speaker recognition, providing weak, moderate or high discrimination power. However, as with any human decision process, it must be stressed that the listener method leads to a subjective decision. Nevertheless, the listener method is still used in some countries for forensic applications as a technique for speaker recognition.

12.7.2 Spectrographic method

The spectrographic method for speaker recognition makes use of an instrument that converts the speech signal into a visual display. For many years the reference instrument was the "Voice Identification Inc., Sound Spectrograph, model 700". This instrument is able to give a permanent record of changing energy-frequency distribution through time of a speech wave. Usually, frequency range is 0-4000Hz, and bandfilter is 300Hz. Because spectrograms are visual representations of the speech signal, they convey information about the message spoken by the speaker as well as about the speaker himself. For this reason, these patterns were thought to be used as a way of identifying speakers. For example, when the recordings of the voice of two individuals are obtained, an examiner may be able to give an opinion about the similarity between two recordings, if there are common phonetic elements between their speech recordings. This method for speaker identification was originally proposed in [Gray 44], but its use for forensic applications was not considered until 1962, when Kersta [Kersta 62] published the results of experiments on one-word spectral comparison in closed-set tests. Further studies were also carried out by Stevens [Stevens68] and by Tosi [Tosi 72], who presented the results of research at Michigan State University on the basis of a "forensic model" with open-set tests. These results have been analysed from Bolt [Bolt70], by observing the error rates of false identification and of false elimination. They observed that the error rate is dependent upon a lot of factors, i.e. different conditions of environment noise, the change in the psychological state of the speaker, his attempts to alter his voice, the recording conditions, the orthophonic or telephonic voice of the talker etc ...; in particular, the error rate is widely dependent on the examiner and is increased by changing from trained to untrained examiners. Owing to these factors and to other restrictive conditions that affect the error rate of examiners, this method is today of no great interest to scientists in speaker recognition tasks.

Both the listening and the spectrographic methods are subjective techniques based, the first one on aural comparison of recordings and the second one on visual examination of spectrograms in order to attribute two voice samples to the same talker. These decisions are taken by one or several experts, according to some process that is clearly impossible to formulate. Moreover, the process that leads to attribute to such or such individual the quality of "expert" is far from being calibrated itself. Very often also, experts are asked to give a probability of confidence on their own judgments. Are forensic speech experts submitted to a bench-mark tests before they are recruited ? Does it make sense to give a figure measuring one's own self-confidence on one's own decision ? There are certainly several issues about forensic applications of human speaker recognition that have to be put in question.

12.7.3 Semi-automatic method

It is clear that some of the research carried out in automatic speaker recognition is targeted towards forensic applications. Most of the methods described in the previous chapter could be adapted to this goal, especially text-independent ones. However, if they offer a certain rate of success (under restrictive conditions), it must be stated clearly that they are far from being faultless, and that the consequence of an erroneous decision can be dramatic. Actually, if some expertises use very basic statistical methods (such as mean and variance computation), there does not seem to be any elaborate speaker recognition system publicly used for forensic applications, but the possibility is not excluded given the certain degree of secret around such applications. Nor it is excluded that experts resort to such systems for making their own decisions, with or without saying it explicitely. It is clear that such systems, whatever level they are used at, should be submitted to objective test protocols designed in agreement with some members of the scientific community. Frequent contacts between speech researchers and members of the Justice and Police institutions would also certainly help to clarify the possibilities and the limits to the use of speech and of speech technology in forensic applications.

12.7.4 Recommendations

For many applications of speaker recognition techniques, a wrong decision has only material consequences. For forensic applications, much more serious aspects are involved.

A database, specific to this area, should be carefully designed, recorded and distributed. It would consist for instance of a large collection of pairs of voice samples, sometimes from the same speaker, sometimes from different speakers. One of the recording could take place in a studio, while the other would be submitted to several kinds of environmental, channel, artificial and intentional distortions. It would be calibrated in such a way that the average auditor can not do better than a random choice.

Such a database would be extremely precious for several purposes. One would be to evaluate forensic speech experts and clarify if they can really do much better than an average auditor. It would also clearly evidence to the forensic professionals, what are the limitations of human and automatic techniques of speaker recognition for such applications.

Some initiative in that direction is currently taking place within the GFCP (Groupe Francophone de la Communication Parlee) of the SFA, but it would certainly be appropriate to start a wider scale action.

12.8 Conclusions

In this chapter, we have identified the main features that characterise the functions of a speaker recognition system. This characterisation leads to distinctions far more accurate than the usual oppositions between speaker identification versus verification on the one hand, and between text-dependent and text-independent systems on the other hand.

We have distinguished 5 tasks covered by the common term of speaker recognition :

- closed-set speaker identification
- open-set speaker identification
- speaker verification
- speaker matching
- speaker change detection

A common distinction of the tasks understood under the general term of speaker recognition, is the opposition between speaker identification and speaker verification, on the one hand, and between text-dependent systems and text-independent systems, on the other hand. If the first opposition is relatively appropriate, the second one does certainly not hold in front of the number of distinct protocols that are covered by each field. A major point of section XXX is that at least 6 different levels of text-dependence must be distinguished. A primary goal when dealing with assessment methodology is to define a non-ambiguous terminology.

We have stressed the existence of several text-dependence levels. As we defined them above, the distinction must be made between :

- text-dependent systems
- user-specific text-dependent systems
- vocabulary-dependent systems
- speech-event-dependent systems
- machine-driven text-independent systems
- user-driven text-independent systems

Especially when a sufficiently large database becomes avalaible, a large amount of basic research is still required to study various aspects of the new material and confirm or refine some previous findings. Knowledge on the exact nature of intra- and inter-speaker variabilities are still very incomplete, and new material tailored to study these phenomena will certainly give a new impulse to fundamental research in this field. A by-product of this research will certainly be some new outcomes in speaker-independent speech recognition, but also in speech synthesis and coding.

Among all the axis of reflexion we have presented here, some may be more prioritary than others, but the definition of such an order can not be decided without deep discussions with many other specialists in the field. We believe however that all of these topics should be taken into account, because they represent different aspects of a same problem, and they should not be addressed independently.

We hope that we have laid significant foundations for further discussion and investigation in the field of speaker recognition assessment methodology, and we strongly recommend that the need of such investigations be clearly acknowledged by concrete support from various institutions, so that it can be carried on.

The section XXX, concerned with algorithmical approaches, illustrates the large amount of work dedicated to speaker recognition in many laboratories across the world, and the wide variety of successful algorithmic approaches to the problem. It shows also that, in spite of this abundance, it is impossible to identify clear advantages for a given family of methods, because of the heterogeneity in databases and test protocols. Conclusive assessment of algorithmic methods relies on the use of common databases and the defininiton of consensual test protocols.

In section YYY, we list and classify some possible applications, together with a few existing prototypes and commercial products, targeted to speaker recognition techniques. Three main fields are considered : telecommunication, on-site and forensic applications. If speaker verification systems are involved in the three of them, applications of speaker identification are mainly linked to the forensic domain. We stress that speaker verification can certainly offer an increased security for user authentication as an additional protection layer, without constituting a factor of discouragement for authorised users : this requires however a careful engineering of the application and a detailed study of ergonomical aspects. In particular, the notion of equal error rate may not be appropriate in a practical application, and a larger rate of false acceptance (combined with a minimal rate of false rejection) may be a good compromise between fraud decrease and user satisfaction. It is certain anyway that speaker recognition will have a role to play in the future, within multi-modal man-machine communication systems. For what concerns forensic applications, we underline the need of actions and contacts between scientists, magistrate and police officials, aimed at clarifying the limits to the use of speaker recognition techniques in this field.

Among the directions mentioned in this chapter toward standardisation of assessment methodologies in speaker verification, we suggest in particular to define standard scoring procedures, common databases (if possible calibrated by human listening tests) and a reference system. The choice of scoring procedures must be the result of a consensus between research laboratories, and can not be done before a wider consultation. Some actions are taking place currently for multi-lingual database collection, and it seems crucial to us that each European language should be represented in this database : this involves a quite large strategic and material effort and requires national and subsidiary European support. The desirable properties of a reference system is its easy implementation and reproductibility. We however do not have enough ground so far to recommend any particular one.

12.9 References

[Atal 76] Atal B.S., "Automatic recognition of speakers from their voices", Proc. IEEE, vol. 64, n. 4, pp. 460, April 1976.

[Doddington 85] Doddington G.R., "Speaker recognition - Identifying people by their voices", Proc. IEEE, vol. 73, n. 11, pp. 1651, November 1985.

[Furui 81] Furui S., "Cepstral analysis technique for automatic speaker verification", IEEE Trans. on ASSP, vol. 29, n. 2, 1981.

[Furui 94] Furui S., "An overview of speaker verification technology", ESCA-ETRW Workshop, Martigny, 1994. Incomplete reference.

[Gish 86] Gish H., Kraner M., Russel W., Wolf J., "Methods and experiments for text-independent speaker recognition over the telephone line", ICASSP, 17.2.1, pp. 865, 1986.

[Oglesby 95] Oglesby J., "What's in a number ? : Moving beyond the equal error rate", To appear in Speech Communication, August 1995. Also published in Martigny ETRW in 1994. Incomplete reference

[Reynolds 95] Reynolds D., "???", to appear in Speech Communication, August 1995. Also published in ETRW Martigny 1994. Incomplete reference.

[Rosenberg 76] Rosenberg A. E., "Automatic speaker verification : a review", Proc. IEEE, vol. 64, n. 4, pp. 475, April 1976.

[Soong 87] Soong F.K., Rosenberg A.E., Juang B.H., Rabiner L.R., "A Vector Quantization approach to speaker recognition", AT&T Technical Journal, vol. 66, issue 2, March-April 1987.

[Abe 90] Abe M., Sagayama S., "Statistical study on voice individuality conservation across different languages", ICSLP, s. 5.8, pp. 157, 1990.

[Artieres 91] Artieres T., Bennani Y., Gallinari P., Montacie C., "Connectionnist and conventional models for text-free talker identification tasks", NEURONIMES, 1991.

* [Atal 68] Atal B.S., "Automatic speaker recognition based on pitch contours", PhD dissertation, Polytechn. Inst. Brooklyn. 1968.

* [Atal 74] Atal B.S., "Effectiveness of LPC characteristics of the speech wave for automatic speaker identification and automatic speaker verification", JASA, vol. 55.

* [Atal 76] Atal B.S., "Automatic recognition of speakers from their voices", Proc. IEEE, vol. 64, n. 4, pp. 460, April 1976.

[Attili 87] Attili J., "On the development of a text-independent real-time speaker verification system", PhD Thesis, ECSE dept, Rensselaer Polytechnic Institute, Sept 1987.

* [Attili 88] Attili J., Savic M.,"A TMS320 based, real-time, text-independent, automatic speaker verification system", ICASSP, s. 13.8, pp. 599, 1988.

[Basztura 91] Basztura C., "Experiments of automatic speaker recognition in open sets", Speech Communication, vol. 10, n. 2, pp. 117, 1991.

[Bennani 90] Bennani Y., Soulie F., Gallinari P., "A connectionist approach to automatic speaker recognition" , ICASSP , s. 5.2, pp. 265, 1990.

* [Bennani 91] Bennani Y., Gallinari P., "On the use of TDNN-extracted features information in a talker identification", ICASSP, s. 6.5, pp. 385, 1991.

* [Bennani 92] Bennani Y. "Speaker identification through a modular connectionist architecture: evaluation on the Timit database", ICSLP, s. 4.4, pp. 607, 1992.

* [Berkley 90] Berkley D., Flanagan J., "Integration of speech recognition, text-to-speech synthesis, and talker verification into a hands free audio/image teleconferencing system (HuMaNet)", ICSLP, s. 20.1, pp. 861, 1990.

[Bernasconi 90] Bernasconi C., "On istantaneous and transitional spectral information for text-dependent speaker verification", Speech Communication, vol. 9, n. 2, pp. 129, 1990.

[Bimbot 92] Bimbot F., Mathan L., De Lima A., Chollet G., "Standard and target driven AR-vector models for speech analysis and speaker recognition", ICASSP, vol. 2, pp. 5, 1992.

** [Bimbot 93] Bimbot F., Mathan L., "Text-free speaker recognition using an arithmetic-harmonic sphericity measure", EUROSPEECH, pp. 169, 1993.

* [Bolt 70] Bolt R.H., "Speaker identification by speech spectrograms : a scientists' view of its reliability for legal purposes", JASA , pp. 597, vol. 47, n. 2 (Part 2), 1970.

[Bonastre 91] Bonastre J., Meloni H., Langlais P., "Analytical strategy for speaker identification", EUROSPEECH, s. 15.3, pp. 435, 1991.

[Bonfiglioli 90] Bonfiglioli M., Lamberti C., "A method for speaker verification based on cepstral measurements", Proc. Verba, Rome, pp. 197, January 1990.

[Bonneau 87] Bonneau H., Gauvain J., "Vector quantization for speaker adaptation", ICASSP , s. 34.6, pp. 1434, 1987.

* [Bricker 71] Bricker P.D., "Statistical techniques for talker identification", Bell System Technical Journal, vol. 50, pp. 1427, 1971.

[Broeders 89] Broeders T., Rietveid T., "Segmental marking as a cue in auditory voice identification of telephone speech", EUROSPEECH, vol. I, pp. 71, 1989.

[Buck 85] Buck J., Burton D., Shore J., "Text-dependent speaker recognition using Vector Quantization", ICASSP, s. 11.5.1, pp. 391, 1985.

[Bugne 79] Bugne E., "Identification judiciaire de la voix par ordinateur", Rev. Int. Pol. Crim. 332: 254, 1979.

[Burr 92] Burr D., "On text-independent talker identification with simulated telephonequality", speech 1st IEEE workshop on I.V.T.T.A., Piscataway, October 1992.

[Carbonell 65] Carbonell J. R., Grignetti M. C., Stevens K. N., Williams C. E., Woods B., "Speaker authentication techniques", Rpt. 1296 prepared under Contract No. DA-28-043-AMC-00116 (E) by Bolt Beranek and Newman, Inc., Cambridge, Mass, 1965.

* [Carey 91] Carey M., Parris E., Bridle J., "A speaker verification system using alpha-nets", ICASSP, s. 6.8, pp. 397, 1991.

[Chang 92] Chang H., "Augmented phonetic map for voice verification", ICASSP, II, pp. 169, 1992.

[Chen 87] Chen S., Lin M., "On the use of pitch contour of mandarine speech in text-independent speaker identification", ICASSP, s. 34.1, pp. 2416, 1987.

[Chibelushi 93] Chibelushi C., Mason J., Deravi F., "Integration of acoustic and visual speech for speaker recognition", EUROSPEECH, pp. 157, 1993.

[Childers 88] Childers D., Wu K., Bae K., Hicks D., "Automatic recognition of gender by voice", ICASSP, s. 13.9, pp. 603, 1988.

[Colombi 93] Colombi J., Anderson T., Rogers S., Ruck D., Warhola G. , "Auditory model representation for speaker recognition", ICASSP, II, pp. 700, 1993.

[Corsi 79] Corsi P., "Reconnaissance automatique du locuteur", these de Docteur-Ingenieur, INPG, Grenoble, 1979.

[Das 69] Das S.D., Mohn W.S. , "Pattern recognition in speaker verification", Fall Joint Computer Conference, 1969.

[De Veth 93] De Veth J., Gallopyn G., Bourlard H., "Speaker verification over telephone channels based on concatenated phonemic hidden markov models", EUROSPEECH, pp. 2279, 1993.

* [Doddington 74] Doddington G.R., "A method of speaker verification", JASA, vol. 49, 1974.

[Doddington 83] Doddington G., "Voice authentication gets the go-ahead for security systems", Speech Technology, pp. 14-23, september 1983.

* [Doddington 85] Doddington G.R., "Speaker recognition - Identifying people by their voices", Proc. IEEE, vol. 73, n. 11, pp. 1651, November 1985.

[Eatock 90] Eatock J., Mason J., "Speaker-dependent speech classification in speaker recognition", Proc. ESCA ETRW on speaker characterisation, Edimburgh, pp. 94, 1990.

[Eatock 91] Eatock J., Mason J., "Automatically focusing on good discriminating speech segments in speaker recognition", ICSLP, s. 5.2, pp. 133, 1990.

[Eatock 92] Eatock J., Mason J., "Phoneme performance in speaker recognition", ICSLP, s. 1.21, pp. 1411, 1992.

[Elvira 93] Elvira J.M., Carrasco R.A., "Neural networks for speech and speaker recognition through a digital telephone exchange", EUROSPEECH, pp. 2291, 1993.

[El Chafei 78] El Chafei C.,"Un systeme de reconnaissance automatique de locuteurs sur mini-ordinateur", These de Docteur-Ingenieur, Universite Paris-Sud, Orsay, 1978.

[Everett 85] Everett S., "Automatic speaker recognition using vocoded speech", ICASSP, 11.2.1, pp. 383, 1985.

[Fakotakis 93] Fakotakis N., Tsopanoglou A., Kokkinakis G., "A text-independent speaker recognition system based on vowel spotting", Speech Communication, vol. 12, n.1, pp.57, 1993.

[Federico 87] Federico A., Ibba G., Paoloni A., "A new automated method for reliable speaker identification and verification over telephone channels", ICASSP, 34.12, pp. 1457, 1987.

* [Federico 89] Federico A., et al. "Comparison between automatic methods and human listeners in speaker recognition tasks", EUROSPEECH, pp. 279-282, 1989.

[Federico 93] Federico A., Paoloni A., "Bayesian decision in the speaker recognition by acoustic parametrization of voice samples over telephone lines", EUROSPEECH, pp. 2307, 1993.

[Feiz 85] Feiz W., Degeorge M., "A speaker verification system for access-control", ICASSP , 11.7.1, pp.399, 1985.

* [Fisher 86] Fisher W., Zue V. Bernstein J., Pallet D., "An acoustic-phonetic database", JASA, suppl. A, 81, S92, 1986.

* [Furui 73] Furui S., Itakura F. "Talker recognition by statistical features of speech", Electron. Comm. Jap., vol. 56A, pp. 62, November 1973.

* [Furui 81] Furui S., "Cepstral analysis technique for automatic speaker verification", IEEE Trans. on ASSP, vol. 29, n. 2, 1981.

[Furui 86] Furui S., "Research on individuality features in speech waves and automatic speakers recognition techniques", Speech Communication, vol. 5, n. 2, pp. 183, 1986.

[Furui 90] Furui S., "Speaker-dependent features extraction, speaker recognition and speaker adaptation techniques", Proc. Verba, Rome, pp. 164-170, January 1990.

[Furui 92] Furui S., Matsui T., "Free-text speaker recognition methods using phoneme class models", 1st IEEE Workshop on I.V.T.T.A., Piscataway, october 1992.

[Fussel 91] Fussel J., "Automatic sex identification from short segment of speech", ICASSP, s. 6.11, pp. 409, 1991.

[Gaganelis 91] Gaganelis D., Frangoulis E., "A novel approach to speaker verification", ICASSP, s. 6.2, pp. 373, 1991.

[Giannini 89] Giannini A., Pettorini M., Cinque U., "Speaker's identification by voice", EUROSPEECH, vol. 1, pp. 283, 1989.

[Gish 85] Gish H., Karnofsky K., Krasner M., Rouco, Schwartz R., Wolf J., "Investigation of text-independent speaker identification over telephone channels", ICASSP, 11.1.1, pp. 379, 1985.

[Gish 86] Gish H., Kraner M., Russel W., Wolf J., "Methods and experiments for text-independent speaker recognition over the telephone line", ICASSP, 17.2.1, pp. 865, 1986.

* [Gish 90] Gish H.,"Robust discrimination in automatic speaker identification", ICASSP, s. 5.9, pp. 289, 1990.

** [Goldman 93] Goldman J.P., Chollet G., "Voice transformations for the evaluation of speaker verification systems", NATO-ASI, pp. 125, Bubion, 1993.

[Gong 90] Gong Y., Haton J., "Text-independent speaker recognition by trajectory space comparison", ICASSP, s. 5.8, pp. 285, 1990.

[Gong 92] Gong Y., Haton J., "Non-linear vectorial interpolation for speaker recognition", ICASSP, II, pp. 173, 1992.

* [Gray 44] Gray C. H, Kopp G. A., "Voiceprint Identification", Bell Telephone Report, Bell Laboratories, 1944.

[Grenier 77] Grenier Y., "Identification du locuteur et adaptation au locuteur d'un systeme de reconnaissance phonemique", These de Docteur-Ingenieur, ENST-E-77005, Paris, 1977.

* [Grenier 80] Grenier Y., "Utilisation de la prediction lineaire en reconnaissance et adaptation au locuteur", Proc. XIth JEP, pp. 163, 1980.

[Hammond 90] Hammond G., "Deploying speech technology in telephone banking services", Speech Technology, pp. 30-33, april 1990.

[Hangai 90] Hangai S., Miyauchi K., "Speaker identification based on multiple excitation and LPC vocal-tract model", ICSLP, s. 27.21, pp. 1269, 1990.

[Hangai 92] Hangai S., Sugiyama S., Miyauchi K., "Speaker verification using locations and sizes of multipulses on neural networks", ICSLP, s. 1.30, pp. 1439, 1992.

[Hannah 93] Hannah M.I., Sapeluk A.T., Damper R.I., Roger I.M., "The effect of utterance length and content on speaker-verifier performance", EUROSPEECH, pp. 2299, 1993.

[Harmegnies 91] Harmegnies B., Bruyninckx M., Llisterri J., Poch D., "Effects of language change on voice quality in bilingual speakers. Corpus content effect", EUROSPEECH, s. 6.8, pp. 165, 1991.

[Hasa 92] Hasa E., Onsy A., "Perceptual bases of speaker recognition", Proc. ESCA ETRW on comp.sp.sig.rep., pp. 321, 1992.

* [Hattori 92] Hattori H., "Text-independent speaker recognition using neural networks", ICASSP, II, pp. 153, 1992.

[Higgins 86] Higgins A., Wohlford R., "A new method of text-independent speaker recognition", ICASSP, 17.3.1, pp. 869, 1986.

[Higgins 90] Higgins A., Lieu K., "A single-board speaker recognition system for the Macintosh", Proc. Speech Tech, pp. 205, New York, April 1990.

[Higgins 91] Higgins A., Bahler L., "Text-independent speaker verification by discriminator counting", ICASSP, s. 6.10, pp. 405.

[Higgins 93] Higgins A., Bahaler L., Porter J., "Voice identification using nearest-neighbor distance measure", ICASSP, II, pp. 375, 1993.

[Hoda 89] Hoda H.,"On the use of the information on individual speaker's position in the parameter space for speaker recognition", ICASSP, s. 10B.1, pp. 516, 1989.

[Homayounpour 93] Homayounpour M.M., Goldman J.P., Chollet G., Vaissiere J., "Performance comparison of machine and human speaker verification", EUROSPEECH, pp. 2295, 1993.

** [Homayounpour 93] Homayounpour M.M., Goldman J.P., Chollet G., "Machine versus human speaker verification", IAFP Conference, Trier, 1993.

* [Hunt 91] Hunt A., "New commercial applications of telephone-network-based speech recognition and speaker verification", EUROSPEECH, s. 15.2, pp. 431, 1991.

* [Ibba 86] Ibba G., Paoloni A., Di Carlo A. , Federico A., "Proposal for a speaker identification methodology by listening session", Proc. of AIA '86, Sorrento, April 1986. (in italian).

[Ibba 90] Ibba G., Paoloni A., "Trends in speaker recognition", Proc. Verba, Rome, pp. 185-196, January 1990.

[Irvine 93] Irvine D.A., Owens F.J., "A comparison of speaker recognition techniques for telephone speech", EUROSPEECH, pp. 2275, 1993.

* [James 90] James P., et al., "A performance evaluation of biometrics identification devices", Sandia National Laboratories, NM 87185, July 1990.

[Jasiuk 87] Jasiuk M., Goncharoff V., Damoulakis J., "Improved speech modification method", ICASSP , 34.17, 1465,1987.

* [Joachim 90] Joachim K., "Overview of telephone based speech I/O application in Germany: present and future", Proc. of Speech Tech 90, New York, pp. 162-167, April 90.

[Jou 90] Jou I., Lee S., Lin S., Tseng C., Yu S., Tsay Y, "A neural network based speaker verification system", ICSLP, s. 27.22, pp. 1237, 1990.

[Juang 90] Juang B., Soong F., "Speaker recognition based on source coding approaches", ICASSP , s. 5.4, pp. 613, 1990.

[Kao 92] Kao Y., Rajasekaran P., Baras J., "Free-text speaker identification over long distance telephone channel using hypothesized phonetic segmentation", ICASSP, II, pp. 177, 1992.

[Kao 92] Kao Y., Rajasekaran P., Baras J., "Robust free-text speaker identification over long distance telephone channels", 1st IEEE workshop on I.V.T.T.A., Piscataway, October 1992.

[Kao 93] Kao Y., Baras J., "Robustness studies in free-text speaker identification and verification", ICASSP, II, pp. 379, 1993.

* [Kersta 62] Kersta L. G., "Voiceprint infallibility", Meeting of Acoust. Soc. Am., Seattle, USA, 1962.

[Kraayevel 91] Kraayevel H., Rietveld A., Van Heuven V., "Speaker characterisation in dutch using prosodic parameters", EUROSPEECH, s. 15.1, pp. 427, 1991.

[Krasner 84] Krasner M., Wolf J., Karnofsky K.,"Investigation of the text-independent speaker identification techniques under conditions of variable data", ICASSP, 18B.5.1, 1984.

[Kreiman 91] Kreiman J., Papcun G., "Comparing discrimination and recognition of unfamiliar voices", Speech Communication, vol. 10, n. 3, pp. 265, 1991.

[Kuo 92] Kuo J., Lee C., Rosenberg A., "Speaker set identification through speaker group modeling", ICSLP, s. 1.16, pp. 1391, 1992.

[Ladefoged 80] Ladefoged P., Ladefoged J., "The ability of listeners to identify voices", UCLA working papers in phonetics, vol. 49, pp. 43, 1980.

[Lhote 90] Lhote E., Haidar L., "Speaker verification by a vocal proxemy cue", Proc. ESCA ETRW on speaker characterisation, Edimburgh , pp. 149, 1990.

[Lhote 91] Lhote E., Haidar L., "An objective and subjective approach to speaker recognition", Proc. of the XIIth ICPhS, vol. 4, pp. 42, Aix-en provence, 1991.

[Li 66] Li K.P. et al., "Experimental studies in speaker verification using an adaptive system", JASA, vol. 40, 1966.

[Li 87] Li K., "Separating phonetic and speaker features of vowels in formant space", ICASSP , 34.18, pp. 1469, 1987.

[Li 88] Li K., Porter J., "Normalization and selection of speech segments for speaker recognition scoring", ICASSP , s. 13.7, pp. 595, 1988.

[Liu 90] Liu C., Lin M., Wang W., Wang H., "Study of Line Spectrum Pair frequencies for speaker recognition", ICASSP , s. 5.6, pp. 277, 1990.

[Li 90] Li K., "Automatic speaker verification / identification : a tutorial review", Proc. Verba, Rome, pp. 171-184, January 1990.

[Lipeika 93] Lipeika A., Lipeikiene J., "The use of pseudostationary segments for speaker identification", EUROSPEECH, pp. 2303, 1993.

[Lummis 72] Lummis R.C., "Speaker verification : a step toward the checkless society", Bell Labs Record, Sept. 1972.

* [Mariani 84] Mariani J., Gauvain J.L., Soury J.L., "Un systeme de verification du locuteur", Proc. XIIIth JEP, pp. 113, Bruxelles, 1984.

[Mason 89] Mason J., Oglesby J., Xu L., "Codebooks to optimise speaker recognition", EUROSPEECH, vol. I, pp. 267, 1989.

[Mason 91] Mason J., Zhang X., "Velocity and acceleration features in speaker recognition", ICASSP, s. 6.1, 1991

[Matsui 90] Matsui T., Furui S., "Text-independent speaker recognition using vocal tract and pitch information", ICSLP, s. 5.3, pp. 137, 1990.

[Matsui 91] Matsui T., Furui S., "A text-independent speaker recognition method robust against utterance variations", ICASSP , s. 6.3, pp. 377, 1991.

[Matsui 92] Matsui T., Furui S., "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs", ICASSP, II, pp. 157, 1992.

[Matsui 92] Matsui T., Furui S., "Speaker recognition using concatenated phoneme models", ICSLP , s. 4.3, pp. 603, 1992.

[Matsui 93] Matsui T., Furui S., "Concatenated phoneme models for text-variable speaker recognition", ICASSP, II, pp. 391, 1993.

[Matsumoto 88] Matsumoto H., "Text-independent speaker identification from short utterances based on piece-wise discriminant analysis", Computer Speech & Language, vol. 3, n. 2, pp. 133, April 1988.

[Matsumoto 90] Matsumoto H., Moue H., "A minimum distortion spectral mapping applied to voice quality conversation", ICSLP, s. 5.9, pp. 161, 1990.

[Maturi 90] Maturi P., "Speaker identification in forensic : a simulation experiment", Proc. ESCA-ETRW on speaker characterisation, Edimburgh, pp. 161, 1990.

[Millar 90] Millar J., Hawkins S., "Selecting representative speakers", Proc. ESCA ETRW on speaker characterisation, Edimburgh, pp. 161, 1990.

[Montacie 92] Montacie C., Deleglise P., Bimbot F., Caraty M.J., "Cinematic techniques for speech processing : temporal decomposition and multivariate linear prediction", ICASSP, vol I, pp. 153, 1992.

* [Montacie 92] Montacie C., Le Floch J., "AR-vector models for free-text speaker recognition", ICSLP, s. 4.5, pp. 611, 1992.

[Montacie 93] Montacie C., Le Floch J.L., "Discriminant AR-vector models for free-text speaker verification", EUROSPEECH, pp. 161, 1993.

* [Moody 91] Moody A., "Speaker verification", Ensigma Ltd. Internal Report, January 1991.

[Naik 86] Naik J., Doddington G., "High performance speaker verification using principal spectral components", ICASSP , 17.6.1, pp. 881, 1986.

[Naik 87] Naik J., Doddington G., "Evaluation of high performance speaker verification system for access control", ICASSP , 34.15, pp. 2392, 1987.

[Naik 89] Naik J., Netsch L., Doddington G., "Speaker verification over long distance telephone lines", ICASSP, s. 10B.3, pp. 524, 1989.

[Nakasone 88] Nakasone H., Melvin C., "Computer assisted voice identification system (CAVIS)", ICASSP , s. 13.5, pp. 587, 1988.

[Netsch 92] Netsch L., Doddington R., "Speaker verification using temporal decorrelation post-processing", ICASSP, II, pp. 181, 1992.

[Ney 82] Ney H., "Automatic speaker recognition using time alignment of spectrograms", Speech Communication, vol. 1, n. 2, pp.135, 1982.

[Noda 90] Noda H., Yanagida M., "Extraction of phoneme-dependent individuality using HMM-based segmentation for text-independent speaker recognition", ICSLP, s. 5.1, pp. 129, 1990.

* [O'Shaughnessy 86] O'Shaughnessy D., "Speaker recognition", IEEE-ASSP Magazine, pp. 4, October 1986.

* [Oglesby 90] Oglesby J., Mason J., "Optimization of neural models for speaker identification", ICASSP, s. 5.1, 261, 1990.

* [Oglesby 91] Oglesby J., Mason J., "Radial basis function networks for speaker recognition", ICASSP, s. 6.7, pp. 395, 1991.

[Openshaw 92] Openshaw J., Sun Z., Mason J., "A comparison of feature performance under degraded speech in speaker recognition", Proc. ESCA-ETRW on speech Proc. in AD. CON., Cannes, pp. 119, 1992.

[Openshaw 93] Openshaw J., Sun Z., Mason J., "A comparison of composite features under degraded speech in speaker recognition", ICASSP, II, pp. 371, 1993.

[Papamichalis 84] Papamichalis P., Doddington G., "A speaker recognizability test", ICASSP , 18B.6.1, 1984.

[Pascal 89] Pascal D., Thill C., Boyer M., "Perceptive similarity of male voices: correlation with acoustic measures", EUROSPEECH, vol. II, pp. 123, 1989.

* [Peckham 90] Peckham J., "An overview of speaker verification technology and application over the telephone", Proc. Voice System Worldwide, pp. 166, 1990.

[Peckham 90] Peckham J., "An overview of telephone Speech I/O Application in UK", Proc. of SPEECHTECH90, New York, pp. 190-198, April 1990.

[Pisoni 90] Pisoni D., "Effects of talker variability on speech perception : implications for current research and theory", ICSLP, vol. 2, pp. 1399, 1990.

* [Poritz 82] Poritz A.B., "Linear predictive Hidden Markov Models and the speech signal", Proc. ICASSP 82, vol. 2, pp. 1291, 1982.

* [Pruzansky 63] Pruzansky S., "Pattern matching procedures for automatic talker recognition", JASA, vol. 35, n. 3, pp. 354, 1963.

* [Reich 81] Reich A., "Detecting the presence of vocal disguise in the male voice", JASA, vol. 69, pp.1458-1461, 1981.

[Repp 85] Repp B., Williams D., "Categorical trends in vowels imitation : preliminary observation from a replication experiment", Speech Communication, vol. 4, n. 1-3, pp. 105, 1985.

[Repp 87] Repp B., Williams D., "Categorical tendencies in imitating self-produced isolated vowels", Speech Communication, vol. 6, n. 1, pp. 1, 1987.

[Reynolds 92] Reynolds D., Rose R., "An integrated speech-background model for robust speaker identification", ICASSP, II, pp. 185, 1992.

[Rocchi 89] Rocchi C., Mumolo E., "A new method for performing weighted distances for speaker authentication", EUROSPEECH, vol. I, pp. 275, 1989.

* [Rosenberg 73] Rosenberg A., "Listener performane in speaker verification tasks", IEEE Trans. Audio Electroacoustic, vol. au-21, pp. 221-225, 1973.

* [Rosenberg 76] Rosenberg A. E., "Automatic speaker verification : a review", Proc. IEEE, vol. 64, n. 4, pp. 475, April 1976.

[Rosenberg 86] Rosenberg A., Soong F.,"Evaluation of a Vector Quantization talker recognition system in a text independent and text dependent modes", ICASSP, 17.4.1, pp. 873, 1986.

[Rosenberg 88] Rosenberg A., Soong F.,"Evaluation of a Vector Quantization talker recognition system in text independent and text dependent modes", Computer Speech & Language, vol. 2, n. 3/4, pp. 143, 1988.

[Rosenberg 90] Rosenberg A., Delong J., Lee C., Juang B., Soong F., "The use of cohort normalised scores for speaker verification", ICSLP, s. 4.2, pp. 599, 1992.

[Rosenberg 90] Rosenberg A., Lee C., Soong F., McGee M., "Experiments in automatic talker verification using sub-word unit Hidden Markov Models", ICSLP, s. 5.4, pp. 141, 1990.

* [Rosenberg 90] Rosenberg A., Lee C., Soong F., "Sub-word unit talker verification using Hidden Markov Models", ICASSP, s. 5.3, pp. 269, 1990.

* [Rosenberg 91] Rosenberg A., Lee C., Gokoen S., "Connected word talker verification using whole word Hidden Markov Models", ICASSP, s. 6.4, pp. 381, 1991.

[Rose 90] Rose R., Reynolds D., "Text independent speaker identification using automatic acoustic segmentation", ICASSP, s. 5.10, pp. 293, 1990.

[Rose 91] Rose R., Fitzmaurice J., Hofstetter E., Reynolds D., "Robust speaker identification in noisy environment using noise adaptive speaker models", ICASSP, s. 6.9, pp. 401, 1991.

* [Rudasi 91] Rudasi L., Zahorian S., "Text-independent talker identification with neural networks", ICASSP, s. 6.6, pp. 389, 1991.

* [Sambur 72] Sambur M.R., "Speaker recognition and verifiation using Linear Prediction Analysis", PhD MIT, Cambridge, 1972.

[Savic 90] Savic M., Gupta S., "On the use of HMM based constrained clustering for speaker verification", Proc. speech tech, pp. 128, New York, April 1990.

* [Savic 90] Savic M., Gupta S., "Variable parameter speaker verification system based on Hidden Markov Modeling", ICASSP, s. 5.7, pp. 281, 1990.

[Savic 92] Savic M., Sorensen J., "Phoneme based speaker verification", ICASSP, II, pp. 165, 1992.

[Shevchenko 91] Shevchenko T., Skopintseva T., "Effects of social and regional backgrounds on LTAS in British English", EUROSPEECH, s. 6.10, pp 169, 1991.

[Shinan 86] Shinan L., Almeida A., "The effects of voice disguise upon formant transition", ICASSP, 17.1.1, pp. 885, 1986.

[Shridhar 82] Shridhar M., Baraniecki M., Mohankrishnan N., "A unified approach to speaker verification with noisy speech inputs", Speech Communication, vol. 1, n. 2, pp. 103, 1982.

[Shridhar 82] Shridhar M., Mohankrishnan N., "Text-independent speaker recognition : a reviev and some new results", Speech Communication, vol. 1, n. 3/4, pp. 257, 1982.

[Siu 92] Siu M., Yu G., Gish H., "An unsupervised, sequential learning algorithm for the segmentation of speech waveform wih multiple speakers", ICASSP, II, pp. 189, 1992.

* [Skinner 92] Skinner T., Holt J., Nguyen N., "Automatic identity confirmation and adaptive solutions", Speech Technology, pp. 106-111, February 1992.

[Soong 85] Soong F., Rosenberg A., Rabiner L., Juang B., "A Vector Quantization approach to speaker recognition", ICASSP, 11.4.1, pp. 387, 1985.

[Soong 86] Soong F., Rosenberg A., "On the use of instantaneous and transitional spectral information in speaker recognition", ICASSP, 17.5.1, pp. 877, 1986.

* [Soong 87] Soong F.K., Rosenberg A.E., Juang B.H., Rabiner L.R., "A Vector Quantization approach to speaker recognition", AT&T Technical Journal, vol. 66, issue 2, March-April 1987.

[Soong 88] Soong F., Rosenberg A., "On the use of instantaneous and transitional spectral information in speaker recognition", IEEE Trans. on ASSP, vol. 36, n. 6, 1988.

* [Sorin 90] Sorin C., Soubigou A., "Overview of telephone speech Input/TTS output application in France : present and future", Proc. of SPEECHTECH90, New York, pp. 178-181, April 1990.

* [Stevens 68] Stevens K., Williams C. E., Carbonell J. L., Woods B., "Speaker authentication and identification : a comparison of spectrographic and auditory presentations of speech material", J.A.S.A., vol. 44, pp. 1596-1607, 1968.

[Sugiyama 93] Sugiyama M., Murakami J., Watanabe H., "Speech segmentation and clustering based on speaker features", ICASSP, II, pp. 395, 1993.

[Sun 93] Sun Z.P., Mason J.S., "Combining features via LDA in speaker recognition", EUROSPEECH, pp. 2287, 1993.

[Suomi 84] Suomi K., "On speaker and phoneme information conveyed by vowels : a whole spectrum approach to the normalization problem", Speech Communication, vol. 3, n. 3, pp.199, 1984.

[Takagi 86] Takagi T., Kuwabara H., "Contributions of pitch, formant frequency and bandwidth to the perception of voice-personality", ICASSP , 17.8.1, pp. 889, 1986.

[Thevenaz 90] Thevenaz P., Hugli H., "Combining four text independent speaker recognition methods", Proc. ESCA-ETRW on speaker characterisation, Edimburgh, pp. 187, 1990.

[Thompson 93] Thompson J., Mason J.S., "Within class optimization of cepstra for speaker recognition", EUROSPEECH, pp. 165, 1993.

[Tishby 88] Tishby N., "Information theoretic factorization of speaker and language in Hidden Markov Models with application to speaker recognition", ICASSP , s. 2.7, pp. 87, 1988.

* [Tosi 72] Tosi O., Oyer H., Asbrook W.L. , Pedrey W. , Nicol C., Nash E., "Experiment of Voice Identification", JASA., 51, pp. 2030-2043, 1972.

[Tseng 92] Tseng B., Soong F., Rosenberg A., "Continuous probabilistic acoustic map for speaker recognition", ICASSP, II, 161, 1992.

[Wang 90] Wang R., He L., Fujisaki H., "A weighted distance measure based on the fine structure of features, application to speaker recognition", ICASSP , s. 5.5, pp. 273, 1990.

[Wang 93] Wang H., Chen M., Yang T., "A novel approach to the speaker identification over telephone networks", ICASSP, II, pp. 407, 1993.

[Webb 93] Webb J., Rissanen E., "Speaker identification experiment using HMMs", ICASSP, II, pp. 387, 1993.

[Wolf 72] Wolf J.J., "Efficient acoustic parameters for speaker recognition", JASA, vol. 51, n. 6, 1972.

[Xu 89] Xu L., Mason J., "Instantaneous and transitional perceptually-based features in speaker identification", EUROSPEECH, vol. I, pp. 271, 1989.

[Xu 89] Xu L., Oglesby J., Mason J., "The optimization of perceptually-based features for speaker identification", ICASSP , s. 10B.2, pp. 520, 1989.

[Xu 91] Xu L., Mason J., "Optimization of perceptually-based spectral transforms in speaker identification", EUROSPEECH, s. 15.4, pp. 439, 1991.

[Yegnanarayama 92] Yegnanarayama B., Madhukumar B., Ramachandran V., "Robust features for application in speech and speaker recognition", Proc. ESCA-ETRW on Speech Proc. in AD. CON., Cannes, pp. 97, 1992.

[Yin 90] Yin H., Zhou T., "Speaker recognition using static and dynamic cepstral feature by a learning neural network", ICSLP, s. 27.23, pp. 1277, 1990.

[Yu 93] Yu G., Gish H., "Identification of speaker engaged in dialog", ICASSP, II, pp. 383, 1993.

[Yuan 93] Yuan Z., Yu C., et al., "Text-independent speaker identification using fuzzy mathematical algorithm", ICASSP, II, pp. 403, 1993.

[Zalewski 89] Zalewski J., "Text dependent speaker recognition in noise", EUROSPEECH, vol. I, pp. 287, 1989.

[Zhang 91] Zhang X., Mason J., Andrews E., "Multiple dynamic features to enhance neural net based speaker verification", EUROSPEECH, s. 49.2, pp. 1407, 1991.

* [Zheng 88] Zheng Y., Yuan B., "Text-dependent speaker identification using circular Hidden Markov Model", ICASSP , s. 13.3, pp. 580, 1988.

Test utterance	Recognised speaker at rank j: $j=1$	$j=2$	$j=3$	$j=4$	$j=5$	id. rank $\rho\,(x_i^k)$	conf. rank at 50 %	conf. rank at 95 %
x_1^1	$\boxed{X_1}$	X_2	X_4	X_3	X_5	1		
x_1^2	$\boxed{X_1}$	X_4	X_2	X_5	X_3	1		
x_1^3	X_4	$\boxed{X_1}$	X_3	X_2	X_5	2		
x_1^4	X_4	$\boxed{X_1}$	X_2	X_5	X_3	2		
x_1^5	$\boxed{X_1}$	X_4	X_5	X_2	X_3	1		
x_1^6	X_4	$\boxed{X_1}$	X_2	X_3	X_5	2		
x_1^7	$\boxed{X_1}$	X_4	X_3	X_2	X_5	1		
x_1^8	X_4	$\boxed{X_1}$	X_3	X_2	X_5	2		
$\eta_{1\,(j)}$	50 %	100 %	100 %	100 %	100 %	$\longrightarrow$	$\rho_{1\,(50\,\%)} = 1$	$\rho_{1\,(95\,\%)} = 2$
x_2^1	X_4	X_5	$\boxed{X_2}$	X_3	X_1	3		
$\eta_{2\,(j)}$	0 %	0 %	100 %	100 %	100 %	$\longrightarrow$	$\rho_{2\,(50\,\%)} = 3$	$\rho_{2\,(95\,\%)} = 3$
x_3^1	X_2	X_1	$\boxed{X_3}$	X_4	X_5	3		
x_3^2	$\boxed{X_3}$	X_2	X_1	X_4	X_5	1		
$\eta_{3\,(j)}$	50 %	50 %	100 %	100 %	100 %	$\longrightarrow$	$\rho_{3\,(50\,\%)} = 1$	$\rho_{3\,(95\,\%)} = 3$
x_4^1	$\boxed{X_4}$	X_1	X_3	X_2	X_5	1		
x_4^2	$\boxed{X_4}$	X_2	X_5	X_3	X_1	1		
x_4^3	$\boxed{X_4}$	X_1	X_5	X_3	X_2	1		
x_4^4	$\boxed{X_4}$	X_2	X_3	X_5	X_1	1		
x_4^5	$\boxed{X_4}$	X_2	X_5	X_3	X_1	1		
$\eta_{4\,(j)}$	100 %	100 %	100 %	100 %	100 %	$\longrightarrow$	$\rho_{4\,(50\,\%)} = 1$	$\rho_{4\,(95\,\%)} = 1$
x_5^1	$\boxed{X_5}$	X_2	X_3	X_4	X_1	1		
x_5^2	$\boxed{X_5}$	X_3	X_1	X_4	X_2	1		
x_5^3	X_2	X_4	X_3	X_1	$\boxed{X_5}$	5		
x_5^4	$\boxed{X_5}$	X_2	X_1	X_4	X_3	1		
$\eta_{5\,(j)}$	75 %	75 %	75 %	75 %	100 %	$\longrightarrow$	$\rho_{5\,(50\,\%)} = 1$	$\rho_{5\,(95\,\%)} = 5$
							$\downarrow$	$\downarrow$
							$\boxed{\bar{\rho}_{\,(50\,\%)} = 1.2}$	$\boxed{\bar{\rho}_{\,(95\,\%)} = 2.8}$
$\eta_{(j)}$	65 %	85 %	95 %	95 %	100 %	$\longrightarrow$	$\boxed{\rho_{\,(50\,\%)} = 1}$	$\boxed{\rho_{\,(95\,\%)} = 3}$

Figure 12.2: An example illustrating how to score confidence ranks in closed-set speaker identification. Only for speaker X_2 is there less than half of the time the right speaker in first position ($\rho_{i\,(50\,\%)} = 1$ except for $i = 2$). On the test set, the list of the 3 best candidates would have to be kept in order to be sure that the right speaker is in the list 95 % of the time ($\rho_{test\,(95\,\%)} = 3$).

	genuine attempts $m = 3$		
	X_1	X_2	X_3
	Male		Female
$\hat{V}$	6	2	4
$\hat{\hat{V}}$	$\boxed{3}$	$\boxed{0}$	$\boxed{3}$
c_i	9	2	7
c	18		
m^*	3		
α_i	$\frac{3}{9} = \frac{1}{3}$	0	$\frac{3}{7}$
$\bar{\alpha}$	$\frac{1}{3}\left[\frac{3}{9} + \frac{0}{2} + \frac{3}{7}\right] = \frac{16}{63} \approx 25\,\%$		
α	$\frac{3+0+3}{18} = \frac{1}{3} \approx 33\,\%$		
$m^*_M \mid m^*_F$	2		1
$\bar{\alpha}_M \mid \bar{\alpha}_F$	$\frac{1}{2}\left[\frac{1}{3} + 0\right] = \frac{1}{6}$		$\frac{3}{7}$
$\bar{\alpha}_{MF}$	$\frac{1}{2}\left[\frac{1}{6} + \frac{3}{7}\right] = \frac{25}{84} \approx 30\,\%$		
$\bar{\mu}_i$	$\frac{7}{16} \approx 0.44$	0	$\frac{9}{16} \approx 0.56$
μ_i	$\frac{1}{2} = 0.50$	0	$\frac{1}{2} = 0.50$

Figure 12.3: Out of 18 genuine attempts, 6 false rejections are observed, hence the *test set false acceptance rate* $\alpha = \frac{1}{3} \approx 33\,\%$. Nevertheless, the 3 false rejections out of 9 trials for X_1 do not have the same impact on the *average false rejection rate* $\bar{\alpha} \approx 25\,\%$ than the 3 false rejections out of 7 trials for X_3. In fact, whereas X_2 seems to be the most *reliable* speaker, X_3 appears more *unreliable* than X_1 in the average, as, for what concerns *relative unreliability* scores, $\bar{\mu}_3 > \bar{\mu}_1$.

	impostor attempts against registered speakers $m = 3$					
	$Y_{\tilde{1}}$		$Y_{\tilde{2}}$		$Y_{\tilde{3}}$	
	Against Male				Against Female	
	$n = 2$		$n = 2$		$n = 2$	
	$Y_{1\,\tilde{1}}$	$Y_{2\,\tilde{1}}$	$Y_{1\,\tilde{2}}$	$Y_{2\,\tilde{2}}$	$Y_{1\,\tilde{3}}$	$Y_{2\,\tilde{3}}$
$\hat{V}$	$\boxed{2}$	$\boxed{0}$	$\boxed{5}$	$\boxed{1}$	-	$\boxed{1}$
$\hat{\hat{V}}$	4	2	1	2	-	4
$d_{j\tilde{i}}$	6	2	6	3	-	5
$d_{\tilde{i}}$	8		9		5	
d	22					
$n^*_{\tilde{i}}$	2		2		1	
$\tilde{m}^*$	3					
$(mn)^*$	5					
$\beta_{j\tilde{i}}$	$\frac{2}{6} = \frac{1}{3}$	0	$\frac{5}{6}$	$\frac{1}{3}$	undef.	$\frac{1}{5}$
$\beta_{\tilde{i}}$	$\frac{1}{2}\left[\frac{1}{3} + 0\right] = \frac{1}{6}$		$\frac{1}{2}\left[\frac{5}{6} + \frac{1}{3}\right] = \frac{7}{12}$		$\frac{1}{1}\left[\frac{1}{5}\right] = \frac{1}{5}$	
$\bar{\beta}$	$\frac{1}{5}\left[\frac{1}{3} + 0 + \frac{5}{6} + \frac{1}{3} + \frac{1}{5}\right] = \frac{17}{50} = 34\,\%$					
$\beta'_{\tilde{i}}$	$\frac{2+0}{8} = \frac{1}{4}$		$\frac{5+1}{9} = \frac{2}{3}$		$\frac{1}{5}$	
$\bar{\beta}'$	$\frac{1}{3}\left[\frac{1}{4} + \frac{2}{3} + \frac{1}{5}\right] = \frac{67}{180} \approx 37\,\%$					
β	$\frac{2+0+5+1+0+1}{22} = \frac{9}{22} \approx 41\,\%$					
$\tilde{m}^*_M \mid \tilde{m}^*_F$	2				1	
$(m_M n)^* \mid (m_F n)^*$	4				1	
$\bar{\beta}_M \mid \bar{\beta}_F$	$\frac{1}{4}\left[\frac{1}{3} + 0 + \frac{5}{6} + \frac{1}{3}\right] = \frac{3}{8}$				$\frac{1}{1}\left[\frac{1}{5}\right] = \frac{1}{5}$	
$\bar{\beta}_{MF}$	$\frac{1}{2}\left[\frac{3}{8} + \frac{1}{5}\right] = \frac{23}{80} \approx 29\,\%$					
$\bar{\beta}'_M \mid \bar{\beta}'_F$	$\frac{1}{2}\left[\frac{1}{4} + \frac{2}{3}\right] = \frac{11}{24}$				$\frac{1}{1}\left[\frac{1}{5}\right] = \frac{1}{5}$	
$\bar{\beta}'_{MF}$	$\frac{1}{2}\left[\frac{11}{24} + \frac{1}{5}\right] = \frac{79}{240} \approx 33\,\%$					
$\bar{\nu}_{\tilde{i}}$	$\frac{10}{51} \approx 0.19$		$\frac{35}{51} \approx 0.69$		$\frac{2}{17} = \frac{6}{51} \approx 0.12$	
$\bar{\nu}'_{\tilde{i}}$	$\frac{15}{67} \approx 0.22$		$\frac{40}{67} \approx 0.60$		$\frac{12}{67} \approx 0.18$	
$\nu_{\tilde{i}}$	$\frac{2}{9} \approx 0.22$		$\frac{2}{3} = \frac{6}{9} \approx 0.67$		$\frac{1}{9} \approx 0.11$	

Figure 12.4: One out of three impostor trials from Y_1 against X_1 were successful while none from Y_2 were. Hence $\beta_{\tilde{1}} = \frac{1}{6}$. But if the identities of impostors are not known, it can only be measured that, out of 8 impostor attempts against X_1 , 2 were successful, i.e $\beta'_{\tilde{1}} = \frac{1}{4}$. As no impostor attempt from Y_1 against X_3 was recorded, the *average false acceptance rate* against X_3 can only be averaged over 1 impostor. Hence $\beta_{\tilde{3}} = \beta'_{\tilde{3}} = \frac{1}{5}$. The 3 ways of computing false acceptance rates, namely the *average false acceptance rate* $\bar{\beta}$, the *average false acceptance rate assuming distinct impostors* $\bar{\beta}'$ and the *test set false acceptance rate* β provide significantly different scores, as the number of test utterances is not balanced across all possible couples (X_i, Y_j). In this example, the *relative vulnerability* scores $\bar{\nu}_{\tilde{i}}$, $\bar{\nu}'_{\tilde{i}}$ and $\nu_{\tilde{i}}$ indicate that speaker X_3 would appear as the most *resistant*, while speaker X_2 would seem to be the most *vulnerable*.

	impostor attempts from impostors $n = 2$					
	Y_1 $m = 3$			Y_2 $m = 3$		
	$Y_{1\,\tilde{1}}$	$Y_{1\,\tilde{2}}$	$Y_{1\,\tilde{3}}$	$Y_{2\,\tilde{1}}$	$Y_{2\,\tilde{2}}$	$Y_{2\,\tilde{3}}$
$\hat{V}$	$\boxed{2}$	$\boxed{5}$	-	$\boxed{0}$	$\boxed{1}$	$\boxed{1}$
$\hat{\hat{V}}$	4	1	-	2	2	4
$d_{j\tilde{i}}$	6	6	-	2	3	5
$\tilde{d}_j$	12			10		
d	22					
$\tilde{m}^*_j$	2			3		
n^*	2					
$(mn)^*$	5					
$\beta_{j\tilde{i}}$	$\frac{2}{6} = \frac{1}{3}$	$\frac{5}{6}$	undef.	0	$\frac{1}{3}$	$\frac{1}{5}$
$\tilde{\beta}_j$	$\frac{1}{2}\left[\frac{1}{3} + \frac{5}{6}\right] = \frac{7}{12}$			$\frac{1}{3}\left[0 + \frac{1}{3} + \frac{1}{5}\right] = \frac{8}{45}$		
$\tilde{\beta}'_j$	$\frac{2+5}{12} = \frac{7}{12}$			$\frac{0+1+1}{10} = \frac{1}{5}$		
$\bar{\tilde{\beta}}$	$\frac{1}{5}\left[\frac{1}{3} + \frac{5}{6} + 0 + \frac{1}{3} + \frac{1}{5}\right] \equiv \bar{\beta} = \frac{17}{50} = 34\,\%$					
$\bar{\tilde{\beta}}'$	$\frac{1}{2}\left[\frac{7}{12} + \frac{1}{5}\right] = \frac{47}{120} \approx 39\,\%$					
$\tilde{\beta}$	$\frac{2+5+0+0+1+1}{22} \equiv \beta = \frac{9}{22} \approx 41\,\%$					
$\bar{\tilde{\nu}}_j$	$\frac{35}{51} \approx 0.69$			$\frac{16}{51} \approx 0.31$		
$\bar{\tilde{\nu}}'_j$	$\frac{35}{47} \approx 0.74$			$\frac{12}{47} \approx 0.26$		
$\tilde{\nu}_j$	$\frac{7}{9} \approx 0.78$			$\frac{2}{9} \approx 0.22$		

Figure 12.5: Out of 6 trials from impostor Y_1 against speaker X_1, 2 of them turned out to be successful, while out of 6 other trials against X_2, 5 lead to a (false) acceptance. As no attempts from Y_1 against X_3 were recorded, the *average imposture rate* from impostor Y_1 can be estimated as $\tilde{\beta}_1 = \frac{1}{2}\left[\frac{2}{6} + \frac{5}{6}\right] = \frac{7}{12}$. If we now ignore the actual identities of violated speakers, and we summarise globally the success of impostor Y_1, we get $\tilde{\beta}'_1 = \frac{2+5}{12}$ which turns out to be also equal to $\frac{7}{12}$. While $\bar{\tilde{\beta}} = \bar{\beta}$ and $\tilde{\beta} = \beta$, the *average imposture rate regardless of the claimed identity* $\bar{\tilde{\beta}}'$ indicates that the "average" impostor is successful almost 2 times out of 5 in his attempts. All estimates of the *relative imitation ability* ($\bar{\tilde{\nu}}_j$, $\bar{\tilde{\nu}}'_j$ and $\tilde{\nu}_j$) agree that Y_1 is a much more *skilled* impostor than Y_2 who seems to be quite *poor*.

Chapter 13

Synthesis assessment

13.1 Introduction

13.1.1 What are speech output systems?

By a speech output system we mean some artifact, whether a dedicated machine or a computer programme, that produces signals that are intended to be functionally equivalent to speech produced by humans. In the present state of affairs speech output systems generally produce audio signals only, but laboratory systems are being developed that supplement the audio signal with the visual image of the (artificial) talker's face (cf. Benoît (1991), Benoît et al. (1992)).[1] Audio-visual (or: bi-modal) speech output is more intelligible than audio-only output, especially when the audio channel is of degraded quality. In the body of this chapter we will not be concerned with bi- or multi-modal speech output systems, and concentrate on audio-only output instead. However, comments on the assessment of the visual component of bi-modal speech output have been provided by Benoît for the present chapter, which will be included in Appendix 2.

We exclude from the domain of speech output systems such devices as tape recorders and other, more advanced, systems that output speech on the basis of complete, pre-stored messages ("canned speech" or "copy synthesis"), irrespective of the type of coding or information compression used to save storage space. We crucially limit our definition to systems that allow the generation of novel messages, either from scratch (i.e. entirely by rule) or by recombining shorter pre-stored units. This definition also includes hybrid synthesis systems where individually stored words (e.g. digits) are substituted in information slots in a carrier sentence (e.g. in time table consultation services).

It seems to us that two basic types of speech output systems have to be distinguished on the basis of their input, namely text-to-speech (TTS) and concept-to-speech (CTS). Other, more complex, systems combine characteristics of these two.

- *Text-to-speech.* The majority of speech output systems is driven by text input. These systems convert text printed in normal orthography (generally stored in a computer memory as ASCII codes) to speech. Conventional spelling provides a reasonable indication of what sounds and words have to be output, but typically underrepresents prosodic properties of the message, such as the positions of accents, melody, and temporal organisation. The prosody of an utterance reflects the communicative intentions of the writer of the input text, which by present-day standards are virtually impossible to fully reconstruct from the text (cf. the title of a much-cited article: 'Accent is predictable, if you're a mind reader'; Bolinger (1972)). The

[1] Moreover, systems have been developed that generate audio output together with visual output (text printed on computer screen, for foreign language learning purposes) and/or tactile output (braille, specifically for the visually handicapped). In such cases the term multi-modal output is often used.

reconstruction of the writer's intentions is an implicit part of the so-called linguistic interface, i.e. the first part of most advanced text-to-speech systems. All errors in the linguistic interface may detract from the quality of the output speech, and are therefore a legitimate object of evaluation.

- *Concept-to-speech.* In other types of speech output systems, especially dialogue systems, the communicative intentions are fully specified at the input stage: the system itself determines what message it wants to get across. It may still be the case, of course, that the dialogue system has misconstrued a user's request, and consequently issues an inappropriate response message, but this should not be considered an error on the part of the output system.

- *Interpreting (or translating) telephony* (SL-TRANS, cf. Morimoto et al. (1990); JANUS, cf. Waibel et al. (1991)) and *face-to-face spoken dialogue translation* (Verbmobil, cf. Wahlster (1993)) combine characteristics of both TTS and CTS. Interpreting telephony, for instance, a spoken utterance in one language (e.g. Japanese) is decomposed into its linguistic message and its speaker specific properties (e.g. voice characteristics, speed, pitch range). The linguistic message is converted to text, and transmitted. At the receiver end the text is automatically translated into another language (e.g. German) and then converted back to speech in the target language setting the synthesiser's speaker specific parameters such that the personal characteristics of the source speaker are approximated in the output signal. Crucially, the sender's intentions do not have to be inferred from the textual representation of the message; the intended focus distribution can be reconstructed directly from the properties of the source language speech signal.

13.1.2 Why speech output assessment?

In spite of the rapid progress that is being made in the field of speech technology, any speech output system available today can still be spotted for what it is: non-human, a machine. Most older systems will fall through immediately due to their robot-like melody and garbled vowels and consonants. Other, more recently developed synthesis methods using short-segment waveform concatenation techniques such as PSOLA (Moulines & Charpentier (1990)) yield segmental quality that is very close to human speech (Portele et al. (1994)), but still suffer from noticeable defects in matters of melody and timing.

As long as synthetic speech is inferior to human speech, speech output assessment will be a major concern. Speech technology development today is typically evaluation-driven. Large scale speech technology programmes have been launched both in the United States and in Europe (for overviews see O'Malley & Caisse (1987); Van Bezooijen & Pols (1989); Pols (1991)). Especially in the European Union, with its many official languages, a strong need was felt for output quality assessment methods and standards that can be applied across languages. With this goal in mind the multi-national EU-ESPRIT SAM-project was set up (Fourcin et al. (1989)). Later, the EU Expert Advisory Group on Language Engineering Standards (EAGLES) programme started, and included a working group on speech output assessment.

Speech output assessment may be of crucial importance to two interested parties, the systems designers on the one hand, and the prospective buyers and end users (possibly represented by consumer organisations) of the system on the other.

- *Designers* are intent on improving their speech output systems. However, designers who have grown up with their system are used to all its habits; they are likely to understand its output much better than first-time users, and will overrate its performance level. Less subjective quality assessment techniques are needed in order to determine how well a system performs relative to a benchmark test, or how favourably it compares with a previous edition of the system or with other designers' systems (*comparative testing* or *performance evaluation*). To the extent that a system performs less than perfect, the designer will have to learn which aspect(s) and/or components of the system are flawed. Designers will therefore also be interested in *diagnostic evaluation*, either by doing detailed error analyses on the test results, or by running component-specific tests.

- The needs of *systems users* (end users and/or systems providers) are different than those of designers but they, too, heavily rely on assessment techniques. Prospective buyers will always have a specific use of their speech output system in mind. Understandably, they will want the simplest, and therefore cheapest, system that satisfies their needs. The buyer (or his consumer organisation) will therefore need an absolute yardstick in order to determine beforehand if the output speech is good enough to get the message across in the given application.

13.1.3 Users of this chapter

This chapter serves the potential needs of several disparate groups of readers. Since not all parts of this chapter will be equally relevant to every reader group, we will identify reader groups, and point out which parts of this chapter, after sections 1–3, have particular relevance for each group. The reader groups, of course, overlap to a large degree with the parties interested in speech output evaluation discussed above, but a more refined classification seems to be in order. We will distinguish the following groups of readers:

- *Developers of speech output tests*
  Speech output assessment is an expanding field. New tests come available in rapid succession, so that test developers want to keep abreast with what is new. Test developers will want to know the advantages and disadvantages of tests proposed in the literature, and need to know what requirements will be made to the next generation of tests. This chapter discusses many alternative tests and testing methodologies, and makes recommendations as to what type of tests are more suited for a specific purpose. The chapter will also indicate what direction speech output testing should take in order to meet the testing requirements of the next generation of speech output systems (section 13.6).

  Readers in this group will be interested in section 13.5 on glass box (diagnostic) testing as well as in section 13.4 on black box testing. When systems are in their early developmental stages, glass box testing will be most relevant; when systems have sufficiently matured, black box tests are in order. Field testing (section 13.4.2) will generally be deferred until the end user groups and their specific applications are known. Field testing will often be conducted by, or at least in close cooperation with, systems procurers for end users (see below).

  The chapter will be especially useful to those readers who are not test developers at this time but aspire at becoming test developers in the near future. These readers include, of course, students at the Ph.D.-level who want to make a career in speech output testing. A second group of readers who are new to the field are developers of speech technology in Eastern Europe and in certain third world countries where computer technology is now widely available at affordable prices, generating an immediate need for the development of speech output systems and tests in the languages of the areas concerned.

- *Developers of speech output systems*
  A lot of speech technology research and development takes place in small high-tech companies. The research staffs are often too small to warrant the appointment of a full-time test evaluation expert, so that a lot of diagnostic do-it-yourself testing is going on. Developers who are not evaluation experts themselves will find this chapter a useful source of information. It identifies standard tests and test suites that are readily available for a range of diagnostic purposes (with addresses, contact persons, and literature references listed in appendices at the end of this chapter). The remarks made above with respect to newcomers to the evaluation field

apply here as well. This type of reader should concentrate on the glass box approach (section 13.5).

- *Procurers of speech output systems*
  At the most user-oriented end of the spectrum, procurers of systems will find our chapter of interest. Procurers, who themselves will more often than not be naive to the field of speech output technology, will not normally be interested in diagnostic testing. They will be looking for a single figure of merit on the basis of which to decide the system that is best for a given range of applications. This reader group is most difficult to deal with since their needs are most divergent. There are no off-the-shelf tests that satisfy their needs. Rather we will provide numerous examples which may serve as guidelines how to go about field testing speech output systems for specific applications. This type of reader should concentrate on those parts of this chapter dealing with black box output testing and field tests (section 13.4).

Given this division of reader groups we will present two types of recommendation, if and when we can.

- The first type suggests what decisions can be made in the present situation with what is available today, or can be made available with little effort in the immediate future. These recommendations will be found throughout this chapter, in concise format and numbered.
- The second type of recommendation that we will be making, outlines possible courses of test development for the mid and long term. Such recommendations, predominantly aimed at the evaluation experts, will be presented in section 13.6 at the end of this chapter, in less explicit format; they will not be numbered.

13.2 Towards a taxonomy of assessment tasks and techniques

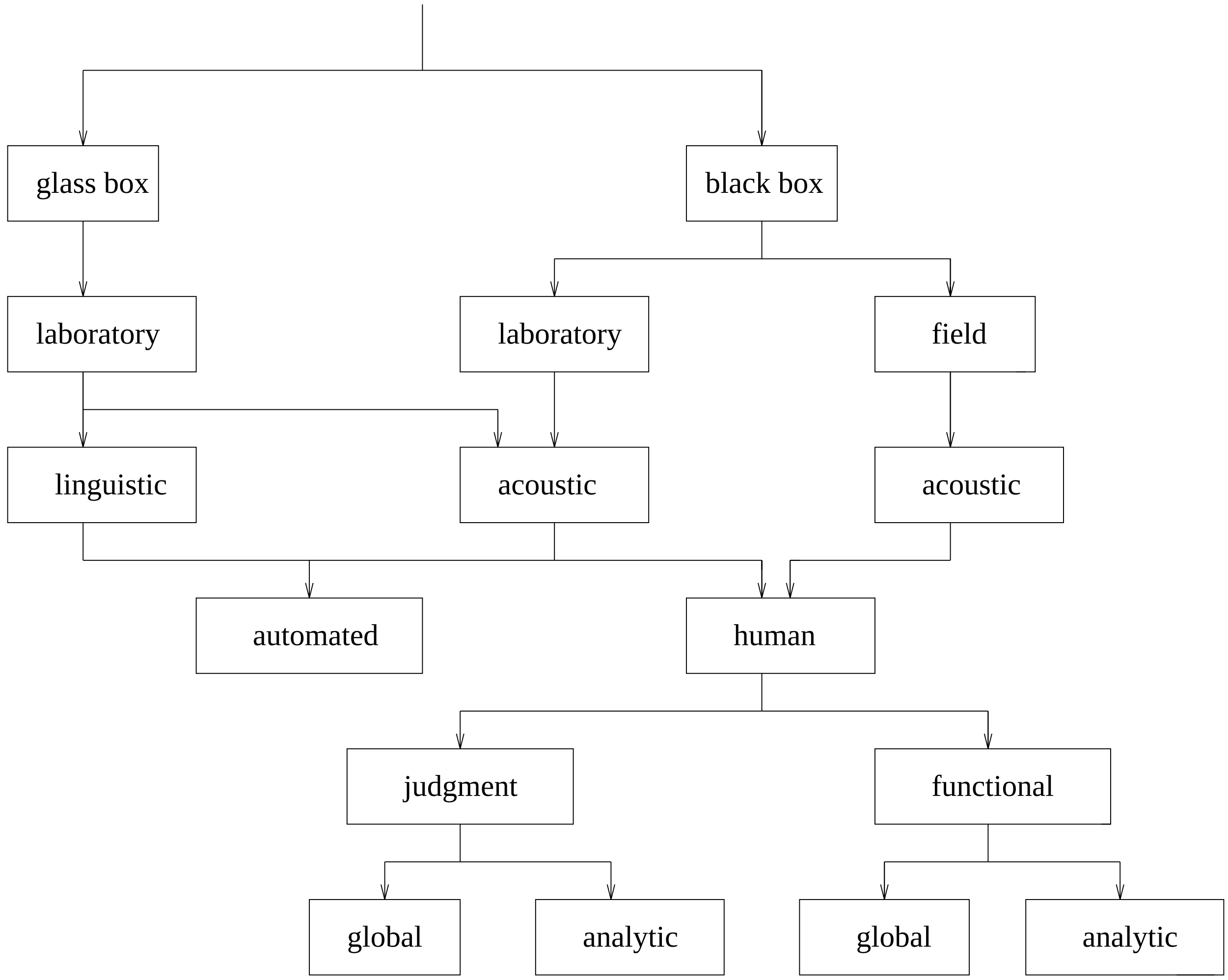
To structure our overview of TTS assessment tests we will discuss a number of useful distinguishing parameters, which partly overlap with earlier attempted taxonomies (see e.g. Van Bezooijen & Pols (1990); Pols (1991); Jekosch & Pols (1994); Goldstein (1995)) and explain the relationships among them, before dealing with any specific assessment techniques. Figure 13.1 illustrates the relationships among the various dichotomies that make up our taxonomy. It will be apparent from the figure that the dichotomies are non-orthogonal. For instance, tests that have been developed to evaluate specific modules within a speech output system (glass box approach), will only be used in a laboratory setting. The subdivision of human listener tasks will only be relevant when tests involve human listeners; therefore no task variables will be specified under automated test techniques. In the subsequent paragraphs we will outline and discuss the taxonomy of speech output assessment techniques following as much as possible the structure of Figure 13.1. Note that the dichotomies used are intended as contrastive dimensions so as to create a multi-dimensional space within which speech output tests can be located. The terms involved in any dichotomy should not be construed as labels identifying mutually exclusive approaches to speech output evaluation.

The levels in Figure 13.1 are dealt with in the following sections:

13.2.1 Glass box $\leftrightarrow$ black box

Text-to-speech systems generally comprise a range of modules that take care of specific tasks. The first module (or complex of modules) converts an orthographic input string to some abstract linguistic code that is explicit in its representation of sounds and prosodic markers. Various modules then act upon this symbolic representation. Typically, one module concatenates the primitive building blocks (phonemes, diphones) in their appropriate order, another implements what coarticulation

Figure 13.1: *Relationships among dimensions involved in a taxonomy of speech output evaluation methods. Any path from the root down to any terminal that does not cross a horizontal gap, constitutes a meaningful combination of test attributes.*



1. glass box vs. black box: 13.2.1

2. laboratory vs. field: 13.2.2

3. linguistic vs. acoustic: 13.2.3

4. automated vs. human: 13.2.4

5. judgment vs. functional: 13.2.5

6. global vs. analytic: 13.2.6

is needed to obtain smooth human-like transitions between successive building blocks. Prosodic modules, taking the positions of word stresses, sentence accents, phrasal and sentence boundaries into account, are then called upon in order to provide an appropriate temporal organisation (local accelerations and decelerations, pauses) and speech melody.

End users will typically be interested in the performance of a system as a whole. They will consider the system as a *black box* that accepts text and outputs speech, a monolith without any internal structure. For them it is only the quality of the output speech that matters. In this way systems developed by different manufacturers can be compared or the improvement of one system relative to an earlier edition can be traced over time (*comparative testing*). However, if the output is less than optimal it will not be possible to pinpoint the exact module or modules that caused the problem. For *diagnostic* purposes, therefore, designers often set up their evaluations in a more experimental ("*glass box*") way. Keeping the effects of all modules but one constant, and systematically varying the characteristics of the free module, any difference in the assessment of the system's output must be caused by the variations in the target module. Glass box testing, of course, presupposes that the researcher has control over the input and output of each individual module.

Recommendation 1
Use a glass box approach if you want diagnostics in order to improve your speech output system.

Recommendation 2
Use a black box approach if you want to assess the overall performance of speech output systems.

The dichotomy between glass box and black box testing is basic to speech output testing, which has led some researchers to propose a strict terminological division whereby "evaluation" signifies glass box testing (or: *diagnostic evaluation*) only, and "assessment" is reserved exclusively for black box testing (or: *performance evaluation*). In this chapter we will use the terms "evaluation" and "assessment" indiscriminately, and use disambiguating adjectives whenever there is a risk of confusion.

13.2.2 Laboratory ↔ field

Ideally, any speech output system should perform at the same level of adequacy as a human speaker. Such a system would be optimal for any application. However, given that systems available today are less than optimal, it is important to know which aspects of a system's performance are essential to a specific application. Speech output systems typically form an element of a larger human-machine interface in an application with a specific, dedicated task. In practice this means that, quite probably, the vocabulary and types of information exchanges are restricted and domain-specific, so that situational redundancy is likely to make up for poor intelligibility. On the other hand, speech output systems will often be used in complex information processing tasks, so that the listener has only limited resources available for attending to the speech input. Also, end users may have different attitudes towards, and motivations for, working with artificial speech than subjects in laboratory experiments, especially when the latter have not been explicitly selected so as to be fully representative of the end users. It is often hazardous, therefore, to predict beforehand, on the basis of *laboratory tests*, how successful a speech output system will be in the practical application. Generally, as an application situation contains more specific aspects, less prediction of field performance is afforded by laboratory tests. Output systems will have to be tested in the field, i.e. in the real situation, with the real users. The use of *field tests* will be limited to one system in one specific application; results of a field test cannot, as a rule, be generalised to other systems and/or other applications.

Recommendation 3
Do not rely on laboratory tests. As soon as there is a discrepancy between the laboratory setting and the true field situation (in terms of environment, tasks, type of listener) field testing is necessary.

13.2.3 Linguistic $\leftrightarrow$ acoustic

The more complex TTS systems can roughly be divided into a linguistic interface that transforms spelling into an abstract phonological code, and an acoustical interface that transduces this symbolic representation to an audible waveform. The quality of the intermediary representation can be tested directly at the *symbolic-linguistic* level or indirectly at the level of the *acoustic* output. Testing the audio output has the advantage that only errors in the symbolic representation that have audible consequences, will affect the evaluation. The disadvantage of audio testing is that it involves the use of human listeners, and is therefore costly and time-consuming. Moreover, the results of acoustic testing are unspecific in that the designer is not informed whether the problems originate at the linguistic or at the acoustic level. As an alternative the intermediate representations in the linguistic interface are often evaluated at the symbolic level. It is, of course, a relatively easy task to compare the symbolic output of a linguistic module with some pre-stored key or model representation and determine the discrepancies, and this is what is normally done. The non-trivial problem is where to obtain the model representations. These will generally have to be compiled manually (or semi-automatically at best), and often involve multiple correct solutions.

13.2.4 Human $\leftrightarrow$ automated

In the large majority of test procedures *human subjects* are called upon in order to determine the quality of a speech output system. This should come as no surprise to us, since the end user of a speech output system is a human listener. However, there are certain drawbacks inherent to the use of human subjects. Firstly, humans, whether acting as single individuals or collectively as a group, are always somewhat noisy in their judgments or task performance, i.e. the results of tests involving human responses are never perfectly reliable in the statistical, psychometric sense of the word. Another drawback of tests involving human subjects it that they are time-consuming and therefore expensive to run.

Recent developments, which are still very much in the laboratory stage, seek to replace human evaluation by *automatic assessment* of speech output systems or modules thereof. Attempts can be (and in fact have been) made to automatically measure the discrepancy in acoustical terms between a system's output and the speech of the human speaker that serves as the model the system is intended to imitate. This is the type of evaluation technique that one would ultimately want to come up with: the use of human listeners is avoided, so that perfectly reproducible noiseless results can be obtained in as little time as it takes a computer to execute the program. At the same time, however, it will be clear that implementation of such techniques as a substitute for human listeners presupposes that we know exactly how human listeners evaluate differences between two realisations of the same linguistic message. Unfortunately, this type of knowledge is largely lacking at the moment; filling the gap would be a research priority. Nevertheless, preliminary automatic comparisons of synthetic and human speech output have been undertaken in the fields of melody and pause distribution (Barry et al. (1989)), long term average spectral characteristics (Pavlovic, Rossi & Espesser (1991)) and dynamics of speech in the frequency and time domains (Houtgast & Verhave (1991), Houtgast & Verhave (1992)). Generally, the results obtained through these techniques show sufficient promise to warrant extension of their scope [2]. We will come back to the possibilities of automated testing in section 13.6.

[2] This intended use of objective quality measurement bears a superficial resemblance to the work by Kryter (1962a), Kryter (1962b)) and associates on the development of the so-called articulation index (AI). The AI accurately predicts the loss in intelligibility of speech due to characteristics of the transmission channel by automatically comparing the long-term average spectral characteristics of the input and output speech. However, no AI-based approach can be used to discriminate between the desirable (speech-like) and the undesirable (noise, distortion) features of the input speech.

13.2.5 Judgment ↔ functional testing

By *judgment testing* (also called *opinion testing* in telecommunication research) we mean a procedure whereby a group of listeners is asked to judge the performance of a speech output system along a number of rating scales. The scales are typically bi-polar adjectives that allow the listeners to express the quality of the output system along a more global or more specific aspect of its performance. Although the construction of an appropriate scaling instrument is by no means a trivial task, a scaling test can be administered with little effort and yields a lot of potentially useful information.

At the other extreme the speech output can be assessed in terms of how well it actually performs its communicative purpose. This is called *functional testing*. For instance, if we want to know to what extent the output speech is intelligible, we may prefer to measure its intelligibility not by asking listeners how intelligible they think the speech is, but by determining, for instance, whether listeners correctly identify the sounds. Consider, as an example on a higher level of communication, the assessment of an information system using speech output. We may ask users to judge the output quality, but we may also functionally determine the system's adequacy by looking at task completion: how often and how efficiently do the users get the information from the system that they need?

One would hope that the results of judgment and functional assessments converge. Obviously, one would like to use the results of functional assessments in order to gauge the validity of judgments, rather than the other way about. As far as we have been able to ascertain, there has been little research into this matter. Yet, there is at least one set of intersubjective and functional data that was collected for the same group of listeners and stimuli, testing two different text-to-speech systems at three different points in time, from which it appeared that the scaling results were highly correlated with the corresponding functional test scores (Pavlovic, Rossi & Espesser (1990)).

13.2.6 Global ↔ analytic assessment

In a sense there is only one ultimate criterion that determines the quality of a speech output system, viz. its overall quality within a given application. Judgment tests usually include one or more rating scales covering such *global* aspects as "overall quality", "naturalness" and "acceptability". A functional approach to global assessment would be to determine whether users of speech output, when given the choice, choose to work with a machine or with the human original the machine is intended to simulate. Or one may determine if the information exchange is as successful in machine-to-human as it is in human-to-human situations.

On the other hand, one may be interested in determining the quality of specific aspects of a speech output system, in an *analytic* listening mode, where listeners are requested to pay particular attention to selected aspects of the speech output. Again, both judgment and functional tests can and have been designed addressing the quality of specific aspects of a speech output system. Listeners may be asked, for instance, to rate the clarity of vowels and consonants, the appropriateness of stresses and accents, pleasantness of voice quality, and tempo. Functional tests have been designed to test the intelligibility of individual sounds (phoneme monitoring), of combinations of sounds (syllable monitoring), of whole words (word monitoring) in isolation as well as in various types of context (e.g. Nusbaum, Greenspan & Pisoni (1986)); Ralston et al. (1991)).

13.3 Methodology

The great majority of speech output assessment techniques use listening experiments involving human subjects, i.e. functional and/or judgment tests of speech output at the acoustic level. In the following subsections we will discuss a number of methodological issues that are relevant especially to this type of testing. The issues concern the choice of subjects, test procedures, benchmarks and reference conditions, and precautions to ensure cross-language comparability. Although there is no a priori reason why this should be so, no accepted methodology seems to exist for other types of speech output evaluation techniques. As will be obvious from later sections, for example, no accepted methodology can be identified in the field of output evaluation at the symbolic linguistic level. It is unclear in this area what kinds of textual materials should be used in tests, what error categories should be distinguished, and what scoring procedures should be used. We will therefore limit the methodological discussion to acoustic output testing techniques involving human listeners.

13.3.1 Subjects

One of the most important aspects of a measuring instrument is its reliability. How reliable, for example, is subjects' performance in functional intelligibility tests when tested several times? Test/retest *intra*subjective reliability of intelligibility was assessed by Logan, Greene & Pisoni (1989)) and Van Bezooijen (1988)); in both cases it was found to be good. More attention has been paid to *subject dimensions* systematically affecting *inter*subjective reliability. This research was motivated by the finding of large variance in the test scores, possibly obscuring effects of the synthesis systems compared. Most studies in this area examined variability in intelligibility scores. Subject dimensions considered relevant include: age, non-expert experience with synthetic speech, expert experience with synthetic speech, and analytic listening.

Within the ESPRIT-SAM project (Howard-Jones (1992a), Howard-Jones (1992b)), the effect of *age* was examined with Italian VCV-items. Five age categories were distinguished (10–19, 20–29, 30–44, 45–59, over 60), with between 5 and 8 subjects per group. The group scores of percentages correct consonant identification ranged from 58%, for the oldest group, to 64%, for the youngest group. So, little evidence was found for an effect of the subject dimension age.

Non-expert experience with synthetic speech was investigated in several studies. Howard-Jones (1992a), Howard-Jones (1992b) compared the performance of 8 subjects experienced with synthetic speech and 24 inexperienced subjects. German VCV-items were presented. The mean score for the experienced subjects was 79%, that for the inexperienced subjects 62%. There is further evidence that the intelligibility of synthetic speech increases as a result of non-expert experience with synthetic speech, both when acquired in the form of training with feedback (e.g. Greenspan, Nusbaum & Pisoni (1985); Schwab, Nusbaum & Greene (1985)) and when acquired in a more natural way without feedback (Pisoni, Greene & Nusbaum (1985), Pisoni, Nusbaum & Greene (1985); Boogaart & Silverman (1992)). The learning effect has been found to manifest itself after only a few minutes of exposure. However, there are indications that the effect of learning depends on the type of synthesis used. Jongenburger & Van Bezooijen (1992)) assessed the intelligibility of two synthesis systems used by visually handicapped for reading a digital daily newspaper in a first confrontation and after one month of experience. An open response CVC identification test was used. For one system, which was allophone based, consonant intelligibility increased from 58% to 79%; for the other system, which was diphone based, intelligibility increased from 63% to 68%. It was hypothesised that the characteristics of allophone-based synthesis are easier to learn because they are rule-governed and therefore more invariant than those of diphone-based synthesis. Moreover, no transfer was found from experience with one type of synthesis to the understanding of the other type of synthesis. This suggests that there is no such thing as general experience in listening to synthetic speech.

The subject dimension *expert experience with synthetic speech* was examined by Howard-Jones (1992a) with English VCV-items. A percentage correct consonant identification of 30% was obtained for the inexperienced subjects versus 49% for the experts. So, again improved performance was found as a function of increased exposure.

The last subject dimension we want to mention is *experience in listening analytically to speech*. On the basis of a reanalysis of the results from a number of their evaluation studies, Van Bezooijen & Pols (1993)) conclude that the more ear-training subjects have, the higher the percentages correct they attain. Furthermore, ear-training was found to result in a reduction of intersubjective differences.

Apart from variance which can be attributed to particular subject dimensions, much apparently *individual variability* is found in test scores. Hazan & Shi (1993) examined the variance in subject scores in various tests, including intelligibility of meaningless VCV-items, intelligibility of Semantically Unpredictable Sentences (SUS, see Appendix 1 G), and speech pattern identification for plosive place and voicing contrasts. A homogeneous group of subjects was used:

- All had normal hearing (average pure tone thresholds of 20 dB HL or better from 0.25 to 8 kHz in both ears.
- All were students at University College London.
- All fell within the age range 18 to 32 years.
- All were native speakers of English.
- None had any training in phonetics.
- All had little or no previous experience with synthetic speech.

Despite the homogeneity of the subject group, a sizeable degree of variability was found in all tests. For the SUS the range (i.e the difference between the best and worst performing subject) was 28%, for the CVC-test the range was 47%. At the level of speech pattern processing, considerable differences were found in the perceptual weighting given to individual cues to plosive place and voicing contrasts. Hazan & Shi attribute the variability not to audiological differences among listeners, but to the development of different perceptual strategies during language acquisition. They distinguish two types of listeners: "auditors" (i.e. users of acoustic information) and "comprehenders" (i.e. users of global contextual information).

Having established that there is much variability in the scores obtained in speech output evaluation tests, part of which can be attributed to clearly identifiable subject dimensions such as previous experience with synthetic speech, one may wonder what implications this has for the selection of subjects in specific tests. We think that the implications for subject selection depend in part on the type of test administered: black box versus glass box, comparative versus diagnostic, application for the general public versus application for a specific group of users, etc. However, some general recommendations can be formulated as well.

Recommendation 4
Always exclude hearing-impaired subjects from speech output assessment. Within the SAM-project (Howard-Jones (1992a), Howard-Jones (1992b)) it is specified that subjects should pass the hearing screening test at 20 dB HL at all octave frequencies from 500 to 4000 Hz.

Recommendation 5
Never use the same subject more than once, unless, of course, one is interested in the effect of repeated exposure.

Recommendation 6
In diagnostic testing only include subjects speaking the same language (variety) as the language (variety) tested.

Recommendation 7
In diagnostic testing, hire a trained phonetician (with a basic understanding of the relationships between articulation and acoustics) in the initial stages of development of a system in order to obtain subtle information (e.g. degree of voicing in plosives), or information that is usually not used for functional purposes in real-life communication (e.g. formal aspects of temporal organisation and intonation, cf. Terken (1993)).

The following approach is recommended not only because of (possible) differences in the perception of the speech output, but also because motivation is known to play an important role in the effort people are willing to spend in order to understand suboptimal speech. If people have a choice between human and synthetic speech, the synthetic speech will have to be good if it wants to have a chance of being accepted. However, if people do not have a choice, e.g. the visually handicapped who without synthesis (or braille) will not have access to a daily newspaper, synthesis will be accepted more easily.

Recommendation 8
In specialised applications, select subjects who are representative of the (prospective) users. For example, synthesis integrated in a reading machine for the blind should be tested with visually handicapped. And synthesis for long-term use should be tested with subjects with different degrees of experience and familiarisation with the type of synthetic speech of interest.

Recommendation 9
Synthesis to be used by the general public for incidental purposes, i.e. which should be functionally adequate in a first confrontation, should be tested with a wide variety of subjects, including people with a limited command of the language, dialect speakers, and people of different ages. However, none of them should have any experience in listening to synthetic speech. In telecommunications research, groups of between 12 and 16 subjects (all with English as their primary language) have been found sufficient to obtain stable mean values in judgment tests.

13.3.2 Test procedures

As indicated in section 13.2, speech output assessment techniques can be differentiated along a number of parameters. No parameters related to the actual test procedure were included there. Test procedures can vary with respect to subjects (see section 13.3.1), stimuli, and response modality.

Stimuli can vary along a large number of parameters, the most important of which are listed below.

- *Length and complexity* (e.g. at the word level: monosyllabic, disyllabic, polysyllabic, including only single consonants and vowels or also sequences of consonants and vowels). The more varied in length and complexity the test items are, the more diagnostic information can be obtained and the more representative the test results are for the perception of unrestricted speech output. However, higher linguistic levels are often less suited for diagnostic purposes because subjects' responses are determined by many other sources of information in addition to the acoustic properties of the stimuli (see section 13.4.1.1).

- *Linguistic level* (word, sentence, paragraph). Again, the higher the linguistic level, the better test results can be generalised to unrestricted speech output.

- *Stimulus set* (fixed set, where all items are presented each time the test is run versus open set, where each time new (combinations of) test items are presented, e.g. the SUS-test in Appendix 1 G). Of course, in the light of learning effects open sets are more useful and flexible than fixed sets.

- *Meaningfulness*, either at the word level or at the sentence level (meaningful, meaningless, or mixed, i.e. lexically or semantically unpredictable). Each choice seems to have both advantages and disadvantages. For example, tests which only use meaningful test items at the word level, such as the DRT and MRT (see Appendix 1 D and E) have the advantage of being reliable and easy to administer. However, intelligibility may be overestimated, there is a risk of a ceiling effect, and they have little diagnostic value. In principle, the mixed approach seems a good choice, because the subjects are not guided in any way as to what constitutes a legal or an illegal response. Nevertheless, there may be a risk of a bias towards meaningful words. For other implications of the choice between meaningful, meaningless, and mixed items at the word level, see section 13.5.2.1.2.1. For implications at the sentence level, see section 13.5.2.1.2.2.

- *Representativeness*, e.g. Phonetically Balanced (PB) stimulus lists, with a frequency of occurrence of phonemes in accordance with the phoneme distribution in the language tested or the specific domain of application at hand, or equal representation of each phoneme. If one wants to obtain a global idea of the intelligibility of a system, PB-lists are to be preferred, if one aims at diagnostic information, one usually opts for equal representation.

In Appendix 1, summary descriptions of tests are given where the stimuli have been categorised along these stimulus parameters.

Response modality can vary along a number of parameters as well. The choice seems to be mainly determined by three factors: comparative versus diagnostic, functional versus judgment, and TTS-development versus psycholinguistic interest. In the five types of response modalities listed below, 1 and 2 are mainly used within the glass box approach (1 in TTS-development, 2 in psycholinguistically oriented research), whereas 3, 4 and 5 are more common in the black box approach. The latter three response modalities can be further differentiated in that 3 and 4 are functional in nature (3 in TTS-development, 4 in psycholinguistically oriented research), whereas 5 represents judgment testing. In the list of response modalities a distinction is made between *off-line* tests, where subjects are given some time to reflect before responding, and *on-line* tests, where an immediate response is expected from the subjects, tapping the perception process before it is finished.

1. *Off-line identification tests*, where subjects are asked to transcribe the separate elements (sounds, words) making up the test items. This response modality can be further differentiated. With respect to the *nature of the set of response categories* there is a choice between:

   - a *closed set*, where subjects are forced to select the appropriate response from a limited number of pregiven categories, and
   - an *open response mode*, where the only restriction are the constraints imposed by the language.

   Transcription can be:

   - in *normal spelling,* leading to problems in the interpretation of the responses in case of meaningless or lexically unpredictable stimuli (e.g. if subjects write down “lead”, have they heard /led/ or /li:d/?), or
   - *unambiguous notation*, placing the burden upon the subjects, since they have to be trained to systematically apply this notation system.

2. *On-line identification tests*, requiring the subject to decide whether the stimulus does or does not exist as a word in the language (so-called lexical decision task, e.g. Pisoni, Greene & Nusbaum (1985), Pisoni, Nusbaum & Greene (1985)).

3. *Off-line comprehension tests*, in which content questions have to be answered in an open or closed response mode (e.g. Pisoni, Greene & Nusbaum (1985), Pisoni, Nusbaum & Greene (1985).

4. *On-line comprehension tests*, requiring the subject to indicate whether a statement is true or not (so-called sentence verification task, e.g. Manous et al. (1985)).

5. *Judgment tests* (also called opinion tests), involving the rating of scales (e.g. Pavlovic, Rossi & Espesser (1990); Delogu et al. (1991); ITU-T (1993)).

The last response modality will be discussed in some more detail. Pavlovic and co-workers have conducted an extensive series of studies (cf. Pavlovic, Rossi & Espesser (1990)) comparing different types of scaling methods that can be used in judgment tests to evaluate speech output. Much attention was paid to:

- the *magnitude estimation method*, where the subject is presented with an auditory stimulus and is asked to express the perceived strength/quality of the relevant attribute (e.g. intelligibility) numerically ("type in a value") or graphically ("draw a line on the computer screen"), and

- the *categorical estimation method*, where the subject has to select a value from a limited range of prespecified values, e.g. 1 representing extremely poor and 10 excellent intelligibility.

Pavlovic et al. stress that there are important differences between the two types of scaling methods, for example the fact that categorical estimation results in an interval scale, whereas magnitude estimation results in a ratio-scale. The former leads to the use of raw ratings, the calculation of the arithmetic mean, and the comparison of conditions in terms of differences, the latter leads to the use of the logarithm of the ratings, the geometric mean, and comparison in terms of ratios. The differences also have implications for the type of conclusions to be drawn from the test results. Both the categorical estimation method (with a 20-point scale) and the magnitude estimation method have been included in SOAP as standard SAM Overall Quality test procedures (see K in Appendix 1).

Recommendation 10
Use categorical estimation in rapid judgment tests with a view of internal comparison, and when you do, use at least a 10-point scale.

Recommendation 11
Use magnitude estimation for test external comparison, and when you do, use line length with imaginary ideal speech as a reference.

13.3.3 Benchmarks

By a *benchmark test* we mean an efficient, easily administered test, or set of tests, that can be used to express the performance of a speech output system (or some module thereof) in numerical terms. The *benchmark* itself is the value that characterises some reference system, against which a newly developed system is (implicitly) set off. The benchmark is preferably chosen such that it represents a performance level that is known to guarantee user satisfaction. Consequently, if the performance of a new product exceeds the benchmark, its designer or prospective buyer is assured of at least a satisfactory product, and probably even better. Obviously, testing against a benchmark is more efficient than pairwise or multiple testing of competing products.

At this time it is too early to talk about either existing benchmarks or benchmark tests. It is clear, however, that the development of benchmarking deserves high priority in the speech output assessment field. As a first step, existing tests should be scrutinised for their potential use as benchmark tests. Choices should be make as to what aspects to include in benchmark tests (overall performance, composite performance by a number of crucial modules), and what system to adopt as the reference on which the benchmark value should be based. In this respect, it seems to us that one should not adopt the performance of human speech as the benchmark. Human speech, at least when produced by professional talkers, will simply be too good for the purpose of benchmarking. Since human speech will always be superior to synthetic speech, the quality of the latter will have to be expressed as a fraction, which makes it hard to compare the relative differences between different types of synthetic speech. What we need is a speech output system of proven, but still imperfect, quality. This is, quite probably, the reason why the quality of many speech output systems for English is often expressed relative to the 'Paul' voice of MITalk/DECTalk, which has long served as the de facto standard in TTS.

13.3.4 Reference conditions

Next to a widely accepted benchmark, it would appear to us that designers of speech output systems should want to know how well their systems perform relative to some optimum, and what performance could be expected of a system that contains no intelligence at all. In other words, the designer is looking for *topline* and *baseline* reference conditions. Reference conditions such as these do not yield diagnostic information in the strict sense of the word. However, they do provide the systems developer with an estimate of how much improvement can still be made to a system as a whole (in a black box approach) or to specific modules (in a glass box approach).

13.3.4.1 Segmental reference conditions

There has been no general practice to include topline and baseline reference conditions in *segmental* quality testing (section 13.5.2.1). Still, it seems to us that it is important to reach consensus on a number of measures. If the output system uses waveform concatenation techniques, the designer will want to know how well the synthesis performs relative to the live human speaker, or to facilitate procedures, to some electronic operationalisation of live speech (e.g. CD quality speech recorded at a short distance from the speaker's mouth in a quiet environment). However, if the system's waveforms have been coded with a lower bitrate than CD quality, the designer should determine to what extent degradation of system performance is due to the synthesis itself as opposed to the non-optimal bitrate. An easy way to determine this, is to adopt a second reference condition using the same (lower) bitrate as the synthesis. This precaution is even more necessary for parametric synthesis. Obviously, no type of parametric synthesis can be better than the maximum quality that is afforded by the analysis-resynthesis coding scheme adopted for the synthesiser. This requirement can generally be fulfilled when LPC synthesis schemes are used. However, for a range of synthesisers (e.g. the Klatt and the JSRU synthesisers) no automatic parameter estimation for straightforward analysis-resynthesis is possible at this time. The optimal parametric representation of human reference materials will then have to be found by trial and error, or else the attempt should be abandoned.

The designer of an output system claims that the intelligence incorporated into the synthesis systems (e.g. through rules) makes the systems perform better than with no intelligence built in at all. In order to establish the extent to which this claim is true, a baseline condition is needed which consists in a type of synthetic speech that has no knowledge of speech processes at all.

Recommendation 12
Absolute segmental topline: In the case of allophone synthesis, use human speech produced by a designated talker, i.e. the same individual on whose speech the table values and synthesis rules were based, or who, in the case of concatenative synthesis, provided the basic synthesis building blocks. The absolute topline reference will then be based on CD-quality digital speech.

Recommendation 13
Relative segmental topline for parametric synthesis: A second useful topline reference is the human

reference speech (see Recommendation 12) but analysed and (re-)synthesised using exactly the same coding scheme that is employed in the speech output system to be tested.

Recommendation 14
Relative segmental topline for waveform concatenation: Use the same (lower) bitrate in the reference condition as in the speech output system.

Recommendation 15
Segmental baseline for allophone synthesis: Use speech in which all segments retain their table values and are strung together merely by smoothing spectral discontinuities at segment boundaries.

Recommendation 16
Segmental baseline for concatenative synthesis: Use speech made by stringing together coarticulatory neutral phones (i.e. stressed vowels spoken between two /s/-es, or stressed consonants preceded by schwa and followed by an unrounded central vowel, cf. the 'neutrone' condition in Van Bezooijen & Pols (1993)). Minimal smoothing should be applied to avoid spectral jumps.

13.3.4.2 Prosodic reference conditions

The need for suitable topline and baseline reference conditions has clearly been recognised in the field of *prosody* (i.e. temporal and melodic structure, cf. section 13.5.2.2) testing. The following are recommendations for prosodic topline and baseline conditions. Note that, in contrast to segmental evaluation, listeners often find it very difficult to differentiate between different prosodic versions of an utterance. Therefore testers often need examples of 'very bad' systems to check whether the listeners are indeed sensitive to prosodic differences.

Recommendation 17
Temporal and melodic topline: Copy, as accurately as possible within the limitations of the synthesiser, the temporal structures and speech melodies of a single designated professional human speaker onto the synthetic speech output.

Recommendation 18
Temporal baseline: Use a condition in which the smallest synthesis building blocks (phoneme, diphone, demisyllable) retain their original, unmanipulated durations as they were copied from the human original from which they were extracted (or, in the case of allophone synthesis, the phoneme duration table values, cf. Carlson, Granström & Klatt (1979)).

This baseline condition, then, contains no intelligence, so that any improvement in the target conditions with duration rules must be due to the added explicit knowledge on duration structure. A reference in which segment durations vary at random (within realistic bounds) can be included for validation purposes, as an example a 'very bad system'. Listeners should rate this condition as poorer than any other condition.

Recommendation 19
Melodic baselines: Synthesise utterances on a monotone, at a pitch level that coincides with the average pitch of the test items. Also, include a random melodic reference for the sake of validation, by introducing random pitch variations (in terms of excursion size, rate of change, and segmental alignment), within physiologically and linguistically reasonable limits and with a mean pitch equal to the average of the test items.

There is a practical problem that not every synthesiser allows the generation of monotonous pitch so that some sort of waveform manipulation (e.g. pitch synchronous overlap and add, PSOLA) may have to be used in order to monotonise the synthetic melody.

13.3.4.3 Voice characteristics reference conditions

In the area of *voice characteristics* (voice quality, section 13.5.2.3), the problem of reference conditions has not been recognised. Generally, there seems to be little point in laying down a baseline reference for voice quality. The choice of a suitable topline would depend on the application of the speech output system. If the goal is personalised speech output (for the vocally handicapped) or automatic speaker conversion (as in interpreting telephony), the obvious topline is the speaker who is being modelled by the system, using the same coding scheme when applicable. When a general purpose (i.e. non-personalised) speech output system is the goal, one would first need to know the desired voice quality, i.e. ideal voices should be defined for specific applications, and speakers should be located who adequately represent the ideal voices. At this time we will refrain from making any further suggestions on this matter. The definition of 'ideal' voices and voice qualities, and the implementation of topline references should be a matter of priority in the near future.

13.3.4.4 Overall quality reference conditions

Given the existence of an overall quality topline reference condition, it would be advantageous to have a set of reference conditions that are poorer than the optimum by a number of calibrated steps until a quality equal to or less than the baseline reference is reached (see also section 13.4.1.2). Such a set of reference conditions would yield a grid within which each type of speech, whether produced by humans or by machines, can be located and compared with other types of speech. Recently, attempts have been made at creating such a continuum of reference conditions by taking high-quality human speech and applying some calibrated distortion to it, such as multiplicative white noise at various signal-to-noise ratios ('Modulated Noise Reference Unit or MNRU', cf. ITU-T Recommendation P.81), or time-frequency warping (TFW, ITU-T Recommendation P.85, cf. Burrell (1991); or T-reference, cf. Cartier et al. (1992)).

TFW introduces greater or lesser (random) deviations from the mean rate of a recorded utterance ($\pm$2.5%, $\pm$5%, ... $\pm$20%) over successive stretches of 150 ms, so that the speech contains potentially disturbing accelerations and decelerations and associated frequency shifts. Fellbaum, Klaus & Sotscheck (1994) showed that the MNRU is not suitable for the evaluation of synthetic speech. TFW of natural speech, however, provided a highly sensitive reference grid within which TTS systems could be clearly differentiated from each other in terms of judged listening effort (Johnston (1993)). Moreover, Johnston showed that the perceived quality ordering among a range of TTS systems interacts with the sound pressure level at which the speech output is presented.

Recommendation 20
Use time-frequency warping of optimal human speech to create a grid of overall quality reference conditions.

13.3.5 Comparability across languages

Although it is generally agreed that in the final analysis all languages are equally complex, it cannot be denied that phonetic and phonological complexity differs widely from one language to the next. Languages differ in the size of their vowel and consonant inventories, in the complexity of syllable structures, stress rules, reduction processes, and so forth.

A number of systems are (commercially) available that provide multi-lingual speech output (e.g. DECTalk, Infovox, Multivox, Apollo). Generally, such systems were primarily developed for one language (American English, Swedish, Hungarian, and British English, respectively), and additional language modules were derived from the original language by making minimal changes to the basic units and rules. As a result it is commonly observed that the derivate languages of multi-lingual systems sound poorer than the original. Yet, it is very difficult to establish this convincingly, since the poorer performance may be due (completely or in part) to the greater intrinsic difficulty of the sound system of the new language. Ultimately one would like to develop speech output assessment techniques that allow us to determine the quality of a system speaking language A and to compare its quality to that of another system speaking language B. In order to reach this objective, we would have to know how to weigh the scores obtained for a language for the intrinsic difficulty or complexity of the relevant aspects in that language.

Such goals will not easily be accomplished. However, steps have been taken in the SAM project to ensure optimal cross-language comparability in the construction of the test materials and administration procedures. For example, in the Semantically Unpredictable Sentence Test (SUS Test, see Appendix 1 G), the same five syntactic structures (defined as linear sequences of functional parts of speech, e.g. Subject–Verb–Direct Object) are used in all languages tested, and words are substituted in each of the designated syntactic slots that are selected from the same lexical categories, and with the shortest word length allowed by the language (see G in Appendix 1). It should be obvious, however, that complete congruence cannot be obtained in this fashion: the shortest content words in Italian and Spanish are typically disyllables, whilst they are monosyllabic in the Germanic languages. Similarly, although all five syntactic structures occur in each of the languages tested, certain structures will be more common in one language than in an other. Given the existence of such intrinsic and unavoidable structural differences between languages, we recommend further research into the development of valid cross-language normalisation measures.

Especially when working within the European Union, with its increasing number of partner countries and languages, speech output products are likely to be produced on a multilingual basis. The further development of efficient testing procedures that can be validly used for all relevant languages is a clear priority. Yet, as we explained in section **??**, we should not raise our hopes too high in this matter, given the existence of intrinsic and unavoidable structural differences between languages. For this reason we recommend parallel research into the development of valid cross-language normalisation measures, that will allow us to realistically compare speech output test results across languages, if the choice of test materials cannot be balanced in all relevant linguistic aspects.

In this effort, ITU recommendation P.85 has potential. Following this procedure (see section 13.3.4.4) a reference grid can be constructed for each (EU) language. One possible outcome could be that some languages prove more resistent to time-frequency warping than others (although we hesitate to make any predictions). Be this as it may, differences in intelligibility between languages would be effectively normalised out when we determine the quality of an output system relative to the reference grid that is applicable for the language being tested.

13.4 Black box approach

13.4.1 Laboratory testing

13.4.1.1 Functional laboratory tests

Black box assessment tests a system's performance as a whole, without considering the performance of modules internal to the system. Ideally, within black box testing, one would want to have at one's disposal a functional test to assess the adequacy of the complete speech output in all respects: does the output function as it should? Such a test does not exist, and is difficult to conceive. In practice, the functional quality of overall speech output has often been equated with comprehensibility: to what extent can synthesised continuous speech be understood by listeners?

Speech comprehension is a complex process involving the interpretation and integration of many sources of information. Important sources of information in complete communication situations, where both auditory and visual information are available to interactants, are:

1. Speech signal information at different levels (segments, prosody, voice characteristics),
2. Segment combinatory probabilities (e.g. /str../ is a permissible consonant sequence at the onset of words in many EU languages, but all other permutations of this sequence are illegal (e.g. /tsr.../,
3. Knowledge of which segment strings are existing words in the language (e.g. the permissible string /strIk/ is not a word in English),
4. Word combinatory probabilities (e.g. the article "the" will tend to be followed by nouns rather than verbs),
5. Semantic coherence (e.g. in the context of "arrive" a word like "train" is more probable than a word like "pain"),
6. Meaning extracted from the preceding linguistic context; due to the repetition of words and the progressive building up of meaning, the last sentence of a text will generally be easier to understand than the first,
7. World knowledge and expectations of the listener based on previous experience,
8. Cues provided by the extra-linguistic context in which the message is spoken (e.g. facial expressions and gestures of the speaker, relevant things happening in the immediate environment).

In normal daily life all these different sources, and others, may be combined by listeners to construct the meaning of a spoken message. As a result, in applied contexts the contributions of separate sources are difficult to assess. Laboratory tests typically try to minimise or control for the effects of at least some of the sources in order to focus on the auditory input. Some segmental intelligibility tests at the word level (such as the SAM Standard Segmental Test, see Appendix I A) try to minimise the effects of all sources except (1) and (2): only meaningless but permissible consonant-vowel-consonant combinations (e.g. /hos/) or even shorter items (/ze, ok/) are presented to the listener. In comprehensibility tests, factor (8) is excluded completely and (7) as far as possible. The latter is done by selecting texts with supposedly novel information for all subjects.

No completely developed standardised test, with fixed test material and fixed response categories, for evaluating comprehension is available, but one wonders whether this would be very useful in the first place, since it is not clear what the "average" text to be used should look like in terms, for example, of the complexity and type of vocabulary, grammatical structures, sentence length, and style. At this level of evaluation it seems a good idea to take the characteristics of the intended application into account [3]. Testing the comprehensibility of speech output destined to provide traffic information asks for a more specific type of test materials (e.g. short sentences, only statements, restricted range of lexical items, formal style) than speech output to be used for reading a digital daily newspaper for the blind, where the test materials should be more varied in all respects. The greatest variation should probably be present in speech material testing text-to-speech systems developed to read novels to the visually handicapped.

As to the type of comprehension test, several general approaches can be outlined. The most obvious one involves the presentation of synthesised texts at the paragraph level, preferably with human produced versions as a topline control, with a series of open or closed (multiple choice) questions.

[3] Clearly, there is a continuum from completely application independent at the one end and completely application specific at the other end. The distinction between section ?? and ?? is therefore somewhat artificial.

Results are expressed in terms of percent correct responses. An example of a closed response approach is Pisoni, Greene & Nusbaum (1985), Pisoni, Nusbaum & Greene (1985), who used 15 narrative passages selected from standardised adult reading comprehension tests. Performance was compared between listening to synthetic speech, listening to human speech, and silent reading. Each condition was tested with 20 subjects. One of the most important findings was a strong learning effect for synthetic speech within a very short time and the absence of clear differences among the test conditions.

At first sight, the results of closed response comprehension tests seem counterintuitive: although the human produced texts sound better than the synthetic version, often no difference in comprehension is revealed (Nye, Ingemann & Donald (1975); Delogu et al. (1992)) or, after a short period of familiarisation, even superior performance for synthetic speech (Pisoni, Greene & Nusbaum (1985), Pisoni, Nusbaum & Greene (1985)) is observed. These results have been tentatively explained by hypothesising that subjects may make more of an effort to understand synthetic speech. This could be expected to lead to:

- Slower reaction times in a sentence verification test.
- A decrease in performance as a function of fatigue.
- Poorer performance for secondary tasks.

Confirmation of the first prediction was found by Manous et al. (1985). The second and third predictions were tested by Luce, Feustel & Pisoni (1983), using a word recall test, and by Boogaart & Silverman (1992), using a tracking task. The first study revealed a significant effect, whereas the second did not.

However, the lack of differentiation in comprehensibility between human and synthetic speech in the above studies may also be due to the use of the closed response approach, where subjects have a fair chance of guessing the correct answer. Open response tests are known to be more sensitive, i.e. more apt to bring to light differences among test conditions. An example of an open response study is Van Bezooijen (1989). She presented five types of texts typically found in daily Dutch newspapers, pertaining to the weather, nature, disasters, small events, and sports, to 16 visually handicapped subjects. An example of a question testing the comprehensibility of the weather forecasts is: *What will the temperature be tomorrow?* The questions were sensitive enough to yield significant differences in comprehensibility among two text-to-speech converted (one automatic and one manually corrected) and a human produced version of the texts. Crucially, the results also suggest that the effect of the supposedly greater effort expended in understanding synthetic speech has its limits. If the synthetic speech is bad enough, increased effort cannot compensate for loss of quality.

The tests described ask subjects to answer questions after that the texts have been presented, thus measuring the final product of text interpretation. In addition to these off-line tests, more psycholinguistically oriented on-line approaches have been developed which request instantaneous reactions to the auditory material being presented. These tests primarily aim at gaining insight into the cognitive processes underlying comprehension: to what extent is synthetic speech processed differently from human speech? To name but a few of these psycholinguistic tests:

- The *word monitoring task*, requiring subjects to press a button as soon as they hear a pre-specified word,
- The *sentence-by-sentence listening task*, in which subjects push a button whenever they are ready for hearing the next sentence (comprehension is checked afterwards but is not part of the test proper),

- The *sentence verification test*, where subjects have to decide whether short sentences are true statements or not (e.g. *Mud is dirty* and *Rockets move slowly*).

All three are on-line measures, the first indexing cognitive workload, the second and third assessing speed of comprehension. On-line tests of this type, which invariably reveal differences between human and synthetic speech, have been hypothesised to be more sensitive than off-line measures (Ralston et al. (1991)). However, the results of such psycholinguistic tests ("subjects responded significantly faster to system A (740 ms) than to system B (930 ms)") are less interpretable for non-scientists than those of comprehension tests ("subjects answered 74% of the system A questions correctly versus 93% of the system B questions). On the other hand, insight into cognitive load may ultimately prove important in double task applications.

Recommendation 21
Try to avoid the use of functional tests to assess overall output quality: on-line reaction time tests are difficult to interpret and off-line comprehension tests are difficult to develop.

Recommendation 22
If determined to develop a comprehension test, beware of the fact that reading tests may be too compact to be used as listening tests; adapt the materials or use materials that are meant to be listened to.

Recommendation 23
Use open comprehension questions rather than closed ones, the former being more sensitive than the latter.

Recommendation 24
When administering a comprehension test, include a top-line reference with a dedicated speaker realising exactly the same texts presented in a synthetic version; use different groups of subjects for the various speech conditions (or better still block conditions over listeners such that no listener hears more than one version of the same text while at the same time each listener gets an equal number of different text versions).

Recommendation 25
When interpreting comprehension results, look at difference scores (synthetic compared to human) rather than at absolute scores to abstract from the intrinsic difficulty of questions.

13.4.1.2 Judgment laboratory tests

The black box tests described so far are functional in nature. However, instead of evaluating overall quality functionally, subjects can also indicate their subjective impression of global quality aspects of synthetic output by means of rating scales. Taking comprehensibility as an example, a functional task would be one where subjects answer a number of questions related to the content of a text passage as described above. Alternatives from a judgment point of view include:

- *Paired comparison*, where subjects indicate which of two synthesisers sounds more comprehensible,
- *Magnitude estimation*, where subjects assign a value expressing, or draw a line of a length which is equal to the magnitude of, their impression of comprehensibility,
- *Categorical estimation*, where subjects rate synthesisers, for instance, along a 10-point scale which runs from 1: extremely incomprehensible to 10: extremely comprehensible.

Some methodological aspects of the second and third method are described in detail in section 13.3.2. There it is also indicated that magnitude estimation is relatively laborious and more fit for test external comparison, whereas categorical estimation is relatively fast and easy, and more fit for test internal comparison.

Both the magnitude (continuous scale) and categorical estimation (20-point scale) methods have been included in SOAP in the form of the SAM Overall Quality Test (see K in Appendix 1). Three scales are recommended, related to:

- *Intelligibility* (How identifiable does the message sound?),
- *Naturalness* (To what extent does the message sound like being produced by a human speaker?),
- *Acceptability* (The overall user's satisfaction with the communicative situation).

The intelligibility and naturalness ratings are based on pairs of (unrelated) sentences. Fixed lists of 160 sentences of varying content and length are available for Dutch, English, French, German, Italian, and Swedish. Examples for English are: *I realise you're having supply problems but this is rather excessive* and *I need to arrive by 10.30 a.m. on Saturday.* For the acceptability ratings, application specific test materials are recommended. The magnitude and categorical estimation procedures have been applied to speech output in a number of studies (e.g. Pavlovic, Rossi & Espesser (1990); Delogu et al. (1991); Goldstein, Lindström & Till (1992)). Methodological aspects, such as the effects of stimulus range and the number of categories, relationships among methods, reliability, and validity, are emphasised.

The importance of application-specific test materials is also stressed by the International Telecommunication Union Telecommunication Standardisation (ITU-T) sector (see L in Appendix 1). They developed a test specifically aimed at evaluating the quality of telephone speech (where synthesis could be the input). It is a categorical estimation judgment test comprising ratings on (a subset of) eight scales:

- *Acceptance*
- *Overall impression*
- *Listening effort*
- *Comprehension problems*
- *Articulation*
- *Pronunciation*
- *Speaking rate*
- *Voice pleasantness*

The first scale is a 2-point scale, the other ones are 5-point scales. Strictly speaking, only the first four scales can be captured under the heading *overall quality*; the other four scales are directed at more specific aspects of the output and require analytic listening. The content of the speech samples presented should be in accordance with the application. Examples of application-specific test items are: *Miss Robert, the running shoes Adidas Edberg Pro Club, colour: white, size: 11, reference: 501-97-52, price 319 francs, will be delivered to you in 3 weeks* (mail order shopping) and *The train number 9783 from Poitiers will arrive at 9:24, platform number 3, track G* (railway traffic information). In addition to rating the eight scales, subjects are required to reproduce information contained in the message. A pilot study has been run by Cartier et al. (1992). Fellbaum, Klaus & Sotscheck (1994) tested 13 synthesis systems for German using the ITU-T Overall Quality Test as well as open response functional intelligibility tests. Waveform concatenative synthesis systems proved measurably better than formant synthesis systems.

Van Bezooijen & Jongenburger (1993) employed a similar series of judgment scales as proposed by the ITU-T in a mixed laboratory/field study which addressed the suitability of synthetic speech within the context of a digital daily newspaper for the blind (see section 13.4.2.2). Their battery comprised ten 10-point scales:

- *Intelligibility*
- *General quality*
- *Naturalness*
- *Precision of articulation*
- *Accuracy of pronunciation*
- *Pleasantness of voice*
- *Adequacy of word stress*
- *Appropriateness of tempo*
- *Liveliness*
- *Fluency*

Again a distinction can be made between scales relating to overall quality (the first three scales), and the other scales, relating to specific aspects of the speech output. A factor analysis yielded two factors, the first with high loadings of intelligibility, general quality, and precision of articulation, the second with high loadings of naturalness, pleasantness of voice, and adequacy of word stress. Intelligibility and naturalness were taken by the authors to be the two central dimensions underlying the evaluative judgments.

Recommendation 26
Since there is no consensus on the most appropriate judgment scales to evaluate overall quality, choose between:

- *Intelligibility, naturalness, and acceptability (SAM Overall Quality Test),*
- *Acceptance, overall impression, listening effort, and comprehension problems (ITU-T), or only listening effort (practice in telephony),*
- *Intelligibility, general quality, and naturalness (Van Bezooijen & Jongenburger (1993))*

Recommendation 27
It is important that the scale positions have a clear meaning for the subjects and that the scale is wide enough to allow differentiation among systems compared. Use at least a 10-point scale.

13.4.2 Field testing

13.4.2.1 Preliminary remarks

In the previous section, the black box approach to speech output evaluation was operationalised within a laboratory context. From an experimental point of view, the main advantage of a laboratory study is the control over possibly interfering factors. However, ultimately it is the functioning of a speech output system in real life, with all its variability, that counts. If overall quality is extended to include all aspects of the synthesis in the context of an application, testing may be necessary in the field. Due to the variety of applications, it is difficult to summarise the aspects which field tests have in common. To illustrate the diversity, below some examples will be given.

13.4.2.2 Field tests

A combined laboratory/field functional/judgment test, with equal attention to the speech output itself and the context within it is used, was done by Van Bezooijen & Jongenburger (1993). They used the following suite of four tests to evaluate the functioning of an electronic newspaper for the visually handicapped:

- An interview enquiring after the subjects' attitudes towards the technology,
- A functional open response identification test with CVC-words,
- Judgments on 10 evaluative scales for continuous text passages (1: extremely bad, 10: extremely good) related to global quality and more specific aspects of the speech output, such as pleasantness of voice (see 13.3.4.3), adequacy of word stress, appropriateness of tempo, liveliness, and fluency (see 13.4.1.2 for further details),
- A functional test assessing the subjects' proficiency in finding their way through the newspaper. This test involved a number of searches, such as *Is there an article on Japan in the economy section?* Proficiency was assessed both in terms of percent correct answers and task completion time.

Each of 24 visually handicapped subjects was visited at home, at three points in time. Since the subjects lived scattered all over the Netherlands, administration of the suite of tests was very time consuming.

Comparable studies have been conducted to evaluate a digital daily newspaper in Sweden (Hjelmquist, Jansson & Torell (1987)). However, the experimental set-up to assess the quality of various aspects of the Swedish speech output was less strict: most information was obtained through interviews. On the other hand, much emphasis was placed upon the reading habits of the users: all keystrokes were registered during long periods of time, so that the frequency of use of all reading commands (e.g. *next sentence, previous sentence, talk faster, talk in letters*) could be determined.

A semi-field study combining function and judgment testing within the context of telephone information services was done by Roelofs (1987). In this test resynthesised human speech was used, but the set-up and results can be generalised to synthetic speech output. Two applications were considered, namely directory assistance (the subject puts his request to an operator and then the number is spoken by the computer twice, thus freeing the operator for the next subscriber) and a service for train departure times (in a single pre-stored message the departure times of a number of trains with different destinations are given). In the former application a human operator served as a reference, in the second high-quality PCM speech was presented. Subjects were sent the instructions in advance and dialled the two services from their homes. The availability of interrupt facilities and speaking rate were examined. Both actual performance (success in writing down the requested data) and subjective reactions were registered (14 5-point scales such as bad–good, impersonal–personal, inefficient–efficient) and two questions, namely: *Do you find this way of information presentation acceptable?* and *Do you think this service could replace the current service?* Due to several factors, the results are of limited value. However, the method is a nice example of how different approaches to testing can be combined.

With a view of exploring the possibilities of synthetic speech for a name and address telephone service, Delogu et al. (1993) tested six Italian TTS systems by presenting lexically unpredictable VCV and CV sequences in an open response format. Intelligibility scores dropped from 31 to 21 percent when the same materials were listened to through a telephone line rather than good quality headphones. Curiously enough, the best TTS systems suffered most from telephone bandwidth limitation.

Finally, an important area of speech output evaluation are applications where people are required to process auditory information while simultaneously performing some other task, involving hands and eyes, for instance to write down a telephone number or land an aircraft. The requirements imposed by double tasks like these have been simulated for instance by having subjects answer simple questions related to the content of short synthesised messages while at the same time tracking a randomly moving square on a video monitor by moving a mouse (Boogaart & Silverman (1992)). This type of laboratory study could and should be extended to more real-life situations. Other important areas are field tests where the functioning of speech output is tested under various noise conditions, and combinations of noise and secondary tasks.

Since field tests will often have to meet specific requirements, it is not realistic to think in terms of standard tests and standard recommendations. Each case will have to be examined in its own right. In order to get an overview of complex test situations that may arise, Jekosch & Pols (1994) recommend a "feature" analysis to define a test set-up, where features are all aspects relevant to the choice of the test. Their analysis comprises three steps, naturally leading to a fourth step:

1. Determine the application conditions (What is to be tested? What are the properties of the material generated?), resulting in a feature profile of the application scenario.
2. Define the best possible test matching this feature profile.
3. Make a comparison of what is desired and what is available in terms of tests.
4. Adapt tests or develop your own test.

Because of the specific nature of some applications, often there will be no ready-made test available, so that it is perhaps better to talk of (suggestions for) test approaches than tests. Moreover, a single test will generally not suffice, but a suite of tests will be needed instead. In this suite both functional and judgment tests can be included. Interviews can be part of the evaluation as well. Moreover, it is possible to administer laboratory type experiments in a field situation. This can be done, for example, by preparing stimulus tapes beforehand and playing them to subjects in the environment where the synthesis system will be used.

13.5 Glass box approach

13.5.1 Linguistic aspects

In this section we shall deal with evaluation procedures that have been, or can be, followed when modules in a text-to-speech system yield some intermediary symbolic output. As was stated above, there are no established methods for evaluating the quality of linguistic modules in speech output testing. As a result there is no agreed-upon methodology in this area nor are there evaluation experts; what little evaluation work is done, is done by the same researchers who developed the modules. In view of the lack of an established methodology we will refrain from making recommendations on the use of specific linguistic tests and test procedures. The need for a more general research effort towards a general methodology in the field of linguistic testing will be discussed in section 13.6.3.

13.5.1.1 Preprocessing

The first stage of a linguistic interface makes decisions on what to do with punctuation marks and other non-alphabetic textual symbols (e.g. parentheses), and expands abbreviations, acronyms, numbers, special symbols, etc. to full-blown orthographic strings, as follows:

abbreviations	"i.e."	→	*that is*
	"viz."	→	*namely*
acronyms	"NATO"	→	*naytoe*
	"UN"	→	*you en*
numbers	"124"	→	*one hundred and twenty four*
	"1:24"	→	*twenty four minutes past one*
special symbols	"#1"	→	*number one*
	"£1.50"	→	*one pound fifty*

There are no standardised tests for determining the adequacy of text preprocessors. Yet is seems that all preprocessors meet with the same classes of transduction problems, so that it would make sense to set up a multi-lingual benchmark for preprocessing. Laver et al. (1988), Laver, McAllister & McAllister (1989)), describing the internal structure of the CSTR text preprocessor, mention a number of transduction problems and present some quantification of their errors in the various categories, which we recapitulate in Figure 13.2. The test was run on a set of anomaly-rich texts taken from newspapers and technical journals.

Figure 13.2: *Percent correct treatment of textual anomalies by CSTR text preprocessor (after Laver et al. (1988: 12–15))*

Anomaly type	No. tested	No. correct	Percent correct
Meta-textuals	95	95	100
Capital-initials	87	87	100
Digit-bearing	35	34	97
Hyphens	15	15	100
Other	24	24	100
Proper names			73
Strings in dictionary			85
Strings not in dictionary			87
Number strings			100

The results in Figure 13.2 are not so much revealing in terms of the numerical information they offer, but by the taxonomy of errors opted for. The only other formal evaluation of a text preprocessor that we have managed to locate uses a completely different set of error categories. Van Holsteijn (1993) presents an account of a text preprocessor for Dutch, and gives the results of a comprehensive evaluation of her module. She observes that the use of abbreviations, acronyms and symbols differs strongly from text to text. She broadly distinguishes three types of newspaper text: editorial text on home and foreign news, editorial text on sports and business, and telegraphic-style text (i.e. classified adds, film & theatre listings, radio & television guide). Text segmentation errors were separately counted for:

1. Sentence demarcation.
2. Expression demarcation (within sentence subunits that need some specific type of preprocessing).

Correctly demarcated expressions could then be characterised further in terms of:

1. Labelling errors,
2. Expansion errors.

Finally, a distinction is made between unavoidable and avoidable errors. The former type would be the result of incorrect or unavailable syntactic/semantic information that would be needed in order to choose between alternative solutions. The latter type is the kind of error that needs correction, either by the addition of new rules or by inclusion in the exceptions lexicon. Figure 13.3 presents some results.

Figure 13.3: *Evaluation results for text preprocessor TextScan (after Van Holsteijn (1993)). Percent of avoidable errors in four categories; percent unavoidable errors in parentheses; N specifies the 100% base per cell.*

	(1) segm. sent	(2) segm. expressions	(3) label. express.	(4) expan.expr
A.	0.1 (0.4) N=786	< 0.01 (< 0.01) N=13699	0.9 (3.8) N=1904	1.5 (0.0) N=479
B.	0.2 (0.4) N=643	< 0.01 (< 0.01) N=10388	0.3 (4.0) N=1683	0.0 (0.0) N=571
C.	0.0 (0.5) N=202	0.0 (2.0) N= 2570	0.9 (2.5) N=1231	0.7 (0.4) N=560

Both Laver et al. (1988)'s and Van Holsteijn (1993)'s proposals represent rather crude, and disparate, approaches towards a taxonomy of errors of a text preprocessor. What is clearly needed for the evaluation of text preprocessors, is a more principled analysis of the various tasks a text preprocessor has to perform, focussing on those classes of difficulties that crop up in any (European) language. Procedures should be devised that automatically extract representative items from large collections of recent text (newspapers) in each of the relevant error categories, so that multi-lingual tests can be set up efficiently. Once the test materials have been selected, the correct solutions to, for instance, expansion problems can be extracted from existing databases, or when missing there, will have to be entered manually.

13.5.1.2 Grapheme-phoneme conversion

By grapheme-phoneme conversion we mean a process that accepts a full-blown orthographic input (i.e. the output of a preprocessor), and outputs a string of phonemes. The output string does not yet contain (word) stress marks, (sentence) accent positions, and boundaries. The correct phonemic representation of a normally spelled word depends on its linear context and hierarchical position (e.g. assimilation to adjacent words: *I have to go* /aI h{*f* t@ g@U/ but *I have two goals* /aI h{*v* tu: g@Ulz/; or the choice between homographs: *I lead* /li:d/ but *made of lead* /led/. Therefore the adequacy of grapheme-phoneme conversion modules should not, in principle, be tested on the basis of isolated word pronunciation (citation forms). In practice, however, this is precisely what is done. The reasons for this are threefold:

- For many languages pronunciation databases (or machine readable pronouncing dictionaries) are available, which are exclusively based on isolated words.
- The adaptation rules for word pronunciation in context are (believed to be) straightforward, exceptionless, and easy to implement, and
- Machine readable phonemic transcriptions of continuous prose are scarce.

Figure 13.4 presents results of a multi-lingual evaluation of grapheme-phoneme converters for seven EU languages, performed within ESPRIT 291/860 "Linguistic analyses of European languages," based on isolated word pronunciation. Since it has often been reported that many more conversion errors occur in proper names than in ordinary words, the evaluation distinguished between four types of materials:

- Newspaper texts of at least 100,000 words,
- The names of the 150 largest towns per country,
- Names of 31 European capitals in the national spelling, and
- The 150 most frequent Christian names for each language.

Figure 13.4: *Percent correct grapheme-phoneme conversion in seven EU languages in four types of materials (after Pols (1991: 394))*

Language	Newspaper	Towns	Capitals	First names
Dutch	98.9	85.3	96.7	89.4
English	90.3	46.0	58.0	58.0
French	96.9-94.5	77.3	74.2	84.7
German	93.0-90.0	81.0	61.0	80.0
Greek	98.7	97.3	93.5	97.3
Italian	85.2	85.3	80.6	86.9
Spanish	98.9	95.3	96.6	98.0

Note: Newspaper scores are weighed for token frequency. Higher first score for French excludes all preprocessing errors; higher first German score is based on the use of an exceptions list.

Incidentally, the results should not be taken to indicate that spelling is harder to convert to phonemes in Italian than in any other language, since different conversion methods were used in each language; however, Italian proper names are no more a problem than ordinary text words. In English and French spelling the proper names do present a serious problem, so that exceptions lists will be a priority there.

In a complementary test Nunn & Van Heuven (1993) compared the performance of three grapheme-phoneme converters for Dutch, i.e. two systems with no or only implicit morphological decomposition (Kerkhoff, Wester & Boves (1984); Berendsen, Langeweg & Van Leeuwen (1986) and one that included the MORPA morphological decomposition module. About 2,000 simplex and complex (see section 13.5.1.4) test words were selected from newspaper texts that did not belong to the 10,000 most frequent Dutch words, so that dictionary look-up would fail. Phoneme, syllabification, and stress placement errors were found by automatic comparison with a hand-made master transcription file. The earlier converters performed at a success rate of 60% and 64%, which is considerably poorer than the newspaper text score in Figure 13.4. The newer system with explicit morphological decomposition was correct in 78 percent.

13.5.1.3 Word stress

Stressed syllables are generally pronounced with greater duration, greater loudness (in terms of acoustical intensity as well as pre-emphasis on higher frequencies), and greater articulatory precision (no consonant deletions, more peripheral vowel formant values). Moreover, when a word is presented in focus, a prominence-lending fast pitch movement is executed on the stressed syllable of that word. With the exception of French, where the stress is always on the last full syllable of the word, the position of the stress varies from word to word in all other EU languages. However, stress position in these languages is predictable to a large extent by rules that look at:

- The internal make-up of words (in terms of their constituent morphemes, the hierarchical relationships between adjacent morphemes, and the lexical categories of these morphemes), and
- The segment structure of the syllables making up the morphemes.

All the EU languages have a proportion of idiosyncratic words that do not comply with the proposed stress rules for diverse reasons. Therefore the coverage of stress rule systems has to be evaluated, and errors have to be corrected by including the problematic words in an exceptions dictionary.

Tests of stress rule modules have been performed only on an ad hoc basis, either checking the output of the rules by hand (Barber et al. (1989) for Italian), or automatically (using the phonemic transcription field in lexical databases containing stress marks (Langeweg (1988) for Dutch), which in turn had been checked by hand in some earlier stage of the database development) [4].

13.5.1.4 Morphological decomposition

In morphological decomposition orthographic words are analysed into morphemes, i.e. elements belonging to the finite set of smallest sub-word parts with an identifiable meaning. Morphological decomposition is necessary when the language/spelling allows words to be strung together without intervening spaces or hyphens so as to form an indefinitely large number of complex, longer words. For many EU languages word-internal morpheme boundaries are referred to by the grapheme-phoneme conversion rules. For instance, if the English letter sequence *sh* is pronounced as /S/ when it occurs morpheme internally as in bishop, but is pronounced as /s/ followed by /h/ when a morpheme boundary intervenes, as in *mishap.*

Obviously, long and complex words will have to be broken up into smaller basic words and affixes (i.e. morphemes) before the parts can be looked up in an exceptions dictionary. If all complex words were to be integrally stored in the lexicon, it would soon grow to unmanageable proportions. For stress placement rules it is sometimes necessary to refer to the hierarchical relationships between the constituent morphemes (e.g. *'lighthouse keeper*, *light 'housekeeper*, where ' denotes main stress) and to the lexical category of the word-final morpheme (which generally determines the lexical category of the complex word as a whole, e.g. *black+bird* is a noun, *pitch+black* is an adjective). Morphological decomposition is a notoriously difficult task, as one input string can often be analysed in a large number of different ways. The hard problem is choosing the correct solution out of the many possible solutions.

[4] English presents a special problem in the assignment of stress. The elements of English compounds are typically separated by spaces, so that each element is erroneously treated as a word by itself. Moreover, the stressing of compounds in English partly depends on the semantic relationship between the words that make up the compound, and in part on purely lexical factors. A comparison of English compound stress rules developed by linguists and decision rules automatically extracted from hand-labelled phonetic databases has been reported by Sproat, Hirschberg & Yarowsky (1992).

As far as we have been able to ascertain, there are no established test procedures for evaluating the performance of morphological decomposition modules. Laver et al. (1988:12–16)) tested the morphological decomposition module of the CSTR TTS on 500 words randomly sampled from a 85,000 word type list, which was compiled from a large text corpus as well as from two machine-readable dictionaries. The output of the module was examined by hand, and proved accurate at 70% (which seems rather low considering the fact that the elements of English compounds are generally separated by spaces or hyphens).

The Dutch morphological decomposition module MORPA (MORphological PArser, cf. Heemskerk & Van Heuven (1993) compared the module's output with pre-stored morphological decompositions in a lexical database. In this comparison only segmentation errors were counted, in a sample of 3,077 (simplex and complex) words taken from weekly newspapers. The results showed that in 3 percent of the input the whole word, or part of it, could not be matched with any entry in the MORPA morpheme lexicon. The frequency of this type of error depends on the coverage of the lexicon. Erroneous analyses were generated in another 1 percent of the input words. In all other cases the correct morphological segmentation was generated, either as the single correct solution (44%), or as the most likely solution in an ordered list of candidate segmentations (48%), or as one of the less probable candidate solutions (3%) [5] Although both the accuracy and the coverage of the MORPA module seems excellent by today's standards, the module proved too slow for realistic text-to-speech applications. Processing speed is therefore an important criterion in the evaluation of morphological parsers. There will be a speed/ accuracy/coverage trade-off in the evaluation of morphological parsers.

13.5.1.5 Syntactic parsing

Syntactic analysis lays the groundwork for the derivation of the prosodic structure needed to demarcate the phonological phrases (whose boundaries block assimilation and stress clash avoidance rules) and intonation domains (whose boundaries are marked by deceleration, pause insertion and boundary marking pitch movements). Syntactic structure also determines (in part) which words have to be accented. Finally, lexical category disambiguation is often a by-product of a syntactic parser.

Although the syntactic parser is an important module in any advanced TTS, we take the view that, in principle, its development and evaluation does not belong to the domain of speech output systems. Syntactic parsing is much more a language engineering challenge, needed in automatic translation systems, grammar checking, and the like. For this reason, we refer to the chapters produced by the EAGLES Working Groups on the evaluation of Automatic Translation and Translation tools.

13.5.1.6 Sentence accent

Appropriate accentuation is necessary to direct the listener's attention to the important words in the sentence. Inappropriate accentuation may lead to misunderstandings and delays in processing time (cf. Terken (1985)). For this reason most TTS-systems provide for accent placement rules. Accentuation rules can be evaluated at the symbolic and the acoustic level.

Monaghan & Ladd (1989), Monaghan & Ladd (1990) tested the symbolic output of a sentence accent assignment algorithm applied to four English 250 word texts (transcripts of radio broadcasts). The algorithm generated primary and secondary accents, which were rated on a 4-point appropriateness scale by three expert judges. Van Bezooijen & Pols (1989) tested a Dutch accent assignment algorithm at the symbolic as well as the acoustic level (only one type of accent is postulated for

[5] On an isolated word basis, this latter category could be considered as an error; however, the error can very often be corrected post-hoc when strings of morphologically segmented words are analysed further by a syntactic parser.

Dutch) using 8 isolated sentences and 8 short newspaper texts. Two important points emerged from this study:

- Correlations between the symbolic and the acoustic evaluations were significant but rather low, which means that tests at the symbolic level are no adequate substitute for acoustic tests, and
- Ratings for isolated sentences were more favourable than for sentences in paragraphs, which means that paragraph testing is necessary if the speech output system has to produce connected text.

Again, these are scattered tests, addressing only a handful of the problems that a linguistic module has to take care of. We would recommend the development of a comprehensive test procedure that identifies categories of accent placement error at the sentence and the paragraph level. The principles that underlie accent placement are largely the same across EU languages, so that it makes sense to develop the test procedure on a multi-lingual basis.

13.5.2 Acoustic aspects

In section 13.4, speech output assessment was approached within a black box context, i.e. with an emphasis on the speech output as a whole. Black box tests are by force acoustic in nature. In the present section acoustic tests are discussed as well, however this time within a glass box context, which means that attention is focussed on the quality of separate modules, mainly with a view of diagnostic testing. The structure of this section is based on traditional views in phonetics (e.g. Abercrombie (1967)) according to which three layers are present in speech: a segmental layer (related to short-term fluctuations in the speech signal), a voice dynamics or prosodic layer (medium-term fluctuations), and a voice characteristics (or voice quality) layer (long-term fluctuations). We will make the same distinction here, 13.5.2.1 being concerned with testing segments, 13.5.2.2 with prosody, and 13.5.2.3 with voice characteristics.

13.5.2.1 Segments

13.5.2.1.1 Functions of segments

The primary function of segments, i.e. the consonants and vowels in the language, is simply to enable listeners to recognise words. Generally, when the segments are sufficiently identifiable, words can be recognised regardless of the durations of the segments and the melodic pattern. In the experience of most researchers good quality (readily identifiable) vowels are afforded by even the simplest speech synthesis systems. One reason is that most coding schemes allow adequate parametrisation of vocalic sounds (narrow band formants slowly varying with time). The synthesis of good quality consonants is an altogether different matter (due to multiple excitation signals, notion of formant not always applicable, abrupt spectral changes), and this is where most (parametric) synthesisers show defects.

Moreover, since speech extends along the time dimension, segments early in the word in practice contribute more to auditory word recognition than later segments. Trailing segments, especially in long (i.e. polysyllabic) words are often not needed to distinguish the word from its competitors. Also, stressed syllables tend to contribute more to a word's identity than segments in unstressed syllables. For these reasons it makes sense to break down the segmental quality of speech output systems for vowels and consonants in various positions (initial, medial, final), within monosyllabic and polysyllabic words, and in stressed versus unstressed syllables.

13.5.2.1.2 Segmental tests

Of all aspects of synthetic speech output, the evaluation of the segmental aspect has received most attention till now, because:

- Good segmental quality is considered to be the main prerequisite for good overall quality.
- There is general agreement about the relevant categories in terms of which quality can be assessed, namely phonemes.
- It is easy to establish accuracy: a phoneme intended to represent /f/ is identified accurately if listeners respond with /f/.

Near perfect segmental quality is essential for applications with a strong emphasis on the transmission of low-predictability information to untrained listeners, for example traffic information services and reverse telephone directory assistance (*What name and address belong to this telephone number?*). Unlike in the case of "normal" words, the pronunciation of names cannot be deduced from the context. Moreover, for names it is particularly important that each consonant and vowel be clearly enunciated because there are many near-homophones, i.e. names that differ in just one sound, and strange names which listeners may never have heard before. In applications like these, where prosody is of minor importance, the required intelligibility level can be attained for instance by making use of canned speech or waveform concatenation. In other applications, where text-to-speech is preferred, it may perhaps not be necessary for each sound to be identified correctly. However, since very little is known as yet on the specific contributions of single sounds to overall intelligibility, synthesis designers have usually taken the pragmatic position that in principle all sounds should be identifiable. In that case detailed diagnostic testing of segmental quality using a glass box approach remains necessary.

As stated above, many tests have been developed to evaluate the quality of synthetic segments. There is a basic distinction between segmental tests at the word level, where single words (meaningful, meaningless or lexically unpredictable) are presented to listeners, and segmental tests at the sentence level, where complete sentences (meaningful, meaningless, or semantically unpredictable) are presented to listeners. Within either category tests can be categorised in functional and judgment studies.

13.5.2.1.2.1 Segmental tests at the word level

Functional segmental tests at the word level

The test approach used to evaluate segments at the word level has been mostly functional, quality being expressed in terms of percent correct phoneme identification. In this section we will discuss the Diagnostic Rhyme Test (DRT), the Modified Rhyme Test (MRT), the SAM Standard Segmental Test, an anonymous type of test which we shall henceforth call the Diphone Test, the Cluster IDentification (CLID) Test, the Bellcore Test, and the (Modified) Minimal Pairs Intelligibility Test. The reasons why these tests were selected to be included are varied: because they are well-known, well-designed, easy and fast to administer and/or promising. Summary information on most tests is provided in Appendix 1.

DRT (Diagnostic Rhyme Test) and MRT (Modified Rhyme Test)

The DRT (see D in Appendix 1) is a closed response test with two response alternatives. Items presented to the subjects are of the form CVC, i.e. an initial Consonant followed by a medial Vowel followed by a final Consonant. The identifiability of the medial vowel and final consonant is not examined, only the identifiability of the initial consonant is tested. All items are meaningful, which means that only factors (1) through (3) as listed in section 13.4.1.1 can influence the results. In order to obtain insight into the precise nature of possibly poor identifiability of initial consonants, the two categories from among which the subjects are forced to select the correct response contain minimal phonemic contrasts. The subject would be asked for instance to indicate whether a synthetic item was intended as *dune* or *tune*.

The MRT (see E in Appendix 1) is an (originally) closed response test with six response alternatives. All items are of the form CVC and (in its original form) meaningful. Both the identifiability of initial and of final consonants are tested, but never simultaneously. An example of response alternatives testing the identifiability of a final consonant would be *peas, peak, peal, peace, peach,* and *peat*.

The use of meaningful test items has some positive effects:

- The DRT and MRT are reliable, fast, and easy to administer and score.
- No training is required of subjects because they can respond in normal spelling.

However, the DRT and MRT have some serious drawbacks and restrictions as well:

- Intelligibility may be overestimated since subjects adjust their perception to the response categories presented to them.
- There is a risk of ceiling effect.
- Due to their restricted coverage and their limitation to meaningful words, they have little diagnostic value.

Both the DRT and MRT have a long tradition in speech output assessment and have been used in many studies, mainly for comparative purposes. The DRT has been employed among others by Pratt (1987), who compared a wide range of synthetic voices/systems and a human reference, both clear and with noise added to give a speech-to-noise ratio of 0 dB(A). Eight subjects participated. The percentages correct for the human voice and five synthesisers are given in Figure 13.5.

Figure 13.5: *Some results obtained with the DRT by Pratt (1987)*

Speech type	Clear	0 dB(A)
Human	96	80
DECtalk Beautiful Betty	92	64
DECtalk Perfect Paul	88	60
Prose 2000 v1.2	81	42
Namal Type and Talk	66	40
Computer Concepts Speech Rom	61	30

All factors (speech system, speech-to-noise ratio, and type of phonemic contrast in the two response categories) had a significant effect on the percents correct identification. More interestingly, all interactions appeared to be significant as well. For example, as can be seen above, the intelligibility of synthetic speech was affected by adding noise to a much higher degree than that of human speech. Moreover, adding noise extended the range of the percentages correct, thus making the test more sensitive. So, if synthesis systems are compared which are rather similar, it might be a good idea to add noise.

The MRT has been employed, among others, by Logan, Pisoni & Greene (1985) to evaluate eight synthesisers and a human reference. On the basis of the results, the systems were grouped into four categories, namely (1) human voice, (2) high-quality DECtalk 1.8 Paul, DECtalk 1.8 Betty, MITalk-79, Prose 3.0, (3) moderate-quality Infovox SA 101, Berkeley, TSI-proto 1, and (4) low-quality Votrax Type'n'Talk, and Echo. Percentages correct are given in Figure 13.6 for the closed response variant.

Figure 13.6: *Some results for the MRT (Logan, Pisoni & Greene (1985))*

Speech type	Initial C	Final C	Overall
Human	99	99	99
High-quality	96	93	95
Moderate-quality	90	82	85
Low-quality	66	71	68

The categories distinguished could be used as benchmarks (although the data are somewhat dated, the set of synthesisers tested is probably representative of the quality range of more recent synthesisers). Methodological matters were considered as well. A test/retest design proved the MRT to be reliable. Moreover, the closed and open response variants (compared for the five best systems) yielded the same rank order.

SAM Standard Segmental Test

For diagnostic purposes the SAM Standard Segmental Test developed within the Speech Assessment Methods (SAM) project of ESPRIT (see A in Appendix 1) is to be preferred above the DRT and MRT tests. The test items in this test consist of meaningless and (sometimes by chance) meaningful, i.e. lexically unpredictable stimuli, which means that factors (1) and (2) as listed in section 13.4.1.1 have an effect on the responses. Items are CV, VC, and VCV stimuli, where C stands for all consonants allowed in the given position in a given language and V for one of the three point vowels of the given language, typically open /a/, close front /i/, and close back /u/. So, all permissible consonants are tested in word initial, word medial, and word final position. Vowels are not tested; they provide varying phonetic contexts within which the consonants to be tested are placed (the identifiability of sounds can vary depending upon neighbouring sounds). Examples of test items are *pa, ap, apa, ki, ik,* and *iki.* An open response format is used, i.e. listeners choose a response from among all consonants.

The SAM Standard Segmental Test has many positive points:

- It is relatively fast.
- It provides precise information on the characteristics of consonants as perceived by listeners, these can be compared to the intended characteristics, thus indicating in which direction improvement should be sought.
- It allows comparison of segmental intelligibility over languages.
- Software is available, allowing automatic stimulus generation and scoring (in fact, the stimulus set is really a subset of the CLID test (see B in Appendix 1), which allows for more vowel contexts and syllable structures).

The main disadvantage of the SAM Standard Segmental Test is that:

- It is restricted in that only single consonants are tested, whilst vowels are left out of consideration completely.

Part of the SAM Standard Segmental Test has been applied to English, German, Swedish, and Dutch synthesisers. Comparative results are available for Swedish medial C produced by a human and two synthesisers as perceived by listeners with perfect and imperfect hearing (Goldstein & Till (1992)). The percentages correct medial C identification are given in Figure 13.7. Of the 54 test items, 3 were found to differ significantly (p=2%) between human and KTH, 9 between human and Infovox, and 3 between KTH and Infovox.

Figure 13.7: *Some results for the SAM Standard Segmental Test (Goldstein & Till (1992))*

Speech type	Perfect hearing (N=24)	Imperfect hearing (N=14)
Human	94	91
KTH-synthesis	91	84
Infovox-synthesis	88	79

Diphone Test

A more complete overview of the performance of segments in a wider variety of contexts is provided by a test which assesses the intelligibility of all permissible (pronounceable) CVC, CVVC, VCV, and VCCV sequences of a given language. Such a test will be referred to as a Diphone Test, because the test items can be constructed by combining all the diphones in a diphone inventory. Just like in the SAM Standard Segmental Test, the test items are lexically unpredictable and the response categories open, so that it is useful for diagnostic purposes. Extra advantages of the Diphone Test over the SAM Standard Segmental Test are the following:

- The Diphone Test is less restricted than the SAM Standard Segmental Test because in addition to consonants vowels are tested,
- Both monosyllabic (CVC) and disyllabic (CV–VC, V–CV, VC–CV) structures are represented,
- Due to the VCCV items, heterosyllabic consonant clusters are included, i.e. sequences of consonants belonging to different syllables.

The main disadvantages of the Diphone Test are:

- It is not incorporated in a software package,
- It does not comprise tautosyllabic consonants, i.e. sequences of consonants within the same syllable.

The Diphone Test has been used to evaluate diphone synthesis in French (Pols et al. (1987)), Italian (Van Son et al. (1988)), and Dutch (Van Bezooijen & Pols (1987)). The Dutch Diphone Test combined all Dutch diphones into a set of 768 test items: 307 CVC, 173 VCV, 267 VCCV, and 21 CVVC. The only thing needed to construct the test material for a particular language is a matrix with the phonotactic constraints operating in that language, i.e. restrictions on the occurrence of all consonants and vowels in various word positions and phonetic contexts. Such matrices have been constructed for a number of European languages within the ESPRIT-SAM project.

Bellcore Test and CLID (CLuster IDentification) Test

As mentioned above, even the Diphone Test is not complete, since no tautosyllabic consonant clusters are included. The importance of this structure should not be underestimated. According to Spiegel et al. (1990), about 40% of all one-syllable words in English begin and 60% end with consonant clusters. The Bellcore Test (see C in Appendix 1) and the CLID Test (see B in Appendix 1), have been developed to fill this gap.

The Bellcore Test has a fixed set of CVC-stimuli, comprising both meaningless and meaningful words, e.g. *frimp* and *friend* or *glurch* and *parch*. Tautosyllabic consonant clusters and single clusters are tested separately in initial and final position. Vowels are not tested. Open response categories are used. Compared to the Diphone Test, the Bellcore Test has some advantages, the main being that:

- It includes tautosyllabic consonant clusters.

However, the Bellcore test has disadvantages as well:

- It is confined to monosyllabic structures.
- Vowels are not tested.
- No test material is available for other languages than English.

The test has been applied to assess the intelligibility of two synthesisers compared with human speech, presented over the telephone (Spiegel et al. (1990)). The syllable score was 88% for human telephone speech and around 70% for the synthetic telephone speech. Consonant clusters had lower intelligibility than single consonants. Intelligibility for meaningful words was higher than for meaningless words, a finding which could not be explained.

The CLID Test is a very flexible architecture which can be used for generating a wide variety of monosyllabic test items in an in principle unlimited number of languages. Both meaningful and meaningless items can be generated as long as matrices are available with the phonotactic constraints to be taken into account. Open response categories are used. Intelligibility can be assessed in whatever way one wants. The CLID test has been applied to test the intelligibility of one German synthesiser (Jekosch (1992)) and the intelligibility of five German synthesisers (Kraft & Portele (1995)).

The CLID test had all the advantages of the SAM Segmental Test listed above, whereas it does not have the disadvantage mentioned. Thus the positive points of the CLID test are the following:

- It is relatively fast.
- Clusters of consonants (tautosyllabic and heterosyllabic) and single consonants are tested as well as clusters of vowels and single vowels.
- It provides detailed diagnostic information.
- It allows comparison of segmental intelligibility over languages.
- Software is available, allowing automatic stimulus generation and scoring.

Although the CLID test allows testing of a wide variety of segments, the range covered is not complete. For example:

- Segments in unstressed syllables are missing.

(Modified) Minimal Pairs Intelligibility Test

The last tests we want to mention in this context, are the so-called Minimal Pairs Intelligibility Test (MPI Test), proposed by Van Santen (1993) as an alternative to the DRT, and a modification to it introduced by Syrdal & Sciacca (1994), the Diagnostic Pairs Sentence Intelligibility Evaluation Test (DPSIE Test). These tests were designed to reduce ceiling effects and expand the coverage of the DRT to include:

- Vowels,
- Tautosyllabic as well as heterosyllabic consonant clusters,
- Unstressed syllables,
- De-accented or cliticised words,
- Words in sentences,
- Polysyllabic words,
- Insertions and deletions.

The MPI Test consists of a fixed set of 256 sentence pairs containing one contrast, e.g. *The horrid courts scorch a revolution* versus *The horrid courts score a revolution.* The minimal pair appears on the screen and the correct sentence has to be identified. Differences between the MPI Test and the DSPIE Test include:

- A change of binary Jakobson, Fant & Halle (1951) distinctive features to multi-valued Ladefoged (1975) distinctive features, so that more realistic perceptual confusions can be tested.
- DPSIE generates its word pairs matched for familiarity (i.e. token frequency) and verb transitivity.

The main advantage of the MPI and DPSIE Tests is that:

- The intelligibility of segments is evaluated in a virtually complete variety of conditions and phonetic contexts, potentially providing a host of diagnostic information.

The main disadvantages of the tests are that:

- They are inefficient, since each response gives information on the identifiability of only one phoneme.
- Inclusion of segments in stimuli is restricted by their occurrence in the language.
- Creating test materials presupposes the availability of large databases.

Recommendation 28
Use the CLID Test for the evaluation of the segmental intelligibility at the word level, both for diagnostic and comparative purposes (in the latter case the stimulus set can be smaller).

Judgment tests at the word level

In principle, in addition to functional intelligibility tests, judgment tests, where subjects rate their subjective impression of the stimuli on scales, are possible for evaluating the segmental quality at the word level as well. For example, Van Bezooijen (1988), in addition to running a consonant cluster identification test, presented 26 Dutch consonant clusters (both initial and final) to be rated on naturalness, intelligibility, and pleasantness. The clusters were embedded in meaningful words. In order to obtain "pure" judgments, unaffected by the quality of the rest of the word, subjects were explicitly asked to pay attention to the clusters only. So, the test required analytic listening. However, one can never be sure to what extent listeners in fact stick to the instructions. Perhaps this is one of the reasons why judgment tests of this type have been rare.

13.5.2.1.2.2 Segmental tests at the sentence level

In addition to the word level, tests for the assessment of segmental quality have been developed at the sentence level as well. Here the effect of prosody could be minimised by presenting the material on a monotone, but in practice, if only for naturalness' sake, prosody is usually included. Compared with the segmental tests at the word level, tests at the sentence level are more similar to speech perception in normal communication but at the same time, as a consequence, less suitable for diagnostic purposes, because of the following reasons:

- With sentences, the intelligibility scores will not only be based on segmental quality but also to some extent on prosodic quality, so that poor intelligibility is more difficult to trace back to specific sources.
- The composition of the test material is somewhat unsystematic, so that no complete confusion matrices can be obtained.
- Especially with semantically normal sentences listeners will not only rely on segmental information but use other information sources as well, related to word combinatory probabilities, and, in the case of meaningful sentences, to semantic coherence (factors (4) and (5) in section 13.4.1.1).

Of course, if the test is not intended as a diagnostic tool but has a purely comparative aim, these consequences of using sentences do not necessarily detract from its value. However, it is important to remember that as soon as complete sentences are presented to listeners, the test is no longer limited to evaluating segmental quality alone. This means that the title of this section "segmental tests at the sentence level" is not completely adequate. In fact, depending on the extent to which restrictions are imposed on the construction of the test materials, tests at the sentence level are in between a glass box approach and a black box approach. So, the main differences among the segmental sentence tests described is their position on the glass box black box continuum.

In this section only functional tests will be discussed. In addition, judgment tests at the sentence level have frequently been carried out. These are described under the heading 'black box approach' in 13.4.1.2, where judgment tests to evaluate *overall output quality* are discussed. These tests entail the rating of scales such as *acceptability, intelligibility* and *naturalness.*

Harvard Psychoacoustic Sentences

One of the most well-known segmental intelligibility tests at the sentence level is the fixed set of 100 semantically and syntactically "normal" Harvard Psychoacoustic Sentences (*Add salt before you fry the egg*) (see H in Appendix 1). Intelligibility is expressed by means of the percent correctly identified keywords (nouns and verbs). In this test no restrictions are placed upon the composition of the test materials, which means that the percent correct responses is determined only to a limited extent by the acoustic characteristics of the individual segments. This test would therefore have to be placed towards the black box end of the continuum. In terms of the factors listed in section 13.4.1.1, only (6), (7) and (8) are excluded.

The main advantages of the Harvard Psychoacoustic Sentences Test are:

- It is easy to administer (no training required of the subjects).
- Responses are easy to score (be it manually).

The main disadvantages of the test are:

- Limited to comparative purposes.
- A strong learning effect (the same sentences can only be used once).
- Restricted number of test items.
- Restricted generalisability, because there is only one syntactic structure.
- Danger of a ceiling effect (so only fit to be used for low-quality systems).

The Harvard Psychoacoustic Sentences were compared with the Haskins sentences by Pisoni, Greene & Nusbaum (1985), Pisoni, Nusbaum & Greene (1985)) for four synthesisers and human speech (see below).

Haskins Syntactic Sentences

Another famous test at the sentence level is the fixed set of 100 Haskins Syntactic Sentences (see F in Appendix 1). These sentences are semantically unpredictable, which means that they do not occur in daily life. An example is *The old farm cost the blood.* In terms of advantages and disadvantages, the Harvard Sentences and Haskins Sentences have much in common. The only difference is that Haskins listeners can rely less on semantic coherence (factor (5) in the list of factors in section 13.4.1.1) so that the role of the acoustic characteristics of the segments is more important. Therefore, the Haskins sentences find themselves somewhat closer to the glass box end of the continuum than the Harvard sentences. The Haskins sentences were applied to four synthesisers and human speech by Pisoni, Greene & Nusbaum (1985), Pisoni, Nusbaum & Greene (1985), and compared with the Harvard sentences. Percent correct keyword identification are given in Figure 13.8.

Figure 13.8: *Some results for the Harvard Psychoacoustic Sentences and Haskins Syntactic Sentences*

Speech type	Harvard (meaningful)	Haskins (unpredictable)
Human	99	98
DEC Paul v1.8	95	87
MITalk-79	93	79
DEC Betty v1.8	90	75
Prose-2000 prototype	84	64

It can be seen that the two tests yield the same rankorder. However, as expected, due to the reduced semantic coherence, the Haskins sentences are more sensitive.

Semantically Unpredictable Sentences (SUS)

Both the Harvard and Haskins sentences had a fixed set of sentences, characterised by a single syntactic structure, which could be used as test materials. More recently, a more flexible approach was opted for in the Semantically Unpredictable Sentences (SUS), developed by SAM (see G in

Appendix 1). The test materials in the SUS consist of a fixed set of five syntactic structures which are common in most Western European languages, such as 'Subject–Verb–Adverbial' (*The table walked through the blue truth*). The lexical slots in these structures are filled with high-frequency words from language specific lexicons. The resulting stimulus sentences are semantically unpredictable, just like the Haskins Syntactic Sentences.

The advantages of the SUS Test are the following:

- Easy to administer (no training required of the subjects).
- Responses can be scored automatically.
- Reasonable generalisability due to the variation in syntactic structures and lexical items.
- Comparability across languages.

The main disadvantages of the test are:

- Limited to comparative purposes.
- Learning effects (despite the large number of test items).
- Risk of ceiling effects (so only fit to be used for low-quality systems).

Pilot studies with the SUS test have been run in French, German, and English (Benoît (1989); Benoît et al. (1989); Hazan & Grice (1989)). Results showed, among other things, that keywords presented in isolation were identified significantly less well than the same words in a sentence context. This is attributed in part to the fact that the syntactic category of the isolated words is not known. Furthermore, the SUS were found to be sensitive enough to discriminate between two synthesisers differing in prosody.

Recommendation 29
Use the SUS Test to evaluate intelligibility for comparative purposes at the sentence level.

13.5.2.2 Prosody

13.5.2.2.1 Functions of prosody

By prosody we mean the ensemble of properties of speech utterances that cannot be derived in a straightforward fashion from the identity of the vowel and consonant phonemes that are strung together in the linguistic representation underlying the speech utterance. Prosody would then comprise the melody of the speech, word and phrase boundaries, (word) stress, (sentence) accent, tempo, and changes in speaking rate. We exclude from the realm of prosody the class of voice characteristics (see section 13.5.2.3).

Prosodic features may be used to differentiate between otherwise identical words in a language (e.g. *trusty* – *trustee*, with initial stress versus final stress, respectively). Yet, word stress is not so much concerned with making lexical distinctions (this is what vowels and consonants are for) as with providing checks and bounds to the word recognition process. Hearing a stressed syllable in languages with more or less fixed stress informs the listener where a new word may begin; error responses in word recognition strongly tend to agree with the stimulus in terms of stress position. In a minority of the EU languages (Swedish, Norwegian) lexical tone (rather than stress) is exploited for the purpose of differentiating between segmentally identical words.

The more important functions of prosody, however, are located at the linguistic levels above the word:

- Prosody offers explicit segmentation cues in the form of phrase boundaries, i.e. it tells the listener which words go together and should be interpreted as making up a coherent chunk of information; also, these cues allow the listener to determine the "depth" of the break between chunks, i.e. whether the end of a word group, clause, sentence, or even a whole paragraph has been reached.
- Prosody provides an indication for the listener which words are presented by the speaker as expressing important information (highlighting or focussing through accentuation).
- Prosody, especially melody, carries some meaning of its own (intonational meaning) which, for example, allows the speaker to present a sentence as a statement or a question, or to express emotions and/or attitudes towards the verbal contents of the message or towards the hearer.

These functions suggest that prosody affects comprehension (establishing the semantic relationships between words) rather than intelligibility (determining the identity of words), and, indeed, this is what most functional tests of prosody aim to evaluate.

13.5.2.2.2 Judgment tests of prosody

Evaluation of the prosody of speech output systems is alternately focussed on the formal or the functional aspects. Only a handful of tests are directed at the formal quality of temporal organisation. An exemplary evaluation study on the duration rules of MITalk (Allen, Hunnicutt & Klatt (1987)) was done by Carlson, Granström & Klatt (1979). They generated six different versions of a set of sentences by including or excluding effects of consonant duration rules, vowel duration rules, and stressed syllable and preboundary lengthening rules in the synthesis. These versions were compared with a topline condition where (normalised) segment durations copied from human versions of the test sentences spoken by the MITalk designated talker were imposed on the synthesis. There were two baseline conditions, one with the neutral (inherent) table values substituted for all segments, and one with random segment duration variation (within realistic bounds). The results showed that the temporal organisation afforded by the complete rule set was judged as natural as the human topline control. Moreover, sentences generated with boundary markers at minor and major breaks were judged more natural than speech without boundary markers. [6]

More work has been done in the field of melodic structure. Let us first consider judgments of formal aspects of speech melody. The formal properties of, for example, pitch movements or complete speech melodies can be tested by asking groups of listeners (either naive or expert) to state their preference in pairwise comparisons or to rate a melody in a more absolute way along some goodness or naturalness scale. At the level of elementary pitch movements (such as accent-lending or boundary marking rises, falls, or rise-fall combinations) the SAM Prosodic Form Test (see I in Appendix 1) is a useful tool. The test was applied to two English and two Italian synthesisers, with 3 contours, 4 levels of segmental complexity, 5 items at each level, 4 repetitions of each token (Grice, Vagges & Hirst (1991)). Significant effects were found of synthesiser and contour, as well as of the interactions between synthesiser and contour and between synthesiser, complexity, and contour. By relating the scores for the contours to those for the monotone reference, the effect of differences in segmental quality on the ratings could be cancelled out.

[6] Later (cf. Allen, Hunnicutt & Klatt (1987)), the duration rules were evaluated directly (objectively) by comparing the predicted segment durations with the segment durations as measured in spectrograms of new paragraphs read by the designated speaker. The rules accounted for 84% of the duration variance with a residual standard deviation of 17 ms (excluding the prediction of pause duration). Seventeen ms is generally less than the just noticeable difference for a duration change in a single segment in a sentence context (Klatt (1976)), which would explain why the human reference and the rule-derived durations were judged equally natural.

Using the same methodology, i.e. rating and pairwise comparisons, the quality of synthetic speech melody can be evaluated at the higher linguistic levels. At the level of isolated sentences pairwise comparisons of competing intonation-by-rule modules is feasible when the number of systems (or versions) is limited (e.g. Akers & Lennig (1985)). When multiple modules are tested using a larger variety of sentences and melodies, scale rating is to be preferred over pairwise comparisons for reasons of efficiency (De Pijper (1983); Willems, Collier & Hart (1988)).

Evaluation of speech melody generators should not stop at the level of isolated sentences. Ratings by expert listeners in Dutch could not reveal any quality differences between synthetic melodies and a human reference when the sentences were listened to in isolation; however, the same synthetic melodies proved inferior to the human reference when they were presented in the context of their full paragraph (Terken & Collier (1989)). Along the same lines, Salza et al. (1993) evaluated the prosody of the Eloquens TTS for Italian in a train schedule consultation application. Three hundred sentences were tested, realistically distributed over seven melodic modalities: Command sentences, Simple declaratives, List sentences, Wh-questions, Yes/no-questions, Yes/no-echo questions, and Yes-no modal questions. Expert listeners' scores did not differ from those of naive subjects, and scores were better for utterances presented as part of a dialogue than for sentences presented in isolation. Clearly, both studies demonstrate that paragraph position or function within a dialogue induces certain perceptually and communicatively relevant adaptations to sentence prosody.

The form tests discussed so far address prosody globally. An analytic approach to prosodic evaluation using judgment testing was proposed by Bladon (1990) and co-workers. They developed an elaborate check list of formal properties that should be satisfied by any speech output system that claims to generate English melodies. Trained (but phonetically naive) judges listen to synthetic utterances, while looking at orthographic transcripts of the utterance with a crucial word or syllable underlined. Their task is to check whether the target syllable does in fact contain the melodic property prescribed by the check list. Although this idea is attractive from a diagnostic point of view (melodic flaws are immediately identified) the system has some drawbacks that should be considered before extending its use to other materials and other languages. First, drawing up a valid checklist presupposes a theory of intonation, or at least a detailed and valid description of the test sentences. Workable theories and descriptions may be provided for English, but will not be available for all (EU) languages. Second, even for English, the criteria for each melodic check point were formulated in rather crude terms, which makes it difficult for the judges to determine whether the utterance does or does not satisfy the criterion. Third, it is impossible to determine the overall quality of the melodies tested, since there is no way of combining the pass/fail scores for the various check points into a weighted overall score. A preliminary experiment revealed that three output systems could be meaningfully rank-ordered along a quality scale, but not at the interval measurement level. Systems that were clearly different as judged by experts were very close to each other in terms of their unweighed overall score, whereas systems that were rated as equally good by experts differed by many points. For these reasons, we do not recommend analytic judgments by naive listeners using a check list as an approach to evaluating prosody.

There is (at least) one judgment test that assesses how well certain communicative functions are signalled by prosody at a higher level. The SAM Prosodic Function Test[7] (see I in Appendix 1) asks for ratings of the communicative appropriateness of melodies in the context of plausible human-machine dialogue situations. The test was applied to human-machine dialogues designed to simulate a telephone enquiry service giving flight information (Grice, Vagges & Hirst (1992a)). A restricted set of contexts and illocutionary acts were included: *asking* (seeking information, seeking confirmation), *assertive* (conclude, put forward, state), *expressive* (greet), and *commissive* (offer, propose to). Two intonation versions were compared, one based on an orthographic input with

[7]The notion 'function test' in this sense has no relationship with our use of the term 'functional test'. In the SAM Prosodic Function Test prosodic quality is not being tested in a functional task: we are still dealing with intuitive judgments (ratings) of how well the melody would fulfil its function without actually testing it.

punctuation (target intonation algorithm) and the other based on a text input edited to conform to the type of text generated by an automatic language generator (reference intonation algorithm). The test should be seen as a first attempt to evaluate paralinguistic appropriateness of intonation in dialogue situations. For general comparative purposes, it would be useful to have an agreed-upon, systematic inventory of situations or speech acts one would want to include, taking as a point of departure, for example, the classification of speech acts proposed by Searle (1979).

Finally, we are not aware of tests asking subjects to judge the quality of the expression of emotions and attitudes in synthetic speech. It would appear that functional testing of these qualities is preferred in all cases.

13.5.2.2.3 Functional tests of prosody

Evaluating speech output prosody using functional tests is even more in its infancy. Since prosody is highly redundant given the segmental information (with the exception of the signalling of sentence type and emotion/attitude), it can be functionally tested only if measures are taken to reduce its redundancy. The first course of action, then, has been to concentrate on atypical, rather contrived materials in which prosody is non-redundant. That is, the materials consist of segmental structures that would be ambiguous without the prosody, and listeners are asked to solve the ambiguity. To the extent that the disambiguation is successful, the speech output system can be said to possess the appropriate prosodic functions. We find examples of such functional tests for the disambiguation of minimal stress pairs (for a survey, see Beckman (1986), word boundaries (for a survey, see Quené (1993), constituent structure (syntactic/prosodic bracketing; e.g. Lehiste, Olive & Streeter (1976), sentence type (e.g. Thorsen (1980)), and focus distribution (e.g. Nooteboom & Kruijt (1987)). However, in these kinds of study speech output assessment typically was not the primary research goal. Rather, speech synthesis was used here by psycholinguists or experimental phoneticians to manipulate the speech parameters in a controlled fashion.

The second route is to make prosody less redundant by degrading the segmental quality, such that without prosody (i.e. in the baseline conditions identified above) the intelligibility of the speech output would be extremely poor. The quality of the prosody would then be measured in terms of the gain in intelligibility, i.e. increase in percent correctly reported linguistic units (phonemes, morphemes, words) due to the addition of prosody. Carlson, Granström & Klatt (1979) measured intelligibility of utterances synthesised by MITalk with and without application of vowel duration, consonant duration and boundary marking rules (see above). They found that adding duration rules improved word intelligibility; adding within-sentence boundaries, however, did not boost intelligibility (even though the result was judged to be more natural, see above). Scharpff & Van Heuven (1988) demonstrate that adding within-sentence boundaries (i.e. changing the temporal organisation) does improve word intelligibility (especially for monosyllabic words) in Dutch diphone synthesis, and that utterances with pauses were judged as more pleasant to listen to (but only when listeners were unfamiliar with the contents of the sentence, (Van Heuven & Scharpff (1991)). Reasoning along the same lines, one would predict that quality differences in speech melody would have an effect on word recognition in segmentally degraded speech. Such effects were, in fact, reported by Maassen & Povel (1985), who used (highly abnormal) speech utterances produced by deaf speakers resynthesised with corrected temporal and/or melodic organisation.

There is a substantial literature on the perception of emotion and attitude in human speech (for a survey, see Murray & Arnott (1993)). Typically, listeners are asked to indicate which emotion they perceive in the stimulus utterance, in open or closed response format. Predictably, the larger the set of response alternatives, the poorer the identification of each emotion. It is not clear, in this context, how many different emotions should be distinguished, and to what extent these can be signalled by phonetic means. Still, results tend to show that the most basic emotions can be identified, in lexically neutral utterances, at better than 50% correct, in a 10 alternative closed response test. Synthesis of emotion in speech output is being attempted by several research groups. Preliminary evaluation of emotion-by-rule in Dutch diphone synthesis was presented by Vroomen, Collier & Mozziconacci (1993), as summarised in Figure 13.9:

Figure 13.9: *Percent correctly recognised emotions-by-rule in Dutch diphone synthesis (two diphone sets, obtained from different speakers) and in human speech (after Vroomen, Collier & Mozziconacci (1993))*

	neutral	joy	boredom	anger	sadness	fear	indignation
set 1	73	73	90	42	40	35	69
set 2	92	50	98	60	54	46	67
human	66	50	97	66	84	78	88

Note: The human reference condition were neutral utterances with temporal and melodic organisation copied from emotional utterances of the same sentences, using PSOLA.

13.5.2.3 Voice characteristics

13.5.2.3.1 Functions of voice characteristics

Whereas the segmental and prosodic features of speech are continuously varying, voice characteristics are taken to refer to aspects of speech which generally remain relatively constant over longer stretches of speech. Voice characteristics, also referred to as voice quality (cf. Laver (1991)), can most easily be viewed as the background against which segmental and prosodic variation is produced and perceived. In our definition, it includes such varied aspects of speech as mean pitch level, mean loudness, mean tempo, harshness, creak, whisper, tongue body orientation, dialect, accent, etc. Voice quality is mainly used by the listener to form a (sometimes incorrect) idea of the speaker's

- *Mood and personality* (cheerful, reliable, dominant),
- *Physical size* (tall, large, strong),
- *Sex* (male, female),
- *Age* (child, young adult, aged),
- *Regional background* (globally "from the North" or more precisely "from London, Paris, or New York"),
- *Socio-economic status* (high/low education),
- *Health* (cold),
- *Identity.*

In principle voice quality is not communicative, i.e. not consciously used by the speaker to make the listener aware of something which he was not previously aware of, but informative, which means that, regardless of the intention of the speaker, it is used by the listener to infer information. This information may have practical consequences for the continuation of the communicative interaction, since it may influence the listener's attitudes towards the speaker in a positive or negative sense and may affect his interpretation of the message (cf. Laver (1994)).

Since recently, increased attention is being paid to voice quality aspects of synthetic speech. In fact, Sorin (1994) regards the successful creation of personalised synthetic voices ("personalised TTS") as one of the most ambitious challenges of the near future. This aspect of synthesis is, for example, relevant in such applications as Translating (Interpreting) Telephony services, where along with translating the content of the message the original voice of the speaker has to be reconstructed (automatic voice conversion). Moreover, the correct encoding of speaker characteristics such as sex, age, and regional background is also relevant for the reading of novels for the blind. Finally, a third application is to be found in non-speaking disabled individuals, who have to use a synthetic speech to replace their own.

With a view of the latter application, Murray & Arnott (1993)) describe a system allowing rapid development of new voice "personalities" for the DECtalk synthesiser with immediate feedback to the user. Voice alteration is done by interpolating between the existing DECtalk voices (five male voices, five females voices, and a unisex child). Thus a voice may be created that sounds "a bit like Paul with a bit of Harry". A somewhat different approach aimed at a somewhat different type of application is described by Yarrington & Foulds (1993), who use original recordings of speakers who know they are going to lose their voice to construct speaker specific diphone sets.

13.5.2.3.2 Voice characteristics tests

Apart from specific requirements imposed by concrete applications, a general requirement of the voice quality of synthetic output is that it should not sound unacceptably unpleasant. *Voice pleasantness* is one of the scales included in the overall quality test proposed by the ITU-T to evaluate synthetic speech transmitted over the telephone (see L in Appendix 1). It has also been used by Van Bezooijen & Jongenburger (1993) in a field test to evaluate the functioning of an electronic newspaper for the blind. In this test, 24 visually handicapped rated the pleasantness of voice of two synthesisers on a 10-point scale (1: extremely bad, 10: extremely good). Ratings were collected at three points in time: (1) in a first confrontation with the synthesis output, (2) after one month, and (3) after two months of "reading" the newspaper. Interestingly, the pleasantness of voice ratings were found not to change over time, in contrast to the intelligibility ratings, which reflected a strong learning effect. From this it was concluded that voice quality has to be good right from the start; one cannot count on the beneficial effect of habituation. Both synthesis systems were generally considered good enough for the reading of popular scientific books and newspapers. However, partly due to the unpleasant voice quality, they were found unfit for the reading of novels or poetry (Jongenburger & Van Bezooijen (1992)). So, voice quality mainly seems to play a role when attention is directed to the form of the message, for recreational purposes. Finally, we hypothesise that perhaps more than for aspects of speech affecting comprehension, motivation and positive attitude might compensate for poor voice quality.

Of course, judgment studies such as these can only provide global information; if results are negative, no diagnostic information is available as to what voice quality component should be improved. There are no standard tests to diagnostically evaluate the voice quality characteristics of speech output. This type of information could in principle be obtained by means of a modular test, where various acoustic parameters affecting voice quality are systematically varied so that their effect on the evaluation of voice quality can be assessed. This would be the most direct approach.

A more indirect approach would involve asking subjects to listen analytically to and rate various aspects of voice quality on separate scales. A potentially useful instrument for obtaining a very detailed description is the Vocal Profile Analysis Protocol developed by Laver (1991). This protocol, which comprises more than 30 voice quality features, requires extensive training. If data are available for several synthesis outputs the descriptive voice quality ratings could be used to predict the overall pleasantness of voice ratings.

It may also be possible to use untrained listeners, although the number of aspects described will necessarily be more limited and less "phonetic". Experience with human speech samples representing various voice quality settings Van Bezooijen (1986) has shown that naive subjects can reliably describe 1-minute speech samples with respect to the following 14 voice quality scales: warm–sharp, smooth–rough, low–high, soft–loud, nasal–free of nasality, clear–dull, trembling–free of trembles, hoarse–free of hoarseness, full–thin, precise–slurred, fast–slow, accentuated–unaccentuated, expressive–flat, and fluent–halting. Again, if descriptive ratings of this type were available for synthetic speech they could be correlated with global ratings of synthesised voice quality. Alternatively, this type of scale could also be used more directly for diagnostic purposes, i.e. subjects could be asked to rate each of these voice quality aspects on a 10-point scale, with 1: extremely bad and 10: extremely good.

However, as mentioned above, experience with detailed perceptual descriptions of voice quality is as yet limited to non-distorted human speech. It remains to be assessed whether such descriptions can also be reliably made for synthetic speech. And even if this proved to be the case, the translation of the results obtained to actual system improvement is not unproblematic, since not much is known about the acoustic basis of perceptual ratings. Attempts in this direction have been rather disappointing (e.g. Boves (1984)).

In addition to judgment tests to evaluate the formal aspects of voice quality, functional tests may be used to assess the adequacy of voice quality. Although here also no standard tests are available, the procedures are rather straightforward and dictated directly by application requirements. One can think, for example, of tests in which subjects are asked, in an open or closed response format, to identify the speaker. This would be useful in an application where one tries to construct a synthetic voice for a given speaker or reconstruct the natural voice of a given speaker. Or one can ask people to identify the speaker's sex, or estimate his age or other characteristics.

In this context, accent and dialect features are relevant as well. For example, for Dutch a new set of diphones was derived from a western speaker, because some non-speaking users complained that the old diphone set had too much of a southern accent to be acceptable for communication in their living environment. To test whether naive listeners were in fact able to discriminate between the two diphone sets, listeners from different parts of the Netherlands rated CVC, VCV, and VCCV stimuli produced with the two systems on a 10-point bipolar *regional accent — standard Dutch* scale. The diphone sets were indeed clearly discriminable (Van Bezooijen (1988)).

Summarising it can be stated that very little experience has as yet been gained with the diagnostic and comparative evaluation of voice quality of speech output systems, either by means of judgment or functional tests. Moreover, except for specific applications where synthesis is closely connected with the identity of a speaker (in a clinical or automatic voice conversion setting), it is not even clear how much importance is attached to voice quality by naive listeners. How much does it people really bother when voice quality is unpleasant? For example, does an unpleasant voice quality prevent them from using a synthetic information service? We think it is too early to give concrete recommendations how to approach the evaluation of voice quality aspects of speech output. It is one of the topics for the near future.

13.5.2.4 Relationships among tests

Knowledge about the relationships among tests is important for at least two reasons:

- It allows a better interpretation of the meaning and validity of the test results obtained.
- It can be used to decide upon the test suite which gives a complete picture of all relevant aspects of speech output without being redundant; it is no use employing two tests which (to a large extent) yield the same information.

What would be needed to assess the relationships among tests is a large scale study which compares the performance of all "serious" tests testing the same aspect (e.g. intelligibility or comprehension) for a wide range of synthesisers. One would then like to know the stable differences among the tests in quality measured (e.g. percentage correct), as well as the correlations among the rank orderings of the synthesisers. In addition, it would be useful to have information on the reliability of "identical" tests developed and applied to a wide variety of different languages.

Some differences between the results obtained with different tests can be predicted to some extent. For example, when considering intelligibility, we think at least four factors will affect the outcomes: Intelligibility can be expected to increase

- as the unit of measurement is smaller (it is easier to identify one phoneme correctly than a sequence of phonemes),
- as the structure of the test items is more predictable (fixed versus open structure),
- as the combination of phonemes is more predictable (meaningful versus meaningless),
- as the number of response categories is smaller (closed versus open).

These predictions can be tested by looking at actual intelligibility results. Jekosch & Pols (1994), for example, assessed the intelligibility of one German synthesiser by means of four different tests (all described in Appendix 1):

- The SAM Standard Segmental Test (open response).
- A German variant of the MRT (closed response).
- The CLID test (open response).
- The SUS test (open response).

Percent correct elements (phonemes in the SAM Standard Segmental Test, clusters in the CLID test, words in the MRT and the SUS test) differed widely, from 19% to 85%. The lowest percentage was obtained for the SUS test, followed by the SAM Standard Segmental Test, the CLID test, and the MRT. The fact that the highest score was obtained with the MRT agrees with our predictions, since this test possesses not a single aspect with a negative effect on intelligibility: The unit of measurement is small (phoneme), the structure is fixed (CVC), the items are meaningful, and the response set is closed (six categories). The results for the other three tests point to complex interactions among the four factors.

Delogu, Paoloni & Sementina (1992) compared four different test methods for evaluating the overall quality of Semantically Unpredictable Sentences produced by a male speaker (once with and once without noise added), three synthesisers, and three vocoders:

- Categorical estimation (5-point scale from "excellent" to "bad").
- Magnitude estimation (quality expressed by any number between 1 and 100).
- Paired comparison ("which of the two realisations do you prefer?").
- Word monitoring ("push a button as soon as a specified target word is recognised").

Very high correlations were obtained among categorical estimation, magnitude estimation, and paired comparison (r>.90); somewhat lower but still high correlations were found between these three test methods and reaction time (r around .80). Reaction time showed the smallest variation in the responses, but the least discriminatory power. The best discrimination among the systems was obtained with paired comparisons.

Silverman, Basson & Levas (1990) compared the results of the Bellcore intelligibility test (see C in Appendix 1) with a comprehension test in which subjects had to answer questions related to the content of synthesised utterances with *yes, no*, or *can't tell from the information provided.* The faster subjects answered questions, the more items they heard. Two synthesisers, A and B, were tested. The intelligibility test yielded higher percentages correct for A than for B (77% versus 70%), whereas the comprehension test yielded higher percentages correct for B than for A (69% versus 63%). A few remarks are in order when attempting to interpret these seemingly contradictory results:

1. The presence of prosody in the comprehension test may have played a role.
2. Silverman et al. rightly state that the intelligibility scores relate to utterance-initial and utterance-final phonemes only, which hardly occur in sentence-length material (and running speech in general, for that matter).
3. It appeared that subjects generally heard fewer questions from B than from A in the test time allotted, a finding the meaning of which is complicated by the fact that B spoke at a slower rate than A.

Whatever the exact basis of the opposite rank orders yielded by the two tests, it is clear that caution should be exercised when generalising from a laboratory-type intelligibility test to a field-type, application-oriented comprehension test. Low correlations between intelligibility (MRT) and comprehension are also reported by Ralston et al. (1991).

In general, studies comparing different tests comprehend only a limited number of systems, which makes it difficult to learn to what extent the different tests rank the systems in the same way. Moreover, the relationship between the results yielded by glass-box and black-box tests deserves more systematic attention. We think that the importance of further studies of the relationships among tests cannot be stressed enough, if one wants to have a good idea of the meaning and generality of results obtained.

13.6 Further developments in speech output testing

13.6.1 Introduction

In this final section we will consider desired developments in the field of speech output testing, and discuss possibilities for further research on a more general level than we did in the preceding sections. The section consists of three parts. In the first part (section 13.6.2) we will be concerned with possibilities of producing more efficient output testing techniques. The general, longer-term strategy proposed here is to replace expensive, time-intensive tests (involving human listeners in field situations) by cheaper, automated tests carried out in a laboratory setting. In order to make this feasible we will have to establish the predictability relationships between the various types of tests discussed in section 13.2:

- How to predict human responses to output systems from automated (computer-simulated) listeners? (section 13.6.2.1)
- How to predict overall performance assessment from evaluation results of system modules? (section 13.6.2.2)

- How to predict functional performance of a system from judgment testing? (section 13.6.2.3)
- How to predict field performance of an output system from laboratory testing? (section 13.6.2.4)

Next, we will discuss (section 13.6.3) developments that we feel are needed in the assessment of linguistic interfaces of speech output systems. Finally, in section 13.6.4, we will propose research for the mid term aimed at improving speech output evaluation at the acoustic level in each of the four areas identified: segmental quality (section 13.6.4.1), prosodic quality (section 13.6.4.2), voice characteristics (section 13.6.4.3), and overall quality (section 13.6.4.4).

13.6.2 Long-term strategy: Towards predictive tests

13.6.2.1 From human to automated testing

The ultimate criterion to decide on the quality of speech output resides with the human listener. Speech output assessment is therefore basically a matter of human perception research. It is commonly acknowledged that the human listener is a noisy measurement instrument, which causes output assessment to be a slow and (therefore) expensive undertaking. There are generally felt to be two ways out of this problem. One is to look for assessment procedures which are optimally efficient, i.e. use perception tasks that are least susceptible to observer noise, and that concentrate on a small set of representative materials from which valid generalisations to all other situations can be made. This line of development has been followed for some time, especially by the SAM consortium, and could fruitfully be extended into the next five years.

The second way out is to replace the human observer by a computer-simulated observer, i.e. to use automated assessment methods. Using automated methods presupposes that we know exactly how human listeners react to speech output. The development of objective methods is therefore necessarily subsequent to the development of human test methods. In those areas of auditory perception where sufficient, consolidated knowledge has been assembled, attempts at computer-simulation can be launched even today, and, in fact, pilot studies have recently been undertaken that show the feasibility of objective testing in selected areas (see section 13.2.4). The field will have to reach agreement on what further aspects of human perception, relevant to speech output assessment, have evolved to the point that computer-simulation of the human listener can realistically be undertaken. Once such areas have been identified, the next step will be to go ahead and implement them.

Candidates that present themselves for automated testing will be:

- *Segmental quality*: Here the problems are not unlike those that have to be solved in automatic speech recognition. In both cases automatic procedures have to be developed that accept/recognise speech as long as the stimulus is within certain prespecified ranges of tolerance, i.e. gets close enough to some prespecified ideal. So, common to both enterprises is the element that we have to be able to determine the (perceptual) distance between input and ideal. However, in automatic speech recognition the tolerances will have to be set as loosely as possible, so as to maximise the chances of deviating exemplars of sounds/words to be recognised. In speech output assessment, ideally, the same perceptual distance metrics can be used. Generally, the closer the speech output hits the ideal norms, the better its quality. Perceptual distance assessment will then become equivalent to quality assessment. Moreover, the results can be used not only at the level of overall assessment (e.g. in terms of some weighed mean perceptual distance), but also for diagnostic purposes: when specific sounds are poorly synthesised, they will show up as local maxima in the plots of perceptual distance as a function of time.

As a first approximation, such computer-simulations should be tried for single speaker situations. That is to say, speech output should be compared only with human ideal speech produced by the same talker, pronouncing the same materials. Note that we assume that even allophone systems are based on a single model talker, since it is generally ill-advised to try and find average values over a larger group of speakers to control the synthesiser's parameters (Loman & Boves (1993: 159)).

- *Temporal structure*: Using the same principles as above, the discrepancies in temporal organisation between system output speech and a human ideal can be determined (e.g. through Dynamic Time Warping).
- *Melodic structure*: Using the same principles as above, perceptual distance between speech output melody and melodic norm can be determined.

  Note that since there will always be (slight) differences in timing between speech output and ideal speech, both segmental and melodic assessment will necessarily involve temporal normalisation. The perceptual evaluation of the discrepancies between output speech and ideal should therefore proceed in — at least — two separate stages: first the penalty that is incurred by deviating durations will have to be determined, and only then can we meaningfully consider the penalty for deviating segmental quality (likewise for melodic structure).

We advocate a two-pronged approach here. The field should concentrate on developing optimally efficient tests involving human listeners, and at the same time begin to work on the development of perceptual distance estimation procedures that can be used later in automated assessment.

13.6.2.2 Predicting functional behaviour from judgment testing

There is a paradox involved in the choice between judgment tasks and functional tests. On the one hand, it could well be argued that a speech output system is adequate if a representative user group judges the system to be adequate for its purpose. Why should the field go to more trouble to improve the system's quality if the users profess to be satisfied? On the other hand, we can predict with near certainty that the users will not be able to estimate precisely the level of adequacy needed for the output system to function smoothly in a concrete application. The relationship between judgments and functional test scores has been studied in the context of segmental quality, but so far similar studies in the field of prosodic quality testing are extant. It would seem a point of immediate concern, therefore, to consider research into the interrelationship between judgments and functional test behaviour, with emphasis on prosodic quality. To what extent do orderings among competing speech output systems, as derived from judgment tests, correspond to orderings derived from functional tests? If we should be able to predict functional test behaviour from judgment test scores, the latter, as a cheaper alternative for functional testing, could be used in all initial stages of speech output assessment. The use of functional testing would then typically be restricted to diagnostic testing.

13.6.2.3 Predicting global from analytic testing

Generally, one would expect the global quality of a speech output system to be a function of the quality of the various system components. One would like to be able to predict and quantify the overall ratings and global performance measures from the scores on the components through some form of regression analysis. Obviously, if system designers have only limited resources available, they would direct their efforts toward improving the quality of those aspects that contribute most (in terms of regression coefficients) to the overall assessment of his system. We suggest that research be undertaken in order to address this type of question.

13.6.2.4 Predicting field performance from laboratory testing

There is general agreement by now that laboratory tests such as are available today do not allow a useful prediction of how well a speech output system will perform in a concrete application. A short-term recommendation is, therefore, to develop a field-test generator, along the same lines as the successful test generators for laboratory intelligibility tests (such as the CLID and SUS tests developed by the SAM consortium). The field-test generator should enable the fast compilation of test materials and adequate simulation of a range of application conditions. For this purpose, an adequate cross-section of applications for speech output has to be inventoried and parametrised along such dimensions as (1) type of users (non-cooperative, children, elderly people, non-native language users), (2) specific aspects of the situation in terms of, for instance, noise, reverberation, telephone channel, and (3) secondary tasks.

On a longer-term basis we advocate a more fundamental solution to the problem of field testing. Ideally, of course, one should not have to go to field every time a new application presents itself. Rather, one would like to be able to predict accurately, on the basis of available results of standard laboratory tests (e.g. intelligibility scores and prosodic adequacy profiles) how a speech output system would perform in a concrete field situation. For this to be the case, it will be necessary to have a valid analysis of the field tasks that have to be accomplished. A task profile will have to be drawn up that analyses the demands that carrying out the task (including and excluding listening to speech output) makes on the user, such as attentional load of the primary task, environmental noise, negative influence of fatigue and boredom, physical strain, etc. Accomplishing this type of prediction calls for cooperation between experts in the field of speech quality assessment experts and human factors studies. We recommend exploratory studies along the lines suggested above, based on quantitative task analyses of a few selected applications.

13.6.3 Linguistic testing: Creating test environments for linguistic interfaces

Generally, we feel that the development and testing of the higher-order linguistic modules of speech output systems should be left to language technology experts. A reasonable division of work would be for speech technology to deal with the linguistic modules that are specific to TTS-applications, i.e. text preprocessing and grapheme-phoneme conversion (including stress position, accent placement and boundary marking). Other linguistic tasks, such as morphological analysis and syntactic parsing, are common to other branches of linguistic engineering (e.g. grammar checking, automatic translation), with much more resources and manpower available. However, even if this division of work could be effected, one would like to see attempts made towards early separation of consequential vs. inconsequential errors in word and sentence parsers. Consequential symbolic errors will audibly affect the (quality of the) acoustic output, whereas inconsequential errors are not reflected at the audio level. This means that part of speech output testing should still be concerned with the higher-order linguistics modules.

We would advocate a more detailed analysis of the various tasks a text-preprocessor has to perform, focussing on those classes of difficulties that crop up in any (European) language. Procedures should be devised that automatically extract representative items from large collections of recent text (newspapers) in each of the relevant error categories, so that multi-lingual tests can be set up efficiently. Once the test materials have been selected, the correct solutions to, for instance, expansion problems can be extracted from existing databases, or when missing there, will have to be entered manually.

A short-term recommendation is to develop multi-lingual machine-readable pronouncing dictionaries at the single word level which list permissible variations. Comparisons of algorithmic output with the model transcriptions requires the development of adequate string alignment procedures. Moreover, not all discrepancies found contribute equally to the overall evaluation. Distance metrics should be developed that allow us to express the differences between two segmentally different phonemic transcriptions in terms of meaningful perceptual distance. Recent work done by Cucchiarini (1993) could serve as a starting point.

The correctness of most symbolic output can only be determined on the basis of connected text at the sentence level. What is dearly needed, therefore, is the availability of large, multi-lingual text corpora with full phonemic annotation, including not only the permissible pronunciation(s) of the words, including the effects of assimilation across word boundaries and stress shifts, but also the indication of accent positions (and degrees of accent), prosodic boundaries (with break indices of various strengths), and some intonation transcription. Moreover, since these corpora will also have to be used for testing morphological and syntactic parsing, hierarchical word and sentence structure should be indicated; or at least provisions should be made for linguists to enter this type of information at a later stage. The development of such corpora, however, is beyond the competence of the speech output subgroup, and is best left to the text corpora subgroup. We refer to the relevant chapters on database development elsewhere in this volume for a report on what has recently been accomplished in this area, and what work has still to be done.

We recommend the development of procedures for strictly modular testing of linguistic interfaces. This means that test materials have to be made available that are specific to each individual module in the linguistic interface. Each module should be given correct input strings, and the correct output string(s) for only the module at hand should be provided. Only in this way can be eliminate the problem of percolating and compounding of errors made by earlier modules. Obviously, such procedures can only be effective if the databases referred to in the previous paragraph contain representations of the correct strings at each of the levels addressed by the various modules.

With the availability of cheap mass memory, the need for highly intricate linguistic interfaces is less strongly felt than some years before. Rather than computing the phonological code that is to be fed to the acoustic modules, the correct code is simply looked up in large lexicons included in the speech output system. If this trend continues, the emphasis of our research efforts will shift from rule development (and testing) to collecting databases. Database collection and annotation will take place regardless of the direction that the field takes in this matter. If choices have to be made, money is spent most safely on the development of corpora, but only if a multi-lingual notation format can be found that can be used for the transcription of segments and prosodic features of all languages dealt with.

Although less important at the isolated word level, it will remain necessary to test grapheme-phoneme conversion. The output of post-lexical rules (changing the pronunciation of words in connected speech, e.g. through assimilation). Also, testing grapheme-phoneme conversion will remain applicable in the development of cheap speech output systems (such as Multivox and Apollo), which do not access large lexicons not perform sophisticated linguistic analyses of the input text.

13.6.4 Acoustic testing: Developments for the near future

13.6.4.1 Segmental quality testing

With a few provisos (see below) there is general consensus that the procedures for testing segmental quality of speech output systems are more or less fully developed (cf. section 13.5.2.1.2 under DRT/MRT, CLID and SUS Tests). Under the auspices of the SAM consortium, efficient test generators have been developed that enable the construction of a large variety of tests that allow quick standardised administration and data analysis of consonant and vowel intelligibility scores, both for isolated word intelligibility and for intelligibility of words in (semantically unpredictable) context. These tools will be very useful in the testing of even the latest generation of parametric synthesisers. However, the upcoming generation of waveform synthesisers (PSOLA based) will have segmental quality that will be tough to discriminate from human speech. Though it may be possible to further refine the discriminatory power of our test procedures, one may well wonder what purpose would be served by such endeavors. A reasonable alternative view would be to consider the quality of waveform concatenation speech output equivalent to the human ideal (if indeed the test shows that no intelligibility difference remains) and leave the matter at that.

A short-term recommendation that should be made, concerns the quality of segments in unstressed syllables. It has rightfully been pointed out by, for instance, Van Santen (1993) that most segmental quality tests consider monosyllabic (or “minisyllabic”) words only. There is a risk involved here that insufficient attention is being paid to the quality of unstressed syllables in longer words. The same, of course, holds true of the quality assessment of (unstressable) function words. Unstressed syllables are generally reduced in human speech, and synthesis-by-rule systems have often neglected to carefully model the reduction processes. In concatenative synthesis, the problem of unstressed syllables can be solved by enlarging the set of acoustic normally unreduced building blocks with a parallel set of reduced building blocks (cf. Drullman & Collier (1993)). The testing problem that crops up in this connection presents an important perceptual question addressing the interaction between segmental and prosodic quality: if unstressed syllables are overarticulated, as would be the case when the reduction processes are not adequately modelled in our synthesis, does the resulting speech output get more intelligible, or does its intelligibility deteriorate? One might predict that, although the identifiability of each individual segment may decrease when reduction is truthfully mimicked, the overall intelligibility, in terms of word scores, will increase, reasoning that the rhythmic structure of words showing natural gradation of strong and weak syllables might be more important to word identifiability than optimal identifiability of each individual phoneme.

On a more general note, we suggest that serious attention be paid to differences in the contribution made to the overall intelligibility of words by the various constituent segments. It is important that we learn to what extent word intelligibility depends on identifying vowels versus consonants, in stressed versus unstressed syllables, in onset, medial, and final position, in short and longer words. Psycholinguistic studies on auditory word recognition have shown that, indeed, stressed segments — because of their greater inherent loudness and duration — have a better chance of contributing to the recognition process, as do segments early in the word. Ideally, we would like to be able to predict the intelligibility of an arbitrary selection of words from the lexicon of a language, just by looking at the identification scores of the constituent vowels and consonants in unpredictable words (i.e. segment strings that are phonotactically legal and may be lexical words or nonsense strings).

13.6.4.2 Prosodic quality testing

With the advent of high-quality segmental speech output (section 13.1.2) a shift from segmental quality testing to prosody seems immanent. It is obvious that there is still a long way ahead of us before the evaluation of prosody will get full coverage. What is needed is a careful taxonomy of prosodic functions at all linguistic and pragmatic levels. We suggest, therefore, that the first priority should be to chart out all the prosodic functions relevant to human-machine communication. We need to know not only what functions are fulfilled by prosody, but also what the communicative importance of each specific function is (if any). Once a reasonably complete view of relevant prosodic functions has been obtained, attempts should be made at defining adequate tests in order to determine to what extent each function is expressed by the speech output system.

It will be difficult to separate the evaluation of prosodic forms from their communicative functions, and perhaps such a dissociation is not even necessary. It seems reasonable to assume that a prosodic feature fulfills its communicative function better as its formal properties are closer to the human model. If this relationship holds, we would not have to test the formal adequacy of speech timing and melody rules in the abstraction of their communicative functions. Once we know the communicative function of each formal prosodic distinction, the prosodic quality of speech output systems can be measured by the effectiveness with which each of the communicative functions is signalled to the human listener. For these reasons we suggested that functional testing of prosody be given priority. Whatever audible flaws remain after the communicative functions have been shown to be signalled as effectively as in human speech, will have to be addressed in a later stage, using judgment tasks.

We propose that the emphasis should be on the functions of prosody, rather than on the details of prosodic form. Our point of departure, for the time being, is that the formal aspects of prosody cannot be too far off the mark if the prosodic functions are all adequately fulfilled. This should not be interpreted in the sense that we consider the details of prosodic form (such as exact pitch movements and timing) unimportant. In fact, there is every reason to believe that prosodic functions such as accentuation are only adequately expressed by very narrowly defined (in terms of direction, excursion size, and segmental alignment) language-specific pitch movements. In this context it seems obvious that adequate prosodic functioning can only be guaranteed if speech output systems are capable of synthesising not only accents and boundaries, but also more subtle degrees within such categories. For instance, the adequacy of prosodic boundary markings should be tested at least at four levels of depth: strong and weaker boundaries within the sentence, as well as sentence and paragraph boundaries, which are signalled in parallel by melody, temporal organisation, and (possibly even) intensity.

Generally, we believe that the identification of prosodic functions (including the expression of emotion) to be tested presents a greater problem than devising tests to determine the functional adequacy of prosody once a particular function has been identified. Still, choices will have to made as to what particular test methodology to adopt. We propose that a pilot study be initiated to examine the pros and cons of the various tests used in the experimental phonetic and psycholinguistic literature (as outlined in section 13.5.2.2.3) that seem relevant to this matter.

As a consequence of claiming priority for prosodic functions, the development of (multi-lingual) prosodic form tests (and test generators) should be postponed until some later stage.

13.6.4.3 Voice characteristics testing

It would appear that the evaluation of voice quality is going to be a matter of increasing concern. Developers of personalised voice speech output will need test procedures in order to determine how convincingly their systems mimic the quality of the model's voice. Simple same-different testing (*Is it Ella? Or is it Memorex?*) will not do, since developers will need the evaluation as a diagnostic tool. We suggest that a test tool be developed that enables the efficient drawing up of voice quality profiles (cf. section 13.5.2.3.2).

Apart from the development of personalised voice synthesis, the voice quality of general purpose speech output systems will get a lot more attention in the coming decade. With the improvement of segmental, and to a lesser extent, prosodic quality of speech output, the need for more natural and pleasant voice quality will be strongly felt. It will be a concern for the evaluation field to develop test procedures in order to determine the appropriateness of voice quality for speech output in general and for specific applications (e.g. alert messages).

13.6.4.4 Overall quality testing

Now that the quality of speech output systems gets closer to that of human speech, assessment should concentrate on other aspects of quality testing than linguistic functions. Synthetic speech may be virtually equivalent to human speech in all aspects, and still be lacking in certain subtle qualities. This aspect of speech output testing should be considered in a special study, looking at the effects of listening to synthetic speech in terms of fatigue and allocation of attention to secondary tasks (cf. section 13.4.1.1). The development of efficient multi-lingual test generators addressing this aspect would be a welcome addition to our arsenal.

13.7 List of recommendations

Recommendation 1
Use a glass box approach if you want diagnostics in order to improve your speech output system.

Recommendation 2
Use a black box approach if you want to assess the overall performance of speech output systems.

Recommendation 3
Do not rely on laboratory tests. As soon as there is a discrepancy between the laboratory setting and the true field situation (in terms of environment, tasks, type of listener) field testing is necessary.

Recommendation 4
Always exclude hearing-impaired subjects from speech output assessment. Within the SAM-project (Howard-Jones (1992a), Howard-Jones (1992b)) it is specified that subjects should pass the hearing screening test at 20 dB HL at all octave frequencies from 500 to 4000 Hz.

Recommendation 5
Never use the same subject more than once, unless, of course, one is interested in the effect of repeated exposure.

Recommendation 6
In diagnostic testing only include subjects speaking the same language (variety) as the language (variety) tested.

Recommendation 7
In diagnostic testing, hire a trained phonetician (with a basic understanding of the relationships between articulation and acoustics) in the initial stages of development of a system in order to obtain subtle information (e.g. degree of voicing in plosives), or information that is usually not used for functional purposes in real-life communication (e.g. formal aspects of temporal organisation and intonation, cf. Terken (1993).

Recommendation 8
In specialised applications, select subjects who are representative of the (prospective) users. For example, synthesis integrated in a reading machine for the blind should be tested with visually handicapped. And synthesis for long-term use should be tested with subjects with different degrees of experience and familiarisation with the type of synthetic speech of interest.

Recommendation 9
Synthesis to be used by the general public for incidental purposes, i.e. which should be functionally adequate in a first confrontation, should be tested with a wide variety of subjects, including people with a limited command of the language, dialect speakers, and people of different ages. However, none of them should have any experience in listening to synthetic speech. In telecommunications research, groups of between 12 and 16 subjects (all with English as their primary language) have been found sufficient to obtain stable mean values in judgment tests.

Recommendation 10
Use categorical estimation in rapid judgment tests with a view of internal comparison, and when you do, use at least a 10-point scale.

Recommendation 11
Use magnitude estimation for test external comparison, and when you do, use line length with imaginary ideal speech as a reference.

Recommendation 12
Absolute segmental topline: In the case of allophone synthesis, use human speech produced by a designated talker, i.e. the same individual on whose speech the table values and synthesis rules were based, or who, in the case of concatenative synthesis, provided the basic synthesis building blocks. The absolute topline reference will then be based on CD-quality digital speech.

Recommendation 13
Relative segmental topline for parametric synthesis: A second useful topline reference is the human reference speech (see Recommendation 12) but analysed and (re-)synthesised using exactly the same coding scheme that is employed in the speech output system to be tested.

Recommendation 14
Relative segmental topline for waveform concatenation: Use the same (lower) bitrate in the reference condition as in the speech output system.

Recommendation 15
Segmental baseline for allophone synthesis: Use speech in which all segments retain their table values and are strung together merely by smoothing spectral discontinuities at segment boundaries.

Recommendation 16
Segmental baseline for concatenative synthesis: Use speech made by stringing together coarticulatory neutral phones (i.e. stressed vowels spoken between two /s/-es, or stressed consonants preceded by schwa and followed by an unrounded central vowel, cf. the 'neutrone' condition in Van Bezooijen & Pols (1993)). Minimal smoothing should be applied to avoid spectral jumps.

Recommendation 17
Temporal and melodic topline: Copy, as accurately as possible within the limitations of the synthesiser, the temporal structures and speech melodies of a single designated professional human speaker onto the synthetic speech output.

Recommendation 18
Temporal baseline: Use a condition in which the smallest synthesis building blocks (phoneme, diphone, demisyllable) retain their original, unmanipulated durations as they were copied from the human original from which they were extracted (or, in the case of allophone synthesis, the phoneme duration table values, cf. Carlson, Granström & Klatt (1979)).

Recommendation 19
Melodic baselines: Synthesise utterances on a monotone, at a pitch level that coincides with the average pitch of the test items. Also, include a random melodic reference for the sake of validation, by introducing random pitch variations (in terms of excursion size, rate of change, and segmental alignment), within physiologically and linguistically reasonable limits and with a mean pitch equal to the average of the test items.

Recommendation 20
Use time-frequency warping of optimal human speech to create a grid of overall quality reference conditions.

Recommendation 21
Try to avoid the use of functional tests to assess overall output quality: on-line reaction time tests are difficult to interpret and off-line comprehension tests are difficult to develop.

Recommendation 22
If determined to develop a comprehension test, beware of the fact that reading tests may be too compact to be used as listening tests; adapt the materials or use materials that are meant to be listened to.

Recommendation 23
Use open comprehension questions rather than closed ones, the former being more sensitive than the latter.

Recommendation 24
When administering a comprehension test, include a top-line reference with a dedicated speaker realising exactly the same texts presented in a synthetic version; use different groups of subjects for the various speech conditions (or better still block conditions over listeners such that no listener hears more than one version of the same text while at the same time each listener gets an equal number of different text versions).

Recommendation 25
When interpreting comprehension results, look at difference scores (synthetic compared to human) rather than at absolute scores to abstract from the intrinsic difficulty of questions.

Recommendation 26
Since there is no consensus on the most appropriate judgment scales to evaluate overall quality, choose between:

- Intelligibility, naturalness, and acceptability (Sam Overall Quality test),
- Acceptance, overall impression, listening effort, and comprehension problems (ITU-T), or only listening effort (practice in telephony),
- Intelligibility, general quality, and naturalness (Van Bezooijen & Jongenburger (1993)).

Recommendation 27
It is important that the scale positions have a clear meaning for the subjects and that the scale is wide enough to allow differentiation among systems compared. Use at least a 10-point scale.

Recommendation 28
Use the CLID Test for the evaluation of the segmental intelligibility at the word level, both for diagnostic and comparative purposes (in the latter case the stimulus set can be smaller).

Recommendation 29
Use the SUS Test to evaluate intelligibility for comparative purposes at the sentence level.

13.8 Appendix 1: Summary of test descriptions

A limited number of tests, older as well as more recent ones, are described in detail below. Parts A through H pertain to intelligibility tests, ordered from tests based on meaningless items of a few phonemes to tests based on meaningful sentences, parts I and J describe two prosody tests, and finally parts K and L are related to two overall quality tests. Tests were included that have been frequently used, or were developed specifically, for evaluating synthesis output. Four tests (SAM Standard Segmental Test, Cluster Identification Test, Semantically Unpredictable Sentences, and the SAM Overall Quality Test) were developed within the ESPRIT-SAM (Speech Assessment Methods) project (1989–1992). Software for these tests is available in SOAP (Speech Output Assessment Package) (cf. Howard-Jones (1992b)). SOAP also includes programs to add noise to speech files, to automatically generate test sequences, to randomise test sequences, and to equalise speech levels. Test descriptions are given below.

A. SAM Standard Segmental Test

Status	Completely developed test with software (SOAP) for item construction, test administration, and data analysis (cf. Howard-Jones (1992b), Chapter 4).
Goal	Diagnostic and comparative evaluation of the intelligibility of single initial, medial, and final consonants.
Languages	In principle any language. The only thing needed are lists of permissible consonants in initial, medial, and final positions.
Items	Open vocabulary, mostly meaningless but also (by chance) meaningful, of the structure CV, VC, and VCV, comprising all phonotactically permissible combinations of initial, medial, and final consonants and three point vowels, e.g. /i/, /u/, and /a/. Examples: *pa, ap, apa* .
Procedure	Open response identification, preferably using unambiguous notation.
Time	With 20 consonants in a particular word position and 3 vowel contexts, a stimulus list contains 60 items. With an interstimulus interval of 4 sec this amounts to about 15 min for three lists. With two randomly ordered lists for each structure (recommended) time is doubled to 30 min per synthesiser.
Analysis	SOAP contains a pre-processing module to "clean" the responses. In addition it offers five scoring methods: (1) consonant confusion matrix, (2) percentage correct per consonant, (3) percentage correct per consonant per stimulus, (4) percentage correct per self-defined group of consonants, (5) group confusion matrix.

B. CLuster IDentification Test (CLID)

Status	Completely developed architecture with software (SOAP) for generating tests to evaluate intelligibility at the word level: item construction, test administration, and data analysis (cf. Jekosch (1992)).
Goal	Diagnostic and comparative evaluation of the intelligibility of sequences of one or more consonants ("consonant clusters") and sequences of one or more vowels ("vowel clusters").
Languages	In principle any language, as long as matrices with permissible phoneme and phoneme combinations are available.
Items	Mostly meaningless but also (by chance) meaningful, open vocabulary taking into account phonotactic constraints; combinatory matrices are available for English, French, German, and Italian. The frequency of the syllable structures (e.g. CCVCC, VCC) and clusters to be included in the test material can be defined. Examples: *storp* and *ast*.
Procedure	Open response identification using unambiguous notation or normal spelling.
Time	With 900 stimulus words (presented by Jekosch (1992)), about 2 hours per synthesiser. However, subsets of stimuli can also be defined.
Analysis	The SOAP scoring module allows, among other things, the calculation of percentages correct of all clusters, of initial, medial, and final clusters separately, and of complete items. No confusion matrices are generated.

C. The Bellcore Test

Status	Completely developed test.
Goal	Diagnostic and comparative evaluation of the intelligibility of sequences of one or more consonants in initial and final word position.
Languages	American English (cf. Spiegel et al. (1990)).
Items	Fixed set of 312 CVC-items, containing two tokens of nearly all onsets (N=62), nearly all plain offsets (N=63), and a subset of 31 (out of 111) offsets with phonetic affixes. Of each pair of tokens one is presented in a meaningful word, the other in a meaningless, but phonotactically legal word. Examples: meaningful versus meaningless onset *swan* versus *swog*, plain offset *warmth* versus *dorth*, offset with affix *dropped* versus *globbed*.
Procedure	Open response identification in normal spelling.
Time	With all 312 items and an interstimulus interval of 6 sec (recommended), about 40 min per synthesiser.
Analysis	Manually. Intelligibility is expressed as the percentage correct onsets and offsets or the percentage of cases where the phonemes occurring in the response correspond with a phoneme in the stimulus.

D. Diagnostic Rhyme Test (DRT)

Status	Completely developed test.
Goal	Diagnostic and comparative evaluation of the intelligibility of single initial consonants.
Languages	Originally English (cf. Voiers, Sharpley & Hehmsoth (1975), Voiers (1983)), adapted (at least) to Dutch (cf. Steeneken (1982)), and French (cf. Peckels & Rossi (1973)).
Items	Fixed set of 192 meaningful CVC-words. Examples: *dune* and *tune*. The Dutch version has 384 stimulus words from which different lists of 192 stimulus words can be composed.
Procedure	Closed response identification among two meaningful alternatives differing by a single phonetic feature in the initial consonant. Both alternatives are presented as a stimulus, so both *dune* and *tune* are presented with the response alternatives *dune* and *tune*. Six contrasts are represented, namely voicing, nasality, sustention, sibilation, graveness, and compactness. Each contrast is included 32 times in the test, combined with 8 different vowels.
Time	With an interstimulus interval of 3 sec, about 15 min per synthesiser.
Analysis	Intelligibility is expressed as the percentage correct initial consonants per contrast and for all initial consonants together. The percentages are corrected for chance.

E. Modified Rhyme Test (MRT)

Status	Completely developed test.
Goal	Diagnostic and comparative evaluation of the intelligibility of single initial and final consonants.
Languages	Originally American English (cf. House et al. (1965)), adapted to German (cf. Sotscheck (1982)).
Items	Fixed set of (originally) 300 meaningful CVC words, consisting of six lists, each with the first 25 items varying in initial consonant and the final 25 items varying in final consonant. Later variants (cf. Logan, Pisoni & Greene (1985)) had mixed instead of blocked presentation.
Procedure	Closed response identification among six meaningful alternatives, differing either in the initial or the final consonant. An example of a response set for a final consonant is *peas, peak, peal, peace, peach*, and *peat*. There is also a variant with open response identification among all meaningful CVC words known to the subject (cf. Logan, Pisoni & Greene (1985)).
Time	With an interstimulus interval of 4 sec, about 25 min per synthesiser.
Analysis	Intelligibility is expressed as the percentage correct initial and final consonants, or overall percentage correct consonants.

F. Haskins Syntactic Sentences

Status	Completely developed test
Goal	Comparative evaluation of sentence intelligibility, minimising the effect of contextual cues.
Languages	American English (cf. Nye & Gaitenby (1974)).
Items	Fixed set of 100 semantically unpredictable sentences of the form *The Adjective Noun1 Verb the Noun2* composed of high-frequency words. Example: *The old farm cost the blood.*
Procedure	Open response identification in normal spelling.
Time	With an interstimulus interval of 15 sec, about 30 min per synthesiser.
Analysis	Manually, intelligibility is expressed in terms of the percentage correct sentences or keywords.

G. SAM Semantically Unpredictable Sentences (SUS)

Status	Completely developed test with software (SOAP) for item construction, test administration, and data analysis (cf. Howard-Jones (1992b), Chapter 5).
Goal	Comparative evaluation of sentence intelligibility, minimising the effect of contextual cues.
Languages	Dutch, English, French, German, Italian, Swedish (can be extended).
Items	Short, semantically unpredictable sentences of five different, common syntactic structures with words randomly selected from lexicons with frequent, "mini-syllabic" words (smallest words available in a given category):

1. Subject – Verb – Adverbial, e.g. *The table walked through the blue truth*
2. Subject – Verb – Direct object, e.g. *The strong way drank the day*
3. Adverbial – Transitive verb – Direct object (imperative), e.g. *Never draw the house and the fact*
4. Q-word – Transitive verb – Subject – Direct object, e.g. *How does the day love the bright word?*
5. Subject – Verb – Complex direct object, e.g. *The place closed the fish that lived.*

Fifty sentences (10 per structure) are recommended per synthesiser.

Procedure	Open response identification in normal spelling.
Time	With 50 sentences and an interstimulus interval of 15 sec (recommended), about 15 min per synthesiser.
Analysis	Manually. Intelligibility is expressed in terms of the percentage correct sentences.

H. Harvard Psychoacoustic Sentences

Status	Completely developed test.
Goal	Comparative evaluation of sentence intelligibility.
Languages	American English (cf. Egan (1948)), French (cf. Combescure (1981)), and Dutch (cf. Plomp & Mimpen (1979)).
Items	Fixed set of 100 meaningful, syntactically varied, phonetically balanced sentences. Example: *Add salt before you fry the egg.*
Procedure	Open response identification in normal spelling.
Time	With an interstimulus interval of 15 sec, about 30 min per synthesiser.
Analysis	Manually. Intelligibility is expressed in terms of the percentage correct keywords.

I. SAM Prosodic Form Test

Status	Proposal (cf. Grice, Vagges & Hirst (1992b), Howard-Jones (1992b), Chapter 6).
Goal	Diagnostic and comparative evaluation of parameters used to characterise specific intonation contours (e.g. falling and rising slope), regardless of the context in which they appear.
Languages	English and Italian.
Items	Meaningful words of varying segmental complexity in terms of phonological vowel length and syllable structure, e.g. (1) *sketched, rain, moralise, salvageable* (English) (2) *re, basta, medico, nominano* (Italian). Several items are included for each level of complexity, with accurate stress placement. Items are placed in a short, neutral carrier phrase, e.g. *The word...*, or *Dico...* Each item should be generated with the contours the system can produce (e.g. those triggered by a period '.', question mark '?', exclamation mark '!', or comma ',') in addition to a reference monotone.
Procedure	Rating of naturalness by naive subjects using magnitude estimation (see **??**).
Time	With 3 contours, 4 levels of segmental complexity, 5 items in each level, 4 repetitions of each token and an interstimulus interval of 5 sec, about 30 min per synthesiser.
Analysis	Automatic calculation of the geometric mean of the responses per synthesiser, segmental complexity, and contour.

J. SAM Prosodic Function Test

Status	Proposal (cf. Grice, Vagges & Hirst (1992a), Howard-Jones (1992b), Chapter 6).
Goal	Diagnostic and comparative evaluation of the appropriateness of intonation contours for use in interactive communication contexts.
Languages	English, but can be applied without effort to other languages.
Items	Two-part, human-machine dialogue excerpts.
Procedure	Rating by naive subjects of the appropriateness of "pronunciation" using magnitude estimation (see **??**). Example: Human: *I'd like to reserve a flight to Paris on Monday morning .* Synthesiser: *Are you travelling from London?* There is a choice between two protocols: The transcript of the excerpt appears on the screen, then (type A) the text produced by the synthesiser is played out to be rated with respect to appropriateness or (type B) both the text produced by the human and the synthesiser is played out, the latter to be rated. It is recommended to present fifteen exchanges per algorithm and to include test materials based on a hand-annotated intonation version as a reference.
Time	Depending on the type of stimuli and the protocol.
Analysis	Automatic calculation of the geometric mean of the responses. Because of possible effects of differences in segmental quality between the dialogues on the ratings, it is advised to calculate the ratio between the text-to-speech version and the hand-annotated version.

K. SAM Overall Quality Test

Status	Completely developed software (in SOAP) allowing the use of the magnitude and categorical estimation scaling methods. Two variants are recommended: 20-point categorical estimation without reference (for test internal comparison) and magnitude estimation by line length, with imaginary ideal speech as a reference (for test external comparison) (cf. Howard-Jones (1992b), Chapter 7).
Goal	Comparative evaluation of overall quality aspects, particularly acceptability, intelligibility, and naturalness, for longer stretches of speech.
Languages	In principle applicable to any language as long as suitable stimulus material is available.
Items	Eight lists of 20 meaningful sentences of varying syntactic structures and length. For the rating of intelligibility and naturalness, speech material is available for Dutch, English, French, German, Italian, and Swedish. One list is sufficient for the evaluation of a synthesiser. Examples: *I realise you're having supply problems, but this is rather excessive* and *I need to arrive by 10.30 a.m. on Saturday.*
Procedure	Each aspect of speech is rated by a different group of subjects (minimally ten). When rating acceptability, it is recommended that application specific speech materials are presented to (prospective) users. The ratings are based on two sentences each time.
Time	With 160 sentences and a 5 sec interstimulus interval, the rating of one scale takes about 20 min.
Analysis	Automatic.

L. ITU-T Overall Quality Test

Status	Proposal (cf. ITU-T (1993)).
Goal	Comparative evaluation of overall quality aspects for longer stretches of speech.
Languages	In principle applicable to any language as long as suitable stimulus material is available.
Items	Speech samples of between 10 and 30 sec, adapted to the application. Example: *Miss Robert, the running shoes Adidas Edberg Pro Club, colour: white, size: 11, reference: 501-97-52, price: 319 francs, will be delivered to you in 1 week* (mail order shopping). It is recommended that a (degraded) human reference is included.
Procedure	Rating of (a subset of) eight categorical estimation scales, namely:

1. Acceptance (*Do you think that this voice could be used for such an information service by telephone?*) 1: yes, 2: no.

2. Overall impression (*How do you rate the quality of the sound of what you have just heard?*) 1: excellent, 2: good, 3: fair, 4: poor, 5: bad.

3. Listening effort (*How would you describe the effort you were required to make in order to understand the message?)* 1: complete relaxation possible, no effort required, 2: attention necessary, no appreciable effort required, 3: moderate effort required, 4: effort required, 5: no meaning understood with any feasible effort.

4. Comprehension problems (*Did you find certain words hard to understand?)* 1: never, 2: rarely, 3: occasionally, 4: often, 5: all of the time.

5. Articulation (*Were the sounds distinguishable*?) 1: yes, very clear, 2: yes, clear enough, 3: fairly clear, 4: no, not very clear, 5: no, not at all.

6. Pronunciation (*Did you notice any anomalies in pronunciation?)* 1: no, 2: yes, but not annoying, 3: yes, annoying, 4: yes, very annoying.

7. Speaking rate (*What do you think of the average speed of delivery?)* 1: much faster than preferred, 2: faster than preferred, 3: preferred, 4: slower than preferred, 5: much slower than preferred.

8. Voice pleasantness (*How would you describe the voice?)* 1: very pleasant, 2: pleasant, 3: fair, 4: unpleasant, 5: very unpleasant.

Subjects hear each test item twice. After the first presentation they answer questions related to the content of the test items (8 sec), after the second time they rate the scales (about 20 sec).

Time	With 4 test items per system, testing 4 synthesis systems with 3 reference conditions (i.e. 7 different sources) takes about one hour for one group of subjects, including instructions to subjects and training session.
Analysis	Histograms and/or cumulative distributions of the ratings per scale and mean ratings. Little attention is paid to the answers related to content.

13.9 Appendix 2: Evaluation of visual aspects of speech output

Chapter 14

Spoken language interactive systems

Chapter 15

Tools

15.1 Introduction

The development of speech technology is dependent upon the availability of data and tools operating in various computer environments. Speech as a sound is applied the digitisation to be computable, and therefore the speech research community pay tribute to the various sectors concerned with the digitisation of sounds. It has to deal with the requirements of its purpose and with the available current technology.

Among these sectors, Audio Industry takes over the Hi-Fi market with the development of Compact Disc technology and Digital Audio Tape. Computer Industry was interested first in audio output, then, with the development of multimedia capacities, both hardware and software (operating systems, I/O boards) were affected to handle progressively speech sampling then speech synthesis and recognition. Last but not least Telephone and Communication Networking play an important role with the transmission and coding of speech and sounds as their worldwide scope implies of course standard protocols.

Each of these sectors has its own requirements, and develop too many standards for hardware and software aspects. Digitised signal is characterised by many parameters which can lead to various encoding on communication channels and to numerous file formats when it is recorded on a computable medium.

15.2 Signal theory

When digitising signal, the following parameters have to be considered:

- The sampling rate (number of samples per second),
- The number of bits per sample (currently 8, or 16),
- The digitisation law (linear, log, U-law, A-Law, etc.),
- Number of channels (1 for mono, 2 for stereo, etc.).

These settings are those of the A/D converter when the data are recorded and must be used when they are played. They vary along the uses of different communities. Usually, the recorded signal is stored in files. When there are N channels, samples may respectively pertain to different files or sometimes to a unique one: in such a case samples or frames of samples are interleaved. The bytes order for each sample in the file (little or big endian) is also a relevant factor in reading the files (for example HI-LO order may be reversed from DOS to some Unix environment). Data compression whenever used, implies the inverse uncompressing scheme to be applied before using the data.

15.3 Speech research

Inside the speech research community itself, formats used for speech databases differ widely, depending on their purpose and the applications they are designed for. The major differences apply on the file format itself and on the sampling frequency.

Sampling frequency:

There is a variety of sampling frequencies used in existing speech databases. Here are a few examples:

Sampling rate	Speech databases
8000 Hz:	Telephone DB as POLYPHONE, CRICUBE (Canada)
10000 Hz:	Read texts corpus (Netherlands)
12000 Hz:	ATR DB (Japan)
12800 Hz:	Telephone DB as COLLECT (Italy)
16000 Hz:	ATR DB, ASJ (Japan), BDSONS (France), GRONINGEN (Netherlands), PHONDAT & ERBA (Germany), TIMIT (USA)
20000 Hz:	ATR DB (Japan), EUROM1 (Europe), BDBRUIT (France)
48000 Hz:	Jeida (Japan)

The higher the sampling rate is, the more space-consuming is the corresponding file, given the same amount of signal. Ten seconds of speech sampled at 10 KHz correspond to a 200 Kb filelength (with a 16 bits quantification), but to a 800 Kb filelength when sampled at 40 KHz. If usual frequencies tend to be higher, because of the high-quality technology being available and of the storage disk space becoming cheaper, many fields in speech research remain attached to middle frequencies for various reasons : because they don't need so high-quality for their purpose (speech synthesis), because they are linked to technology standards (European telephony standard: PCM A-Law (POLYPHONE)), or because their purpose is to deal with low-quality speech (speech recognition) for real applications.

File format:

It is very unlikely that getting a speech file one can guess what is its sampling rate, coding, conditions of recording and even less the age of the speaker. So it is of crucial evidence that information ABOUT the speech signal file must be somehow available in order to use it properly. The minimum requested information concerns of course the way to access the file (byte order, quantification, sampling rate). But information about the recording conditions, the speaker characteristics, the text of the utterances, and various parameters, is more than usefull in real speech studies. There are two main philosophies in force: storing information WITHIN the speech file or OUTSIDE the speech file i.e. in an external file. These two approaches have both pros and cons, and are well-represented respectively by the NIST/SPHERE format and the SAM format.

* NIST/SHERE: This format is provided by the National Institute for Standard and Technology in USA, and claims for a WITHIN approach using a SPHERE header. It consists of an "object-oriented, 1024-byte blocked, ASCII structure which is prepended to the waveform data. The header is composed of a fixed-format portion followed by an object-oriented variable portion." "The fixed portion is as follows: NIST_1A 1024 The first line specifies the header type and the second line specifies the header length." The remaining object-oriented variable portion is composed of object-type-value "triple" lines which have the following format: $< TRIPLE >::=< OBJECT >< space >< TYPE >< space >< VALUE >< OPT - SPACES >$ The currently defined objects cover database identification and version, utterance identification, channel count, samples

count, sampling rate, min and max level, and A/D settings. “The list may be expanded for future databases, since the grammar does not impose any limit on the number of objects. The file is simply a repository for ”standard” object definitions. The single object “end_head” marks the end of the active header and the remaining unused header space is undefined” (but within the 1024 bytes limit)”. (For further information see ftp site in the Appendices).

The NIST/SPHERE format is widely used in US and elsewhere, also for US & Dutch Polyphone and French BREF. It is supported by NIST, maintenance path exists, and it is provided with a set of tools to handle the header (access, update, remove, replace...). The header approach minimises risk of losing track of data identity; header can support both prompt and transliteration texts BUT requires data files to be changed after collection for annotation and also if an upgrade/correction is issued. Header is fixed-length and unpromptable through a text editor.

* SAM:
This format is a European 'standard', defined by SAM consortium (Esprit Project “SAM”: Speech Assessment & Methodologies). SAM claims for an OUTSIDE approach (headerless) using an associated description file. It consists of a SPEECH FILE + ASSOCIATED DESCRIPTION FILE a speech file contains only speech waveforms an associated description file (ascii) is linked to the speech file.

The files are going within a pair, and their names are identical, the extension last letter excepted, according to SAM terminology. The associated description file is a standard label file with a header an a body. It contains all information which can be usually needed by people working on the files without the database management system. A label file is made of a header and a body (or several). Each line consists of a specific mnemonic followed by the corresponding value:

<MNEMONIC> : <space><value> (optional <coma><space><value>...)

The label header: from LHD: ... to LBD: the label body: from LBD: to ELF: or a new labelbody LBD: ...

In a current annotation file the header contains database identification, file localisation, file production, A/D settings, sampling rate, start and end samples, number of channels, speaker information, and pointers to prompt textfile, recording conditions and protocol. As the format is potentially adapted to store several items in a file, the body contains on-the-field labels for the one or several items recorded in the speech file: sequence beginning (in sample), sequence end, input gain on recording, minimum sample value, maximum sample value, orthographic text promt are present for each item. Discontinuities between the items are indicated if any. Both the content of the header and of the body can be extended, to store new relevant descriptors or labels, provided that adequate mnemonics are created and no contradiction occurs with existing ones.

Further information can be obtained from ??(UCL ?, ELRA ?) The SAM format is widely used in EUROPE for multilingual databases (EUROM1) and for national ones (French BDSONS, English SCRIBE, Italian COLLECT, Spanish ALBAYZIN). The current SpeechDat consortium adopted SAM format for its telephone recordings. (SAM provided a conversion routine from NIST to SAM format on the DARPA/TIMIT CDROMs). The Associated description file implies to have files going together by pair and increases the risk of losing files. But he headerless system keeps data files unchanged after collection during database transcription correction/up- grade. It supports multiple annotation levels. Filelength is not limited, an information is available through a single text editor. The newcoming ELRA (European Language Resources Association) should take care of the maintenance/upgrades of this format.

* Other format The VERBMOBIL project in Germany has developed its own format, handling especially dialog. Examples of databases in Japan:(as JEIDA, ASJ) show no header.

Discussion

So far we have considered that a correct description of a sound data file includes a lot of mandatory fields. The first (and minimal) one contains information to use the file:

- file format: Header/no-header (and its length)
- number of significant bits, bytes order, sampling rate
- type of encoding : Linear (signed/unsigned), logarithmically, compression scheme ...
- channel number: mono/stereo, ...

But researchers in their job need a lot more about the speech they study:

- information about the corpus: name, prompt text
- information about the speaker: sex, age, accent, etc.
- information about recording conditions: quiet room, office, car, noise environment...

The development of speech applications in new domains imply many others descriptors being available. Description of new data types (multisensor, multimodal, dialogue) are needed, as well as more complex and complete descriptions of data(dialogue, e.g. in WOZ techniques; multimodal synchronisation; timing notations; additional descriptions as dialogue flow, emotional state, man-machine situation). In an other way the forthcoming development of databases distribution and networking will require information about the sources of the data to be available so as the way to obtain it and the rigth to use it.

The standardisation carried out in previous large collaborative projects must be clearly enhanced; efforts must be devoted to the representation of more complex information on speech data, with associated description files and pointers to various descriptors (including location of the data, source of the data, transformations applied to the sources, country of provenance acknowledgement, restrictions on use, derived information...).

15.4 Computer hardware and software

Current hardware uses a variety of in-house or more widely-spread standards. For example coding is 8 or 16 bits at Apple (Mac) and PCs, U-LAW 8 or 16 bits at Sun (Sparc), Next, VAX, DEC, U-LAW or A-LAW at HP. Available sampling rates are often limited to 8 KHz in the Unix world, but higher rates may be available in the world of PC (DOS/Windows) and Mac depending on the curent or professional I/O boards.

File formats are often indicated with the filename extension they bear. Computer manufacturers as Next and Sun deal with .au (AU) or .snd (SND)files, Apple and Silicon Graphics with .aif (AIF); I/O boards manufacturers may promote their own format (as .voc for SoundBlaster board) and the developper of the Windows operating system Microsoft try of course to impose its .wav (WAVE) format. This situation is complicated by the encoding mean (linear, compressed, data and information intermingled...) and even for the same filename extension, the implementation may vary a bit under different operating systems (WAVE in WINDOWS or UNIX environments, SND in Next or PC/MAC environments). A standardisation initiative comes trough the development of INTERNET to promote an interchange format called MIME.

A major example of the constraints imposed to the speech research community by the market can be demonstrated in looking at the implications of the multimedia standard development in the PC world.

Multimedia standard

The world of PCs has considerably evolved during the past few years along two relevant dimensions:

- Operating system: Windows operating system is now world-widely used, and it provides a suitable graphic interface.
- I/O boards: the development of Multimedia functionalities implied the availability of low-cost I/O boards to be easily included in a low-level PC configuration (as SoundBlaster, Pro Audio SPectrum...).

The point is now wether these current boards, primarily dedicated to audio output, can satisfy the needs of speech research and applications in term of:

- Quality of signal (signal/noise ratio...)
- Sampling frequency: the multimedia standard is basically derived from CD Audio standard (44.1KHz) or DAT one (48 KHz). So most of the multimedia compatible I/O boards use sampling rates that are provided through successive entire divisors of this basic frequency (22.5, 11.25 KHz etc...). But sampling rates used in our current speech databases are [actually] at present 16 KHz, 20 KHz... Care should be taken that a continuum of sampling frequencies could be available (lets say from 5 to 50 KHz) on these boards, to satisfy the requirements of the speech research community. It is foreseen that all current cheap boards will not be convenient. Otherwise, on-line re-sampling techniques would be requested (*) to maintain compatibility with existing databases, and for future databases the speech community is to adopt a standard "audio" sampling rate.
- File format: the multimedia standard apply the same. Most of the boards use "standard" (or peculiar) file format definition, for example the main one is WAVE format (.wav). It means that these boards are not able to play the files from our existing databases (SAM or national) which are in a 16 bit linear format, as the WAVE one consists of chunks of data intermingled with chunks of encoding info. The files of these databases would have to be converted (**) on output from one format to the other in order to be played. Future databases should either adopt a new "market" standard, or have their files converted on input and output.
- Number of channels available (2 or more channels may be requested for various microphones or sensor data recordings).

(*)(**) Using I/O boards without DSP implies that some signal processing will be deported to the PC (speech level detection, min/max measurement, eventual over- or under-sampling). These on-line procedures, augmented with on-line format conversion routines, could increase the CPU load in such a way that low-level SESAM workstation could not be able to support running with a high speech sampling rate for example (or using two channels).

One topic is background compatibility with existing databases, an other one is WHICH format is going to be THE STANDARD, i.e. the worldwide audio/ computer/speech standard. Such a topic is to be considered during the SPEECHDAT project, but it is foreseen that no unique standard will emerge and that conversion routines will remain a big issue. A lot of tools are available but as an example, even for the RIFF WAVE format the conversion between Windows and Unix worlds is all but trivial. At the moment, it is not sure whether the current inter-changeable standard I/O boards in the market will satisfactorily meet the speech research needs or not depending from the target application.

15.5 Conclusion

At the time of a huge involvement in the development of linguistic resources, of their wide dissemination and re-usability, the problem of the standardisation of tools and data is especially relevant in the speech domain. The useability of speech databases and the portability of speech tools is very dependent upon the functionnalities provided or imposed by hardware and software manufacturers. Audio industry particularly (with the high sampling frequency used in DAT, CD) and computer manufacturers (with their devoted I/O boards or sound systems formats (WAV, AU, SND files...)) don't make things easier. Although some powerfull conversion tools do exist (as "sox" or "sfconvert"), there is a general lack of reliable and easy-to-use conversion routines. A close relationship between industrials and public research organisations should take care of the bridges among their respective standards. If no agreement about a common standard can be aimed at, at least import and export drivers for conversion to the different formats must be handled.

15.6 Appendix: Useful anonymous ftp sites

AUDIO FILES FORMATS:

- ftp.cwi.nl:/pub/audio (part 1 and part 2),or
- rtfm.mit.edu:/pub/usenet/news.answers audio-fmts/part[12]

FILE CONVERSIONS (SOX: "Sound Exchange")

- ftp.cwi.nl:/pub/audio/SOX¡version¿.tar.Z (Unix) and sox5dos.zip (PC)

NIST-SPHERE

- jaguar.ncsl.nist.gov in compressed tar form as "sphere-v.tar.Z"
- ftp.cwi.nl:/pub/audio/NIST-SPHERE (short description)

RIFF WAVE (.WAV)

- ftp.cwi.nl:/pub/audio/RIFF-format
- ftp.microsoft.com:/MSDN/CD8/RIFFNE.ZIP (complete description)
- teeri.ouli.fi:/pub/msdos/programming/*
- ftp.ircam.dr:/pub/music

References

Abercrombie, D. (1967) *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Aho, A. V., Kernighan, B. W. & Weinberger, P. J. (1987) *The AWK programming language*. Reading, Mass., etc.: Addison-Wesley Publishing Company.

Ainsworth, W. A., ed. (1992) *Advances in speech, hearing and language processing: A research annual*. Volume 2. London: Jai Pr..

Aitchison, J. (1994) *Words in the mind. An introduction to the mental lexicon*. Oxford: Blackwell.

Akers, G. & Lennig, M. (1985) Intonation in text-to-speech synthesis: Evaluation of algorithms. *Journal of the Acoustical Society of America, JASA* 77:2157–2165.

Akmajian, A. (1984) Linguistics: An introduction to language and communication. 2nd edition. Cambridge, Massachusetts: The MIT Press.

Allen, G. D. (1988) The PHONASCII system. *Journal of the International Phonetic Association* 18(1):9–25.

Allen, J., Hunnicutt, M. S. & Klatt, D. H. (1987) *From text to speech: The MITalk system*. Cambridge: Cambridge University Press.

Allen, J. F. (1991) Discourse structures in the TRAINS project. Proc. 4th DARPA Speech and Natural Language Workshop.

Alleva, F., Hon, H., Huang, X., Hwang, M., Rosenfeld, R. & Weide, R. (1992) Applying SPHINX-II to the DARPA Wall Street Journal CSR task. Speech and Natural Language Workshop, Harriman, New York. 393–398.

Andernach, T., Deville, G. & Mortier, L. (1993) The design of a real world Wizard of Oz experiment for a speech driven telephone directory information system . Proceedings of the 3rd European Conference on speech communication and technology, Berlin, September . 1165–1168.

Andry, F., Bilange, E., Charpentier, F., Choukri, K., Ponamali, M. & Soudoplatoff, S. (1990) Computerised simulation tools for the design of an oral dialogue system. Proceedings of the ESPRIT Technical Conference, Brussels, November.

Andry, F., McGlashan, S., Youd, N. J., Fraser, N. M. & Thornton, S. (1992) Making DATR work for speech: Lexicon compilation in SUNDIAL. *Computational Linguistics* 18(3):245–267.

Argente, J. A. (1991) From speech to speaking styles. *Proceedings of the ESCA Workshop 'Phonetics and phonology of speaking styles: Reduction and elaboration in speech communication', Barcelona* 1-1:1–12.

Aubergé, V. (1992) Developing a structured lexicon for synthesis of prosody. In: Bailly, G., Benoît, C. & Sawallis, T. R. (eds.) *Talking machines: Theories, models and designs*. Amsterdam: North-Holland. 307–321.

Austin, J. L. (1962) *How to do things with words*. Oxford: Oxford University Press.

Autesserre, D., Pérennou, G. & Rossi, M. (1989) Methodology for the transcription and labeling of a speech corpus. *Journal of the International Phonetic Association* 19(1):2–15.

Averbuch, A., Bahl, L. & Bakis, R. (1987) Experiments with the TANGORA 20000 word speech recognizer. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP . 701–704.

Averbuch, A., Bahl, L. R., Bakis, R., Brown, P., Cole, A., Daggett, G., Das, S., Davies, K., De Gennaro, S., De Souza, P. V., Epstein, E., Fraleigh, D., Jelinek, F., Katz, S., Lewis, B., Mercer, R., Nadas, A., Nahamoo, D., Picheny, M., Shichman, G. & Spinelli, P. (1986) An IBM PC-based large-vocabulary isolated-utterance speech recognizer. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP . 53–56.

BMDP (1990) BMDP Statistical Software manual, Vol. 2.

Baayen, H. (1991) De CELEX lexicale databank. *Forum der Letteren* 32(3):221–231.

Bach, E. & Harms, R., eds. (1968) *Universals in linguistic theory*. New York: Holt, Rinehart and Winston.

Bahl, L. R., Brown, P. F., De Souza, P. V. & Mercer, R. L. (1989) A tree-based statistical language model for natural language speech recognition . Proc. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-37(7) . 1001–1008. Also in: A. Waibel, K.-F. Lee, eds. (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 507–514 .

Bailleul, C. (1987) Evaluation des performances d'un systhème de reconnaissance vocale dans des Tâches de Contrôle airiens. Note Interne, CENA/N87083, 22 June.

Bailly, G. (1994) Rule compilers and text-to-speech systems. *Les Cahiers de l'ICP* 3:87–91.

Bailly, G. & Benoît, C., eds. (1992) *Talking machines: Theories, models and designs*. Amsterdam: North-Holland, Elsevier Science Publishers.

Baker, J. (1975a) Stochastic modeling for automatic speech understanding. In: Reddy, D. (ed.) *Speech recognition*. New York, N.Y.: Academic Press. 521–541. Also in: A. Waibel, K.-F. Lee, eds. (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 297–307.

Baker, J. (1975b) The DRAGON system — An overview. IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-23. 24–29.

Baker, J. (1989) DragonDictate-30k: Natural language speech recognition with 30000 words. *Proc. European Conference on Speech Technology* 2:161–163.

Baker, J., Bamberg, P., Bishop, K., Gillick, L., Helman, V., Huang, Z., Ito, Y., Lowe, S., Peskin, B., Roth, R. & Scattone, F. (1992) Large vocabulary recognition of Wall Street Journal sentences at Dragon systems. Speech and Natural Language Workshop, Harriman, New York, 23–26 February. 387–392.

Ball, M. J. (1991) Computer coding of the IPA: Extensions to the IPA. *Journal of the International Phonetic Association* 21(1):36–41.

Barber, S., Carlson, R., Cosi, P., Benedetto, M. G. D., Granström, B. & Vagges, K. (1989) A rule-based Italian text-to-speech system. *Proc. Eurospeech '89, Paris* 2:517–520.

Barry, W. J. & Fourcin, A. J. (1992) Levels of labelling. *Computer Speech and Language* 6:1–14.

Barry, W. J., Grice, M., Hazan, V. & Fourcin, A. J. (1989) Excitation distributions for synthesised speech. *Proc. Eurospeech '89, Paris* 1:353–356.

Bates, M. & Ayuso, D. (1991) A proposal for incremental dialogue evaluation. Proceedings of the DARPA Workshop on Speech and Natural Language, Pacific Grove, CA, February. 319–322.

Bates, M., Boisen, S. & Makhoul, J. (1990) Developing an evaluation methodology for spoken language systems. Proceedings of the DARPA Workshop on Speech and Natural Language, Hidden Valley, PA, June. 102–108.

Baum, F. (1900) *The wizard of Oz.* London: Collins. New edition of 1974.

Beckman, M. E. (1986) *Stress and non-stress accent.* Dordrecht: Foris.

Belina, F. & Hogrefe, D. (1988) The CCITT specification and design language SDL. *Computer networks and ISDN systems* 16:311–341.

Bennis, H. & Van Lessen-Kloecke, W. U. S., eds. (1984) *Linguistics in The Netherlands 1984.* Dordrecht: Foris.

Benoît, C. (1989) Intelligibility test for the assessment of French synthesizers using semantically unpredictable sentences. Proc. of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, 1.7.1–1.7.4.

Benoît, C. (1991) On the assessment of audio-visual speech synthesis. Proc. Workshop on International Cooperation and Standardisations of Speech Databases and Speech I/O Assessment Methods, Chiavari, Italy.

Benoît, C., Lallouache, T., Mohamadi, T. & Abry, C. (1992) A set of French visemes for visual speech synthesis. In: Bailly, G. & Benoît, C. (eds.) *Talking machines: Theories, models, and design.* Amsterdam: North Holland, Elsevier Science Publishers. 485–504.

Benoît, C., Van Erp, A., Grice, M., Hazan, V. & Jekosch, U. (1989) Multilingual synthesizer assessment using semantically unpredictable sentences. *Proc. Eurospeech '89, Paris* 2:633–636.

Bentler, P. M. (1985) *Theory and implemetation of EQS, a structural equations program.* Los Angeles: BMDP Statistical Software Inc..

Berendsen, E., Langeweg, S. & Van Leeuwen, H. C. (1986) Computational phonology: Merged not mixed. Proc. International Conference on Computational Linguistics '86. 612–614.

Beukema, F. & Hulk, A., eds. (1986) *Linguistics in the Netherlands 1986.* Dordrecht: Foris.

Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, R., Santorini, B. & Strazalkowski, T. (1991) A procedure for quantitatively comparing the syntactic coverage of English grammars. Proceedings of the DARPA Workshop on Speech and Natural Language, Pacific Grove, CA, February. 306–311.

Bladon, A. (1990) Evaluating the prosody of text-to-speech synthesizers. Proc. Speech Tech '90. 215–220.

Bleiching, D. (1992) Prosodisches Wissen im Lexikon. In: Görz, G. (ed.) *KONVENS 92, 1. Konferenz "Verarbeitung natürlicher Sprache", Nürnberg, 7–9 October 1992.* Berlin: Springer-Verlag. 59–68.

Bleiching, D. & Gibbon, D. (1994) Handbuch zur Demonstrator-Wortliste. V1.1. May 1994. Bielefeld, Germany: Bielefeld University.

Bobrow, D. G. & Winograd, T. (1977) An Overview of KRL, a knowledge representation language. *Cognitive Science* 1:3–46.

Boguraev, B. K., Carroll, J., Pulman, S., Russell, G., Ritchie, G. D., Black, A., Briscoe, E. J. & Grover, C. (1988) The lexical component of a natural language toolkit. In: Walker, D., Zampolli, A. & Calzolari, N. (eds.) *Automating the lexicon: Research and practice in a multilingual environment*. Cambridge: Cambridge University Press.

Bolinger, D. (1972) Accent is predictable (if you're a mind-reader). *Language* 48:633–644.

Boogaart, T. & Silverman, K. (1992) Evaluating the overall comprehensibility of speech synthesizers. Proc. 2nd International Conference on Spoken Language Processing, ICSLP, Banff. 1207–1210.

Boogart, T., Van Alphen, P. & Doll, J. (1993) Application oriented assessment of dialogue systems. Joint ESCA — NATO/RSG10 Tutorial and Research Workshop on Applications of Speech Technology, Lautrach, September.

Boves, L. (1984) *The phonetic basis of perceptual ratings of running speech*. Dordrecht: Foris.

Brachman, R. J. & Levesque, H. (1985) *Readings in knowledge representation*. Los Altos, California: Morgan Kaufmann Publishers, Inc..

Branigan, G. (1979) Some reasons why successive single word utterances are not. *Journal of Child Language* 6:411–421.

Brietzmann, A., Hein, H., Niemann, H. & Regel, P. (1983) The Erlangen system for understanding continuous German speech. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Boston. 304–307.

Brouwer, D. & Haan, D. D., eds. (1987) *Woman's language, socialization and self-image*. Dordrecht: Foris Publications.

Browman, C. P. (1980) Rules for demisyllable synthesis using Lingua, a language interpreter. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Denver. 561–564.

Bruce, G. (1989) Report from the IPA Working Group on suprasegmental categories. Working Papers 35, Lund University, Department of Linguistics, Lund. 25–40.

Bunt, H. C. & al. (1985) The TENDUM dialogue system and its theoretical basis. *IPO Annual Progress Report* 19:105–113.

Burrell, M. D. (1991) Assessment of the degradations of synthetic speech and Time Frequency Warping over different listening levels. *Proc. Institute of Acoustics* 13, Pt. 2.

Button, G. (1990) Going up a blind alley: Conflating conversation analysis and computational modelling. In: Luff, P., Gilbert, G. N. & Frohlich, D. M. (eds.) *Computers and conversation*. London: Academic Press. 67–90.

Button, G., ed. (1992) *Technology in Working Order: Studies of work, interaction and technology*. London: Routeledge.

CCITT (1988) Objective measurement of active speech level. Rec. P. 56 Melbourne.

Cahill, L. J. (1993) Morphonology in the lexicon. Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics, Utrecht. 87–96.

Cahill, L. J. & Evans, R. (1990) An application of DATR: The TIC lexicon. In: Evans, R. & Gazdar, G. (eds.) *The DATR Papers*. 2nd edition. Brighton: School of Cognitive and Computing Science, University of Sussex. 31–39.

Carbonell, N. & Pierrel, J. M. (1986) Architecture and knowledge sources in a human computer oral dialog system. Proc. of NATO workshop: Structure of multimodal dialogues including voice, Corsica, France.

Carlson, R., Granström, B. & Klatt, D. H. (1979) Some notes on the perception of temporal patterns in speech. Proc. 9th International Congress of Phonetics Sciences, Copenhagen, vol. 2. 260–267.

Carre, R. (1992) Speech databases. In: Ainsworth, W. A. (ed.) *Advances in speech, hearing and language processing: A research annual*. Vol. 2. London: Jai Pr.. 199–215.

Carson, R. (1984) A digital speech voltmeter — the SV6. *British Telecommunications Engineering* 3:23–30.

Carson-Berndsen, J. (1993) Time map phonology and the projection problem in spoken language recognition. Doctoral dissertation. Bielefeld, Germany: University of Bielefeld.

Cartier, M., Emerald, F., Pascal, D., Combescure, P. & Soubigou, A. (1992) Une méthode d'évaluation multicritère de sorties vocales: Application au test de 4 systèmes de synthèse à partir du texte. 19èmes Journées d'Étude sur la Parole, Brussels.

Ceri, S., Gottlob, G. & Tanca, L. (1990) *Logic programming and databases*. Berlin: Springer Verlag.

Chafe, W. (1992) The importance of corpus linguistics to understanding the nature of language. In: Svartvik, J. (ed.) *Directions in corpus linguistics: Proceedings of the Nobel Symposium 82, New York* . Berlin: Mouton de Gruyter. 79–97.

Charniak, E. & McDermott, D. (1985) *Introduction to Artificial Intelligence*. Reading, Massachusetts: Addison-Wesley.

Chomsky, N. (1965) *Aspects of the theory of syntax*. Cambridge, MA: The MIT Press.

Chomsky, N. & Halle, M. (1968) *The sound pattern of English*. New York, Evanston, London: Harper and Row.

Church, K. W. (1987a) *Phonological parsing in speech recognition*. Boston, Dordrecht, Lancaster: Kluwer Academic Publishers.

Church, K. W. (1987b) Phonological parsing and lexical retrieval. *Cognition* 25:53–69.

Coates, J. (1986) *Women, men and language: A sociolinguistic account of sex differences in language*. London: Longman.

Cole (1995) The challenge of spoken language systems: Research directions for the nineties. *IEEE Transactions on speech and audio processing* 3:1–20.

Cole, P. & Morgan, J., eds. (1975) *Syntax and semantics 3: Pragmatics*. New York: Academic Press.

Combescure, P. (1981) 20 listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique* 56:34–38.

Content, A., Mousty, P. & Radeau, M. (1990) Brulex, une base de données lexicales informatise pour le français écrit et parlé. *L'Année Psychologique* 90:551–566.

Cookson, S. (1988) Final evaluation of VODIS voice operated database inquiry system. Proceedings of Speech-88, 7th FASE Symposium, Edinburgh, August. 1311–1320.

Cooper, W. & Sorensen, J. (1981) *Fundamental frequency in sentence production*. New York: Springer-Verlag.

Cosi, P. & Omologo, M. (1991) Caratterizzazione Statistica della Segmentazione Manuale del Segnale Vocale. Associazione Italiana Acustica (AIA) Meeting. Napoli, Italy, 10–12 April. Cited in Barry and Fourcin 1992.

Crowdy, S. (1993) *Spoken corpus design and transcription*. Harlow: Longman.

Cruse, D. A. (1986) *Lexical semantics*. Cambridge: CUP.

Crystal, D. (1980) *Introduction to language pathology*. London: Edward Arnold Ltd..

Crystal, D. (1985) *A dictionary of linguistics and phonetics*. Oxford, UK: Basil Blackwell.

Cucchiarini, C. (1993) Phonetic transcription: A methodological and emperical study. Doctoral thesis. Nijmegen: University of Nijmegen.

Dahlbäck, N. & Jönsson, A. (1986) A system for studying human-computer dialogues in natural language. Research Report, LiTH-IDA-R-86-42. Linkoping: Department of Computer and Information Science, Linkoping University.

Dahlbäck, N. & Jönsson, A. (1989) Empirical studies of discourse representations for natural language interfaces. Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics, Manchester. 291–298.

Dalsgaard, P. & Baekgaard, A. (1994) Spoken language dialogue systems. In: Niemann, H., De Mori, R. & Hanreider, G. (eds.) *Progress and prospects in speech and language technology*. Sankt Augustin: Infix. 178–191.

Damhuis, M., Boogaart, T., Veld, C. in't, Versteijlen, M., Schelvis, W., Bos, L. & Boves, L. (1994) Creation & analysis of the Dutch Polyphone Corpus. Proceedings International Conference on Spoken Language Processing, ICSLP, Yokohama. 1803–1806.

Date, C. J. (1987) *An introduction to database systems*. Reading, Massachusetts: Addison-Wesley.

De Mori, R., Gilloux, M., Mercier, G., Simon, M. A., Tarrides, C. & Vaissiere, J. (1984) Integration of acoustic, phonetic, prosodic and lexical knowledge in an expert system for speech understanding. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. 42.9.1–42.9.4.

De Mori, R. & Suen, C. Y., eds. (1985) *New systems and architectures for automatic speech recognition and synthesis*. NATO ASI Series F, vol.16. Berlin: Springer-Verlag.

De Pijper, J. R. (1983) *Modelling British English intonation*. Dordrecht: Foris.

Delogu, C., Carlo, A. D., Sementino, C. & Stecconi, S. (1993) A methodology for evaluating human-machine spoken language interaction. Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin, September. 1427–1430.

Delogu, C., Paoloni, A., Pocci, P. & Sementina, C. (1991) Quality evaluations of text-to-speech synthesizers using magnitude estimation, categorical estimation, pair comparison and reaction time methods. Proc. Eurospeech '91, Genova. 353–355.

Delogu, C., Paoloni, A., Ridolfi, P. & Vagges, K. (1993) Intelligibility of Italian text-to-speech synthesizers over ortophonic and telephonic channel. Proc. Eurospeech '93, Berlin, vol. 3. 1893–1896.

Delogu, C., Paoloni, A. & Sementina, C. (1992) Comprehension of natural and synthetic speech: Preliminary studies. In: SAM (ed.) *Final report. Year three: 1.III.91–28.II.1992*. London: ESPRIT Poject 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation. SAM Internal Report II.c.

Delogu, C., Paoloni, P., Pocci, P. & Sementina, C. (1992) A comparison among different methodologies for evaluating the quality of text-to-speech synthesis systems. In: SAM (ed.) *Final report. Year three: 1.III.91–28.II.1992*. London: ESPRIT Poject 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation. SAM Internal Report II.d.

Delomier, D., Meunier, A. & Morel, M. -A. (1989) Linguistic features of human-machine oral interaction. Eurospeech '89, Paris, vol. 2. 236–239.

Den Os, E. A. (1994) Transliteration of the Dutch Speech Styles Corpus. *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam* 18:87–94.

Diaper, D. (1986) Identifying the knowledge requirements of an expert system's natural language processing interface. In: Harrison, M. D. & Monk, A. F. (eds.) *People and Computers V: Proceedings of the 2nd Conference of the British Computer Society Human-Computer Interaction Specialist Group*. Cambridge: Cambridge University Press.

Diaper, D. (1989) The Wizard's apprentice: A program to help analyse natural language dialogues. In: Sutcliffe, A. & Macaulay, L. (eds.) *People and Computers: Designing for usability. Proceedings of the 2nd Conference of the British Computer Society Human-Computer Interaction Specialist Group*. Cambridge: Cambridge University Press.

Dolmazon, J. -M., Caërou, J. -C. & Barry, W. (1990) Initial development of SAM standard workstation. SAM-UCL-022, December (Appendix Se.10). London: University College London.

Dougherty, D. (1990) *sed & awk*. Sebastopol, CA: O'Reilly & Associates Inc..

Draxler, C. (1995) An introduction to the PhonDat-Verbmobil Database of Spoken German; Practical Applications of Prolog 95, Paris, 1995 .

Dreckschmidt, G. (1987) The linguistic component in the speech understanding system SPICOS. In: Tillmann, H. & Willée, G. (eds.) *Analyse und Synthese gesprochener Sprache, Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung, Bonn*. Hildesheim: Olms. 96–101.

Drullman, R. & Collier, R. (1993) Speech synthesis with accented and unaccented diphones. In: Van Heuven, V. J. & Pols, L. C. W. (eds.) *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation*. Berlin: Mouton de Gruyter. 147–156.

Duncan Jr., S. & Fiske, D. (1977) *Face-to-face interaction: Research, methods and theory*. Hillsdale, N.J.: Erlbaum.

Dybkjaer, H., Bernsen, N. O. & Dybkjaer, L. (1993) Wizard-of-Oz and the trade-off between naturalness and recognizer constraints. Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin, September. 947–950.

Edwards, J. A. & Lampert, M. D., eds. (1993) *Talking data: Transcription and coding in discourse research*. Hillsdale: Lawrence Erlbaum.

Egan, J. P. (1948) Articulation testing methods. *Laryngoscope* 58:955–991.

Ehrlich, U. (1986) Ein Lexikon für das natürlich-sprachliche Dialogsystem EVAR. Arbeitsberichte des IMMD, vol. 19. Erlangen, Germany: University of Erlangen-Nürnberg.

Eisen, B. (1993) Reliability of speech segmentation and labelling at different levels of transcription. *Proceedings of the Third European Conference on Speech Communication and Technology. 21–23 September 1993, Berlin, Germany* 1:673–676.

Elmasri, R. & Navathe, S. B. (1989) *Fundamentals of database systems.* Redwood City, California: The Benjamin/Cummings Publishing Company, Inc..

Erman, L. D. (1977) A functional description of the HEARSAY II speech understanding system. Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Hartford.

Erman, L. D. & Hayes-Roth, F. (1981) The Hearsay-II speech understanding system: Integrating knowledge to resolve uncertainty. In: Webber, B. L. & Nilsson, N. J. (eds.) *Readings in Artificial Intelligence*. Palo Alto, CA: Tioga. 349–389.

Erman, L. D. & Lesser, V. R. (1980) The Hearsay-II speech understanding system: A tutorial. In: Lea, W. A. (ed.) *Trends in speech recognition.* Englewood Cliffs, NJ: Prentice Hall. 361–381. Also in: A. Waibel & K.-F. Lee, eds. (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 235–245.

Esling, J. (1990) Computer coding of the IPA: Supplementary report. *Journal of the International Phonetic Association* 20(1):22–26.

Esling, J. & Gaylord, H. (1993) Computer Codes for phonetic symbols. *Journal of the International Phonetic Association* 23(2):83–97.

Esling, J. H. (1988) 7.1 Computer coding of IPA symbols and 7.3 detailed phonetic representation of computer data bases. *Journal of the International Phonetic Association* 18(2):99–106.

Evans, R. & Gazdar, G., eds. (1989) The DATR papers. Research Report: May 1989. Brighton: School of Cognitive and Computing Science, University of Sussex.

Evans, R. & Gazdar, G., eds. (1990) The DATR papers. Research Report: February 1990. Brighton: School of Cognitive and Computing Science, University of Sussex.

Fellbaum, K., Klaus, H. & Sotscheck, J. (1994) Hörversuche zur Beurteilung der Sprachqualität von Sprachsynthesesystemen für die deutsche Sprache. Fortschritte der Akustik, Plenarvorträge und Fachbeiträge der 20. Deutschen Jahrestagung für Akustik. 117–122. Dresden, DPG GmbH.

Ferguson, G. A. (1976) *Statistical analysis in psychology and education.* Tokyo: McGraw-Hill.

Ferrané, I., De Calmès, M., Cotto, D., Pécatte, J. -M. & Pérennou, G. (1992) Statistiques lexicales sur le corpus de textes utilisés dans le projet BREF: Questions de couverture lexicale. Proc. Communication Homme-Machine, Séminaire LEXIQUE, 21–22 January 1992, IRIT-UPS Toulouse. 217–226.

Fillmore, C. J. (1968) The case for case. In: Bach, E. & Harms, R. (eds.) *Universals in linguistic theory*. New York: Holt, Rinehart and Winston. 1–88.

Flanagan, J. L., Berkley, D. A., Elko, G. W. & Sondhi, M. M. (1991) Autodirective microphone systems. *Acoustica* 73:58–71.

Fourcin, A. J., Harland, G., Barry, W. & Hazan, V., eds. (1989) *Speech input and output assessment. Multilingual methods and standards.* Chichester: Ellis Horwood Ltd..

Fraser, N. M. (1991) Corpus-based evaluation of the SUNDIAL system. In: Neal, J. G. & Walter, S. M. (eds.) *Proceedings of the Natural Language Processing Systems Evaluation Workshop.* Rome: Rome Laboratory. Technical Report RL-TR-91-362.

Fraser, N. M. & Gilbert, G. N. (1991a) Effects of system voice quality on user utterances in speech dialogue systems. Proceedings of the Second European Conference on Speech Communication and Technology, Genova, September. 57–60.

Fraser, N. M. & Gilbert, G. N. (1991b) Simulating speech systems. *Computer Speech and Language* 5:81–99.

Fraser, N. M., Gilbert, N. & McDermid, C. (1992) The value of simulation data. Proceedings of the Workshop on Empirical Models and Methodology for Natural Language Dialogue Systems, Trento, April.

French, J. P. (1991) Updated notes for soundprint transcribers + one page sample text from COBUILD corpus. NERC-WP4-47, October. Birmingham: J.P. French Associated, York and COBUILD.

French, J. P. (1992) Transcription proposals: Multi-level system. Working paper, NERC-WP 4-50, October. Birmingham: University of Birmigham.

Furui, S. & Sondhi, M. M., eds. (1991) *Advances in Speech Signal Processing*. New York: Marcel Dekker Inc..

Gerbino, E., Baggia, P., Ciaramella, A. & Rullent, C. (1993) Test and evaluation of a spoken dialogue system. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP'93, Minneapolis, April.

Geutner, P. (1995) *Using morphology towards better large-vocabulary speech recognition systems.* Karlsruhe, Germany: Interactive Systems Laboratories, University of Karlsruhe. To appear in the Proceedings of ICASSP 1995.

Gibbon, D. (1991) Lexical signs and lexicon structure: Phonology and prosody in the ASL-lexicon. Research Report, ASL-MEMO-20-91/UBI. Bielefeld, Germany: University of Bielefeld.

Gibbon, D. (1992) ILEX: A linguistic approach to computational lexica. In: Klenk, U. (ed.) *Computatio linguae. Aufsätze zur algorithmischen und quantitativen Analyse der Sprache*. Stuttgart: Franz Steiner Verlag. 32–51.

Gibbon, D. (1993) Generalized DATR for flexible lexical access: PROLOG specification. Verbmobil Report 2, October 1993. Bielefeld, Germany: Bielefeld University.

Gibbon, D. (1995) The VERBMOBIL lexicon: Bielefeld lexicon database V2.1. VERBMOBIL Technisches Dokument 21, 31 January 1995. Bielefeld, Germany: Bielefeld University.

Gibbon, D. & Ehrlich, U. (1995) Spezifikationen für ein VERBMOBIL-Lexikondatenbankkonzept. VERBMOBIL Memo 69. Bielefeld, Ulm: Bielefeld University & Daimler Benz AG.

Gilbert, H. R. & Weismer, G. G. (1974) The effect of smoking on the speaking fundamental frequency of adult women. *Journal of Psycholinguistic Research* 3:225–231.

Gish, H. & Schmidt, M. (1994) Text-independent speaker identification. *IEEE Signal Processing* 11:18–32.

Goldsmith, J. (1990) *Autosegmental and metrical phonology*. Bloomington, Indiana: Indiana University Linguistics Club.

Goldstein, M. (1995) Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener . *Speech Communication* 16:225–244.

Goldstein, M., Lindström, B. & Till, O. (1992) Assessing global performance of speech synthesizers: Context effects when assessing Naturalness of Swedish sentence-pairs generated by 4 systems using 3 different assessment procedures (free number Magnitude Estimation, 5- and 11-point category scales). In: SAM (ed.) *Final report. Year three: 1.III.91–28.II.1992*. London: ESPRIT Poject 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation. SAM Internal Report II.a.

Goldstein, M. & Till, O. (1992) Assessing segmental intelligibility of two rule-based synthesizers and Natural Speech using the ESPRIT/SAMVCV test procedures (SOAP v3.0) in Swedish and testing for differences between two correlated proportions. In: SAM (ed.) *Final report. Year three: 1.III.91–28.II.1992*. London: ESPRIT Poject 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation. SAM Internal Report II.b.

Gong, Y. (1995) Speech recognition in noisy environments: A survey. *Speech Communication* 16:261–291.

Goodine, D., Hirschman, L., Polifroni, J., Seneff, S. & Zue, V. (1992) Evaluating interactive spoken language systems. Proceedings of the International Conference on Spoken Language Processing, ICSLP'92, Banff, October. 201–204.

Goorfin, L. (1989) Electronic dictionary pronounces over 83,000 words. *Speech Technology* 4(4):49–51.

Görz, G., ed. (1992) *KONVENS 92, 1. Konferenz "Verarbeitung natürlicher Sprache", Nürnberg, 7–9 October 1992*. Berlin: Springer-Verlag.

Green, D. (1986) Control, activation and resource: A framework and a model for the control of speech in bilinguals. *Brain and Language* 27:210–223.

Greenspan, S. L., Nusbaum, H. C. & Pisoni, D. B. (1985) Perception of speech generated by rule: Effects of training and attentional limitations. Research on Speech Perception Progress Report 11, pages 263–287. Indianapolis: Indiana University.

Grice, H. P. (1975) Logic and conversation. In: Cole, P. & Morgan, J. (eds.) *Syntax and semantics 3: Pragmatics*. New York: Academic Press. 41–58.

Grice, M., Vagges, K. & Hirst, D. (1991) Assessment of intonation in text-to-speech synthesis systems — A pilot test in English and Italian. *Proc. Eurospeech '91, Genova* 2:879–882.

Grice, M., Vagges, K. & Hirst, D. (1992a) Prosodic form tests. In: SAM (ed.) *Final report. Year three: 1.III.91–28.II.1992*. London: ESPRIT Poject 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation. Stage report So. 5, Part One.

Grice, M., Vagges, K. & Hirst, D. (1992b) Prosodic function tests. In: SAM (ed.) *Final report. Year three: 1.III.91–28.II.1992*. London: ESPRIT Poject 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation. Stage report So. 5, Part Two.

Grosz, B. J. (1977) The representation and use of focus in dialogue understanding. Unpublished doctoral thesis. Berkely: University of California.

Guindon, R., ed. (1988a) Cognitive Science and its applications for human-computer interaction.

Guindon, R. (1988b) A multidisciplinary perspective on dialogue structure in user-advisor dialogues. In: Guindon, R. (ed.) *Cognitive Science and its applications for human-computer interaction.* . 163–200.

Guindon, R., Shuldberg, K. & Connor, J. (1987) Grammatical and ungrammatical structures in user-advisor dialogues: Evidence for sufficiency of restricted languages in natural language interfaces to advisory systems. Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, Stanford. 41–44.

Guindon, R., Sladky, P., Brunner, H. & Connor, J. (1986) The structure of user-adviser dialogues: Is there method in their madness?. Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics. 224–230.

Guyomard, M. & Siroux, J. (1986a) PALABRE Phase 2 Experimental protocol. CNET/TSS/RCP WP4 task 3, May.

Guyomard, M. & Siroux, J. (1986b) PALABRE Phase 1 Experimental protocol. CNET/TSS/RCP WP4 task 3, April.

Guyomard, M. & Siroux, J. (1987) Experimentation in the specification of oral dialogue. NATO-ASI Recent Advances in Speech Understanding and Dialogue Systems, July.

Guyomard, M. & Siroux, J. (1988) Constitution incrementale d'un corpus de dialogues oraux cooperatifs. *Journal Acoustique* 1.

Hansen, J., Pelaez, C., Solana, L. & Vossen, P. (1992) Performance assessment and evaluation: Specification document. SUNSTAR Report II.4.

Harrison, M. D. & Monk, A. F., eds. (1986) *People and Computers V: Proceedings of the 2nd Conference of the British Computer Society Human-Computer Interaction Specialist Group.* Cambridge: Cambridge University Press.

't Hart, J., Collier, R. & Cohen, A. (1990) *A perceptual study of intonation.* Cambridge: Cambridge University Press.

Haton, J. -P., Pierrel, J. -M., Pérennou, G., Caelen, J. & Gauvain, J. -L. (1991) *Reconnaissance automatique de la parole.* Paris: Bordas.

Hauptmann, A. G. & Rudnicky, A. I. (1988) Talking to computers: An empirical investigation. *International Journal of Man-Machine Studies* 28:583–604.

Hayes, W. L. (1963) Statistics. (pages 269–280). New York: Holt, Rinehart and Winston, Inc..

Hazan, V. & Grice, M. (1989) The assessment of synthetic speech intelligibility using semantically unpredictable sentences. Proc. of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, 1.6.1–1.6.4.

Hazan, V. & Shi, B. (1993) Individual variability in the perception of synthetic speech. *Proc. Eurospeech '93, Berlin* 3:1849–1852.

Heemskerk, J. & Van Heuven, V. J. (1993) MORPA, a morpheme lexicon based morphological parser. In: Van Heuven, V. J. & Pols, L. C. W. (eds.) *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation.* Berlin: Mouton de Gruyter. 67–85.

Helfrich, H. (1979) Age markers in speech. In: Scherer, K. R. & Giles, H. (eds.) *Social markers in speech.* Cambridge: Cambridge University Press. 63–107.

Hertz, S., Kadin, J. & Karplus, K. (1985) The DELTA rule development system for speech synthesis from text. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. 1589–1601.

Heyer, G., Waldhur, K. & Khatchadourian, H. (1991) Motivation, goals and milestones of ESPRIT II MULTILEX. Génie Linguistique 91, vol. 1, Versailles, France, 16–17 January.

Hieronymus, J., Alexander, H., Bennett, C., Cohen, I., Davies, D., Dalby, J., Laver, J., Barry, W., Fourcin, A. & Wells, J. (1990) Proposed speech segmentation criteria for the SCRIBE project. SCRIBE Project Report.

Hirschman, L., Dahl, D. A., McKay, D. P., Norton, L. M. & Linebarger, M. C. (1990) Beyond class A: A proposal for automatic evaluation of discourse. Proceedings of the DARPA Workshop on Speech and Natural Language, Hidden Valley, PA, June. 109–112.

Hjelmquist, E., Jansson, B. & Torell, G. (1987) Psychological aspects on blind people's reading of radio-distributed daily newspapers. In: Knave, B. & Widebäck, P. G. (eds.) *Work with display units 86*. Amsterdam: North-Holland, Elsevier Science Publishers. 187–201.

Hockett, C. F. (1958) *A course in modern linguistics*. New York: Macmillan.

Höge, H., Marschall, E., Schmidbauer, O. & al. (1985) Worthypothesengenerierung im Projekt SPICOS. In: Niemann, H. (ed.) *Mustererkennung 85, 7. DAGM-Symposium Erlangen, Informatik-Fachberichte, vol. 107*. Berlin: Springer-Verlag. 175–179.

Holmes, J. N. (1988) *Speech synthesis and recognition*. Berkshire: Van Nostrand Reinhold.

House, A. S., Williams, C. E., Hecker, M. H. L. & Kryter, K. D. (1965) Articulation testing methods: Consonantal differentiation with a closed response set. *Journal of the Acoustical Society of America, JASA* 37:158–166.

House, J., Shitara, Y., Grice, M. & Howard-Jones, P. (1992) Evaluation of prosody in dialogue synthesis. *Speech, Hearing and Language* 6:89–108.

Houtgast, T. & Verhave, J. A. (1991) A physical approach to speech quality assessment: Correlation patterns in the speech spectrogram. Proc. Eurospeech '91, Genova, vol. 1. 285–288.

Houtgast, T. & Verhave, J. A. (1992) An objective approach to speech quality. In: SAM (ed.) *Final report. Year three: 1.III.91–28.II.1992*. London: ESPRIT Poject 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation. Stage report So. 9.

Howard-Jones, P. (1992a) SOAP, Speech Output Assessment Package. Version 4.0, ESPRIT SAM-UCL-042.

Howard-Jones, P. (1992b) Specification of listener dimensions. In: SAM (ed.) *Final report. Year three: 1.III.91–28.II.1992*. London: ESPRIT Poject 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation. Stage report So. 8, Part One.

ITU-T (1993) Draft recommendation P.8S — Subjective performance assessment of the quality of speech voice output devices . Study group 12 — contribution 6.

Jakobson, R., Fant, G. & Halle, M. (1951) *Preliminaries to speech analysis*. Cambridge: The MIT Press.

Jassem, W. & Lobacz, P. (1989) IPA phonemic transcription using an IBM PC and compatibles. *Journal of the International Phonetic Association* 19(1):16–23.

Jekosch, U. (1992) The Cluster-Identification Test. In: SAM (ed.) *Final report. Year three: 1.III.91–28.II.1992*. London: ESPRIT Poject 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation. Internal report II.e.

Jekosch, U. (1994) Speech intelligibility testing: On the interpretation of results. *Journal of the American Voice I/O Society* 15:63–79.

Jekosch, U. & Pols, L. C. W. (1994) A feature-profile for application-specific speech synthesis assessment and devaluation. Proc. 3rd International Conference on Spoken Language Processing, ICSLP, Yokohama.

Jelinek, F. (1985) A real-time, isolated-word, speech recognition system for dictation transcription. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. 858–861.

Johnston, R. D. (1993) An on-going series of subjective experiments to assess speech output from text-to-speech systems. Unpublished report to CCITT Study Group 12.

Jongenburger, W. & Van Bezooijen, R. (1992) *Evaluatie van ELK: Attitudes van de gebruikers, verstaanbaarheid en acceptabiliteit van de spraaksynthese, bruikbaarheid van het zoeksysteem.* Utrecht: Stichting Spraaktechnologie.

Jönsson, A. & Dalbäck, N. (1988) Talking to your computer is not like talking to your best friend. Proceedings of the First Scandanavian Conference on Artificial Intelligence, Tromso, Norway.

Joreskog, J. G. & Sorbom, D. (1984) *Lisrel VI. Analysis of linear structural relationships by maximum likelihood, instrument variables, and least squares methods* . Mooreville, IN: Scientific software.

Karttunen, L. (1983) KIMMO: A general morphological processor. *Texas Linguistic Forum* 22:165–186.

Kasschau, R. A., Lachman, R. & Laughery, K. R., eds. (1982) *Information technology and psychology: Prospects for the future.* New York: Praeger.

Kelley, J. F. (1983a) Natural language and computers: Six steps for writing an easy-to-use computer application. Unpublished doctoral dissertation. Baltimore: The Johns Hopkins University.

Kelley, J. F. (1983b) An empirical methodology for writing user-friendly natural language computer applications. Proceedings of the International Conference of Computer-Human Interaction, CHI '83.

Kelley, J. F. (1984) An interactive design methodology for user-friendly natural language office information applications. *Association for Computing Machinery Transactions on Office Information Systems* 2:26–41.

Kerkhoff, J., Wester, J. & Boves, L. (1984) A compiler for implementing the linguistic phase of a text-to-speech conversion system. In: Bennis, H. & Van Lessen-Kloecke, W. U. S. (eds.) *Linguistics in The Netherlands 1984*. Dordrecht: Foris. 111–119.

Klatt, D. H. (1976) The linguistics uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, JASA* 59:1208–1221.

Klatt, D. H. (1977) Review of the ARPA speech understanding project. *Journal of the Acoustical Society of America, JASA* 62(6):1345–1366. Also in: A. Waibel, K.-F. Lee, eds., (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 554–575.

Klatt, D. H. (1982) The KLATTalk text-to-speech conversion system. IEEE Conference on Acoustics, Speech and Signal Processing. 1589–1592.

Klatt, D. H. (1987) Review of text-to-speech conversion in English. *Journal of the Acoustic Society of America* 82:737–793.

Klenk, U., ed. (1992) *Computatio linguae. Aufsätze zur algorithmischen und quantitativen Analyse der Sprache.* Stuttgart: Franz Steiner Verlag.

Knave, B. & Widebäck, P. G., eds. (1987) *Work with display units 86.* Amsterdam: North-Holland, Elsevier Science Publishers.

Knowles, G. & Alderson, P. (1995) *Working with speech: The computational analysis of formal British English speech.* London: Longmans.

Knowles, G., Taylor, L. & Williams, B. (1995) *A Corpus of formal British English speech.* London: Longmans.

Knuth, D. (1973) *The art of computer programming 3: Sorting and searching.* Reading, Massachusetts: Addison-Wesley.

Kohler, K., Lex, G., Pätzold, M., Scheffers, M., Simpson, A. & Thon, W. (1994) Handbuch zur Datenaufnahme in TP14 von Verbmobil – 3.0. Technisches Dokument Nr.11.

Kohler, K., Pätzold, M. & Simpson, A. (1994) Handbuch zur Segmentation und Etikettierung von Spontansprache – 2.3. Technisches Dokument Nr.16.

Kornai, A. (1991) Formal phonology. Doctoral dissertation. Stanford: Stanford University.

Koskenniemi, K. (1983) *Two-level morphology: A general computational model for word-form recognition and production.* Helsinki, Finland: University of Helsinki, Department of General Linguistics.

Kraft, V. & Portele, T. (1995) Quality evaluation of five German speech synthesis systems. Acta Acustica. to appear.

Kryter, K. D. (1962a) Validation of the Articulation Index. *Journal of the Acoustical Society of America, JASA* 34:1698–1702.

Kryter, K. D. (1962b) Methods for the calculation and use of the Articulation Index. *Journal of the Acoustical Society of America, JASA* 34:1689–1697.

Labov, W. (1972) *Sociolinguistic patterns.* Pennsylvania: University of Pennsylvania Press.

Labov, W. (1994) *Principles of linguistic change. Volume 1: Internal factors.* Oxford: Blackwell.

Labrador, C. & Dinesh, P. (1984) Experiments in speech interaction with conventional data services. *Interact '84*:104–108.

Lacouture, R. & Normandin, Y. (1993) Efficient lexical access strategies. Proc. European Conference on Speech Technology.

Ladefoged, P. (1975) *A course in phonetics.* New York: Harcourt, Brace, Jovanovich.

Langer, H. & Gibbon, D. (1992) DATR as a graph representation language for ILEX speech oriented lexica. Research Report, March 1992, ASL-TR-43-92/UBI. Bielefeld, Germany: University of Bielefeld.

Langeweg, S. J. (1988) The stress system of Dutch. Doctoral dissertation. Leiden: Leiden University.

Laver, J. (1991) The gift of speech. Papers in the analysis of speech and voice. Edinburgh: Edinburgh University Press.

Laver, J. (1994) *Principles of phonetics.* Cambridge: Cambridge University Press.

Laver, J., McAllister, J., McAllister, M. & Jack, M. (1988) A Prolog-based automatic text-to-phoneme conversion system for British English. Proceedings of the Second Symposium on Advanced Man-Machine Interface through Spoken Language, November 19–22, Hawaii.

Laver, J., McAllister, M. & McAllister, J. (1989) Pre-processing of anomalous text-strings in an automatic text-to-speech system. In: Ramsaran, S. (ed.) *Studies in the pronunciation of English: A commemorative volume in memory of A.C. Gimson.* London: Croon Helm.

Lea, W. A., ed. (1980) *Trends in speech recognition.* Englewood Cliffs, NJ: Prentice Hall.

Lee, K. -F., Hon, H. -W. & Reddy, R. (1990) An overview of the SHPINX speech recognition system. In: Waibel, A. & Lee, K. -F. (eds.) *Readings in speech recognition.* San Mateo, California: Morgan Kaufmann Publishers. 600–610. appeared first in IEEE Transactions on Acoustics, Speech and Signal Processing 1990.

Leggett, J. & Williams, G. (1984) An empirical investigation of voice as an input modality for computer programming. *International Journal of Man-Machine Studies* 21:493–520.

Lehiste, I. (1970) *Suprasegmentals.* Cambridge, Mass.: MIT Press.

Lehiste, I., Olive, J. & Streeter, L. A. (1976) Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America, JASA* 60:1199–1202.

Lesser, V. R., Fennell, R. D., Erman, L. D. & Reddy, D. R. (1975) Organization of the Hearsay-II speech understanding system. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-23. 11–23.

Levelt, J. M. (1989) *Speaking: From intonation to articulation.* ACL-MIT Press Series in Natural Language Processing. Cambridge Massachusetts, London, England: Bradford Book—The MIT-Press.

Life, M. A., Lee, M. A. & Long, J. B. (1988) Assessing the usability of future speech technology: Towards a method. Speech '88: 7th FASE Symposium, Edinburgh.

Likert, R. (1932) A technique for the measurement of attitudes. *Archives of Psychology* 140.

Lindberg, B., Andersen, B., Baekgaard, A., Brøndsted, T., Dalsgaard, P. & Kristiansen, J. (1992) An integrated dialogue design and continuous speech recognition system environment. Proceedings of the International Conference on Spoken Language Processing, ICSLP, Banff. 1553–1556.

Llisterri, J. (1994) Prosody Encoding Survey, Multext – LRE Project 62–050.

Llisterri, J. & Mariño, J. B. (1993) Spanish adaptation of SAMPA and automatic phonetic transcription. In: SAM-A (ed.) *SAM-A periodic progress report.* London: ESPRIT Project 6819 (SAM-A) Speech technology assessment in multilingual applications, Year 1, 1 April 1993–30 September 1993. 1–9. Document No: SAM-A/UPC/001/V1.

Logan, J. S., Greene, B. G. & Pisoni, D. B. (1989) Measuring the segmental intelligibility of synthetic speech produced by ten text-to-speech systems. *Journal of the Acoustical Society of America, JASA* 86:566–581.

Logan, J. S., Pisoni, D. B. & Greene, B. G. (1985) Measuring the segmental intelligibility of synthetic speech: Results from eight text-to-speech systems. Research on speech perception Progress Report 11, 3–31. Indianapolis: Indiana University.

Loman, H. & Boves, L. (1993) Development of rule based synthesis for text-to-speech. In: Van Heuven, V. J. & Pols, L. C. W. (eds.) *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation.* Berlin: Mouton de Gruyter. 157–168.

Lowerre, B. & Reddy, R. (1980) The Harpy speech understanding system. In: Lea, W. A. (ed.) *Trends in speech recognition*. Englewood Cliffs, NJ: Prentice Hall. 340–360. Also in: A. Waibel, K.-F. Lee, eds., (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 576–586.

Luce, P. A., Feustel, T. C. & Pisoni, D. B. (1983) Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors* 25:17–32.

Luzzati, D. & Neel, F. (1989) Dialogue behaviour induced by machine. Eurospeech '89, Paris, vol. 2. 601–604.

Lyons, J. (1977) *Semantics. Volumes I and II.* Cambridge: Cambridge University Press.

Maassen, B. & Povel, D. -J. (1985) The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech. *Journal of the Acoustical Society of America, JASA* 78:877–886.

MacDermid, C. (1993) Features of naive callers' dialogues with a simulated speech understanding and dialogue system. Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin, September. 955–958.

Manous, L. M., Dedina, M. J., Nusbaum, H. C. & Pisoni, D. B. (1985) Speeded sentence verification of natural and synthetic speech. Research on Speech Perception Progress Report 11. Indianapolis: Indiana University.

Marascuilo, L. A. & Serlin, R. C. (1988) *Statistical methods for the social, and behavioral sciences.* New York: Freeman and company.

Marslen-Wilson, W., ed. (1989) *Lexical representation and process.* Cambridge, Massachusetts and London, England: The MIT Press.

Martin, D. (1986) *Advanced database techniques.* Cambridge: The MIT Press.

Megaw, E. D., ed. (1984) *Contemporary ergonomics.* London: Taylor and Francis.

Mérialdo, B. (1988) Multi-level decoding for very-large-size-dictionary speech recognition. *IBM Journal of Research and Development* 32(2):169–301.

Millar, B., Vonwiller, J., Harrington, J. & Dermody, P. (1994) The Australian National Database of Spoken Language. IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP, Adelaide, vol. 1. 97–100.

Monaghan, A. I. C. & Ladd, D. R. (1989) Evaluating intonation in the CSTR text-to-speech system. Proc. ESCA Workshop on Speech I/O Assessment and speech databases, Noordwijkerhout. 3.6.1–3.6.4.

Monaghan, A. I. C. & Ladd, D. R. (1990) Symbolic output as the basis for evaluating intonation in text-to-speech systems. *Speech Communication* 9:305–314.

Morimoto, T., Shikano, K., Iida, H. & Kurematsu, A. (1990) Integration of speech recognition and language processing in the spoken language translation system SL-TRANS. Proc. International Conference on Spoken Language Processing, ICSLP, Kyoto. 921–928.

Moulines, E. & Charpentier, F. (1990) Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9:453–467.

Müller, C. & Runge, F. (1993) Dialogue design principles — key for usability of voice processing. Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin, September. 943–946.

Murray, I. R. & Arnott, J. L. (1993a) A tool for the rapid development of new synthetic personalities. Proc. ESCA Workshop on Speech and Language Technologies for Disabled Persons, Stockholm. 111–114.

Murray, I. R. & Arnott, J. L. (1993b) Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America, JASA* 93:1097–1108.

Nespor, M. & Vogel, I. (1986) *Prosodic phonology*. Dordrecht: Foris.

Newell, A. F. (1978) The palantype transcription unit — its history and progress to date. *Hearing*:99–104. May/June.

Newell, A. F. (1989) Speech simulation studies — performance and dialogue specification. In: Peckham, J. (ed.) *Recent developments and applications of natural language processing*. London: Kogan Page. 141–157.

Ney, H., Mergel, D., Noll, A. & Paeseler, A. (1988) Overview of speech recognition in the SPICOS system. In: Niemann, H., Lang, M. & Sagerer, G. (eds.) *Recent advances in speech understanding and dialog systems*. NATO ASI Series F. Vol. 46. Berlin: Springer-Verlag. 305–310.

Nielsen, P. B. & Baekgaard, A. (1992) Experiences with a dialogue description formalism for realistic applications. Proceedings of the International Conference on Spoken Language Processing, ICSLP, Banff. 719–722.

Niemann, H., ed. (1985) *Mustererkennung 85, 7. DAGM-Symposium Erlangen, Informatik-Fachberichte, vol. 107*. Berlin: Springer-Verlag.

Niemann, H., Brietzmann, A., Mühlfeld, R., Regel, P. & Schukat, G. (1985) The speech understanding and dialog system EVAR. In: De Mori, R. & Suen, C. Y. (eds.) *New systems and architectures for automatic speech recognition and synthesis, NATO ASI Series F, vol.16*. Berlin: Springer-Verlag. 271–302.

Niemann, H., De Mori, R. & Hanreider, G., eds. (1994) *Progress and prospects in speech and language technology*. Sankt Augustin: Infix.

Niemann, H., Lang, M. & Sagerer, G., eds. (1988) *Recent advances in speech understanding and dialog systems*. NATO ASI Series F, Volume. Vol. 46. Berlin: Springer-Verlag.

Niemann, H., Nöth, E., Mast, M. & Schukat-Talamazzini, E. G. (1992) Ein Lexikon für ein natürlichsprachliches Dialogsystem. Beiträge des ASL-Lexikonworkshops, Wandlitz, 26–27 November. 15–18. ASL-TR-40-92/ZSB.

Nolan, F. (1987) The limits of segmental description. *Proceedings of the Eleventh International Conference of Phonetic Sciences. 1–7 August 1987, Tallinn, Estonia* 5:411–414.

Nooteboom, S. G. & Kruijt, J. G. (1987) Accents, focus distribution, and the perceived distribution of given and new information. *Journal of the Acoustical Society of America, JASA* 82:1512–1524.

Nossin, M. (1991) Le projet GENELEX: EUREKA pour les dictionnaires genériques. Génie Linguistique 91, vol. 1, Versailles, France, 16–17 January 1991.

Nunn, A. M. & Van Heuven, V. J. (1993) MORPHON: Lexicon-based text-to-phoneme conversion and phonological rules. In: Van Heuven, V. J. & Pols, L. C. W. (eds.) *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation*. Berlin: Mouton de Gruyter. 88–113.

Nusbaum, H. C., Greenspan, S. L. & Pisoni, D. B. (1986) Perceptual attention in monitoring natural and synthetic speech. Research on Speech Perception Progress Report 12. Indianapolis: Indiana University.

Nye, P. W. & Gaitenby, J. H. (1974) The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research, 37/38, pages 169–190.

Nye, P. W., Ingemann, F. & Donald, L. (1975) Synthetic speech comprehension: A comparison of listener performances with and preferences among different speech forms. Haskins Laboratories Status Report on Speech Research, 41.

O'Malley, M. H. & Caisse, M. (1987) How to evaluate text-to-speech systems. *Speech Technology* 3:66–75.

O'Neill, J. J. (1975) Measurement of hearing by tests of speech and language. In: Singh, S. (ed.) *Measurement procedures in speech, hearing, and language*. Baltimore: University Park Press. 219–252.

O'Shaughnessey, D. (1979) Linguistic features in fundamental frequency patterns. *Journal of Phonetics* 7:119–145.

O'Shaughnessy, D. (1986) Speaker recognition. IEEE ASSP Magazine. 4–17.

Oerder, M. & Ney, H. (1993) Word graphs: An efficient interface between continuous-speech recognition and language understanding. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol. 2. 119–122.

Olsen, G. M. & Olsen, J. S. (1990) *User-centered design of Collaborative Technology*. Cognitive Science and Machine Intelligence Laboratory. To appear in Organizational Computing 32.

Pallet, D. S., Fisher, W. M. & Garofolo, J. S. (1990) DARPA ATIS Results, June 1990. Proceedings of the DARPA Workshop on Speech and Natural Language, Hidden Valley, PA, June. 114–121.

Parducci, A. (1965) Category judgement: A range-frequency model. *Psychological Review* 72:407–418.

Parducci, A. (1974) Category-judgment: A range-frequency model. *Psychological Review* 72:407–418.

Paulus, E., ed. (1987) *Mustererkennung 87, 9. DAGM-Symposium Braunschweig, Informatik-Fachberichte*. Berlin: Springer-Verlag.

Pavlovic, C. V., Rossi, M. & Espesser, R. (1990) Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis system. *Journal of the Acoustical Society of America, JASA* 87:373–381.

Pavlovic, C. V., Rossi, M. & Espesser, R. (1991) Perceived spectral energy distributions for Eurom-0 speech and for some synthetic speech. Proc. 12th International Congress of Phonetic Sciences, Aix-en-Provence, vol. 5. 418–421.

Peckels, J. P. & Rossi, M. (1973) Le test diagnostic par paires minimales. Adaptation au Français du "Diagnostic Rhyme Test" de W.D. Voiers. *Revue d'Acoustique* 27:245–262.

Peckham, J. (1993) A new generation of spoken dialogue systems: Results and lessons from the SUNDIAL project. Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin, September. 33–40.

Pérennou, G., Cotto, D., De Calmès, M., Ferrané, I. & Pécatte, J. -M. (1992) Le projet BDLEX de base de données lexicales du Français écrit et parlé. Proc. Communication Homme-Machine, Séminaire LEXIQUE, 21–22 January 1992, IRIT-UPS Toulouse. 153–171.

Pérennou, G., Cotto, D., De Calmès, M., Ferrané, I., Pécatte, J. M. & Tihoni, J. (1991) Composantes phonologique et orthographique de BDLEX. Deuxièmes Journées Nationales du GRECO-PRC Communication Homme-Machine, Toulouse, 29–30 January 1991. 351–362.

Pérennou, G. & De Calmès, M. (1987) BDLEX lexical data and knowledge base of spoken and written French. European Conference on Speech Technology, Edinburgh, vol. 1. 393–396.

Pérennou, G. & Tihoni, J. (1992) Lexique et phonologie en reconnaissance de la parole. Proc. Communication Homme-Machine, Séminaire LEXIQUE, 21–22 January 1992, IRIT-UPS Toulouse. 41–57.

Perkins, W. H. (1977) *Speech pathology, an applied behavioral science*. Saint Louis: The C.V. Mosby Company.

Philips, S. U., Stelle, S. & Tanz, C. (1987) *Language, gender and sex in comparative perspective*. Cambridge: Cambridge University Press.

Pisoni, D. B., Greene, B. G. & Nusbaum, H. C. (1985) Some human factors issues in the perception of synthetic speech. Proc. Speech Tech '85, New York. 57–61.

Pisoni, D. B., Nusbaum, H. C. & Greene, B. G. (1985) Perception of synthetic speech generated by rule. Proceedings of the IEEE, 73. 1665–1676.

Pitrelli, J. F., Beckman, M. E. & Hirschnerg, J. (1994) Evaluation of prosodic transcription labeling reliability in the ToBI framework. Proceedings of the International Conference on Spoken Language Processing (ICSLP). 18–22 September 1994, Yokohama, Japan .

Plenat, M. (1991) Vers d'une phonémisation des sigles. Deuxièmes journées du GDR-PRC Communication Homme-Machine, EC2 Editeur, Toulouse, 29–30 January 1991. 363–371.

Plomp, R. & Mimpen, A. M. (1979) Improving the reliability of testing the speech reception threshold for sentences. *Audiology* 8:43–52.

Pols, L. C. W. (1991) Quality assessment of text-to-speech synthesis-by-rule. In: Furui, S. & Sondhi, M. M. (eds.) *Advances in speech signal processing*. New York: Marcel Dekker Inc.. 387–416.

Pols, L. C. W., Lefevre, J. -P., Boxelaar, G. & Van Son, N. (1987) Word intelligibility of a rule synthesis system for French. Proc. European Conference on Speech Technology, Edinburgh, vol. 1. 179–182.

Ponamale, M., Bilange, E., Choukri, K. & Soudoplatoff, S. (1990) A computer-aided approach to the design of an oral dialogue system. Proceedings of Eastern Multiconference, Nashville.

Portele, T., Heuft, B., Höfer, F., Meyer, H. & Hess, W. (1994) A new high quality speech synthesis system for German. Proc. Yokohama/New Paltz.

Pratt, R. L. (1987) Quantifying the performance of text-to-speech synthesizers. *Speech Technology*:54–64.

Price, P. J. (1990) Evaluation of spoken language systems: The ATIS domain. Proceedings of the DARPA Workshop on Speech and Natural Language, Hidden Valley, PA, June. 91–95.

Quené, H. (1993) Segment durations and accent as cues to word segmentation in Dutch. *Journal of the Acoustical Society of America, JASA* 94:2027–2035.

Radford, A. (1988) *Transformational grammar: A first course*. Cambridge: CUP.

Ralston, J. V., Pisoni, D. B., Lively, S. E., Greene, B. G. & Mullennix, J. W. (1991) Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors* 33:471–491.

Rayner, M., Alshawi, H., Breton, I., Carter, D., Digalakis, V., Gamback, B., Kaja, J., Karlgren, J., Lyberg, B., Pulman, S., Price, P. & Samuelsson, C. (1993) A speech to speech translation system built from standard components. Proc. of a Workshop: Human Language Technology, Princeton, NJ, 21–24 March. 217–222.

Reddy, D., ed. (1975) *Speech recognition*. New York, N.Y.: Academic Press.

Reilly, R. (1987) Ill-formedness and mis-communication in person-machine dialogue. *Information and Software Technology* 29:69–74.

Richards, M. A. & Underwood, K. M. (1984a) Talking to machines. How are people naturally inclined to speak?. In: Megaw, E. D. (ed.) *Contemporary Ergonomics*. London: Taylor and Francis.

Richards, M. A. & Underwood, K. M. (1984b) How should people and computers speak to each other?. *Interact '84*:33–36.

Ritchie, G. D., Black, A. W., Russell, G. J. & Pulman, S. G. (1992) *Computational morphology*. Cambridge, Massachusetts and London: The MIT Press.

Roach, P., Knowles, G., Varadi, T. & Arnfield, S. (1993) MARSEC: A machine-readable Spoken English corpus. *Journal of the International Phonetic Association* 23(2):47–53.

Roach, P., Roach, H., Dew, A. & Rowlands, P. (1990) Phonetic analysis and the automatic segmentation and labeling of speech sounds. *Journal of the International Phonetic Association* 20(1):15–21.

Roelofs, J. C. (1987) Synthetic speech in practice: Acceptance and efficiency. *Behaviour and Information Technology* 6:403–410.

Rudnicky, A., Baumeister, L., De Graff, K. & Lehmann, E. (1987) The lexical access component of the CMU continuous speech recognition system. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP.

Ruske, G. (1985) Demisyllables as processing units for automatic speech recognition and lexical access. In: De Mori, R. & Suen, C. Y. (eds.) *New systems and architectures for automatic speech recognition and synthesis*. NATO ASI Series F. Vol. 16. Berlin: Springer-Verlag. 593–611.

Ruske, G. & Schotola, T. (1981) The efficiency of demisyllable segmentation in the recognition of spoken words. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Atlanta. 971–974.

SAM-A, ed. (1993) *SAM-A periodic progress report*. London: ESPRIT Project 6819 (SAM-A) Speech technology assessment in multilingual applications, Year 1, 1 April 1993–30 September 1993 .

Sagerer, G. (1990) *Automatisches Verstehen gesprochener Sprache*. Reihe Informatik. Vol. 74. Mannheim: Bibliographisches Institut.

Salza, P. L., Di Fabbrizio, G., Oreglia, M., Falcone, M., Sementina, C. & Delogu, C. (1993) Development of a context dependent methodology for Text-to-Speech synthesis evaluation in interactive Dialogue systems. In: SAM-A (ed.) *Speech technology assessment in multilingual applications*. London: ESPRIT Project 6819 (SAM-A), SAM-A Periodic Progress Report. Year 1, 1 April 1993–30 September 1993. Report R2.

Scharpff, P. J. & Van Heuven, V. J. (1988) Effects of pause insertion on the intelligibility of low quality speech. Proceedings 7th FASE/Speech '88 Symposium, Edinburgh. 261–269.

Scherer, K. R. & Giles, H., eds. (1979) *Social markers in speech*. Cambridge: Cambridge University Press.

Schmidt, M. S. & Watson, G. S. (1991) The evaluation and optimization of automatic speech segmentation.

Schröder, S., Sagerer, G. & Niemann, H. (1987) Wissensakquisition mit semantischen Netzwerken. In: Paulus, E. (ed.) *Mustererkennung 87, 9. DAGM-Symposium Braunschweig, Informatik-Fachberichte*. Berlin: Springer-Verlag. 305–309.

Schukat-Talamazzini, E. G. (1993) Automatische Spracherkennung. Habilitationsschrift. Erlangen, Germany: Erlangen University.

Schwab, E. C., Nusbaum, H. C. & Greene, B. G. (1985) Perception of synthetic speech generated by rule. Proc. IEEE, 73. 1665–1676.

Searle, J. R. (1969) *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Searle, J. R. (1979) *Expression and meaning*. Cambridge: Cambridge University Press.

Sells, P. (1985) *Lectures on contemporary syntactic theories: An introduction to Government-Binding theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar* . Stanford, California: CSLI Center for the Study of Language and Information.

Shackel, B., ed. (1981) *Man-computer interaction: Human factors aspects of computers and people*. Rockville, MD: Sijthoff and Noordhoff.

Siegel, S. (1956) *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Silverman, K., Basson, S. & Levas, S. (1990) Evaluating synthesizer performance: Is segmental intelligibility enough?. Proc. International Conference on Spoken Language Processing, ICSLP, Kobe. 981–984.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992) ToBI: A standard for labeling English prosody. *Proceedings of the 1992 International Conference on Spoken Language Processing, ICSLP. 12–16 October 1992, Banff, Canada* 2:867–870.

Simpson, A. & Fraser, N. M. (1993) Black box and glass box evaluation of the Sundial system. Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin, September. 1423–1426.

Singh, S., ed. (1975) *Measurement procedures in speech, hearing, and language*. Baltimore: University Park Press.

Smith, P. M. (1979) Sex markers in speech. In: Scherer, K. R. & Giles, H. (eds.) *Social markers in speech*. Cambridge: Cambridge University Press. 109–146.

Smith, R. W., Hipp, D. R. & Biermann, A. W. (1992) A dialog control algorithm and its performance. Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, April. 9–16.

Soclof, M. S. (1990) A comparison of spontaneous speech and read speech in human-machine problem solving dialogues. Unpublished Master of Science thesis. Cambridge, MA: Massachusetts Institute of Technology.

Sorin, C. (1994) Towards high-quality multilingual text-to-speech. Proc. CRIM/FORWISS workshop, Munich. 53–62. Also to appear in H. Nieman, ed., Progress and prospects in research and technology, Infix Publishing Company, Sankt Augustin.

Sotscheck, J. (1982) Ein Reimtest fuer Verstaendlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachuebertragungsgeraete. *Der Fernmeldung* 36:1–84.

Sperberg-McQueen, C. M. & Burnard, L., eds. (1994) *Guidelines for electronic text encoding and interchange.* TEI P3. Chapter 1 Transcription of Speech. Chicago and Oxford: Association for Computational Linguistics, Association for Computers and the Humanities, Association for Literary and Linguistic Computing.

Spiegel, M. F., Altom, M. J., Macchi, M. J. & Wallace, K. L. (1990) Comprehensive assessment of the telephone intelligibility of synthesized and natural speech. *Speech Communication* 9:279–291.

Sproat, R., Hirschberg, J. & Yarowsky, D. (1992) A corpus-based synthesizer. Proc. 2nd International Conference on Spoken Language Processing, ICSLP, Banff, vol. 1. 563–566.

Steeneken, H. J. M. (1982) Ontwikkeling en toetsing van een Nederlandstalige Diagnostische Rijmtest voor het testen van spraakcommunicatiekanalen. Rapport IZF 1982–13. Soesterberg: IZF.

Stubbs, M. (1984) *Discourse analysis. The sociolinguistic analysis of natural language.* Oxford: Blackwell.

Sundheim, B. M. (1991) Third message understanding evaluation and conference (MUC-3): Phase 1 status report. Proceedings of the DARPA Workshop on Speech and Natural Language, Pacific Grove, CA, February. 301–305.

Sutcliffe, A. & Macaulay, L., eds. (1989) *People and computers: Designing for usability. Proceedings of the 2nd Conference of the British Computer Society Human-Computer Interaction Specialist Group.* Cambridge: Cambridge University Press.

Svartvik, J., ed. (1992) *Directions in corpus linguistics: Proceedings of the Nobel Symposium 82, New York* . Berlin: Mouton de Gruyter.

Syrdal, A. K. & Sciacca, B. A. (1994) Testing the intelligibility of text-to-speech output with the Diagnostic Pairs Sentence Intelligibility Evaluation. ITD-94-23828A, Technical Memorandum. AT&T Bell Laboratories. Submitted to the Journal of the Acoustical Society of America, JASA.

Taylor, M. M., Neel, F. & Bouwhuis, D. G., eds. (1987) *The structure of multimodal dialogue.* Amsterdam: North Holland.

Terken, J. M. B. (1985) Use and function of accentuation. Some experiments. Doctoral dissertation. Leiden: Leiden University.

Terken, J. M. B. (1993) Human and synthetic intonation: A case study. In: Van Heuven, V. J. & Pols, L. C. W. (eds.) *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation.* Berlin: Mouton de Gruyter. 241–259.

Terken, J. M. B. & Collier, R. (1989) Automatic synthesis of natural-sounding intonation for text-to-speech conversion in Dutch. Proc. Eurospeech '89, vol. 1, Paris. 357–359.

Thomas, T. J. & Winski, R. (1987) Speech recognizers, assessment in the laboratory, not in the field. *Speech Technology* 3(4):88–93.

Thorsen, N. (1980) A study of the perception of sentence intonation — Evidence from Danish. *Journal of the Acoustical Society of America, JASA* 67:1014–1030.

Thurmair, G. (1986) Linguistische Analyse im Projekt SPICOS.

Tillmann, H. & Willée, G., eds. (1987) *Analyse und Synthese gesprochener Sprache, Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung, Bonn.* Hildesheim: Olms.

Tubach, J. -P. & Bok, L. -J. (1985) *ZUT — Petit dictionnaire français.* Grenoble: Institut de Phonitique de Grenoble, avec le concours du CNRS (GRECO Comm. Parlie).

Tubach, J. P. & Doignon, P. (1991) A system for natural spoken language queries: Design, implementation and assessment. Proceedings of the 2nd European Conference on Speech Communication and Technology, Genova, September. 1473–1476.

Turing, A. M. (1950) Computing machinery and intelligence. *Mind* 59:433–460.

Van Bezooijen, R. (1984) *Characteristics and recognizability of vocal expressions of emotion.* Dordrecht: Foris.

Van Bezooijen, R. (1986) Lay ratings of long-term voice-and-speech characteristics. In: Beukema, F. & Hulk, A. (eds.) *Linguistics in the Netherlands 1986.* Dordrecht: Foris. 1–7.

Van Bezooijen, R. (1988a) Evaluation of two synthesis systems for Dutch — Development and applications of intelligibility tests. SPIN-ASSP Report No. 5. Utrecht: Stichting Spraaktechnologie.

Van Bezooijen, R. (1988b) Evaluation of two synthesis systems for Dutch — Intelligibility and overall quality of initial and final consonant clusters. SPIN-ASSP Report No. 3. Utrecht: Stichting Spraaktechnologie.

Van Bezooijen, R. (1989) Evaluation of the suitability of Dutch text-to-speech conversion for application in a digital daily newspaper. Proc. ESCA Workshop Speech I/O Assessment and Speech Databases, Noordwijkerhout, 6.3.1–6.3.4.

Van Bezooijen, R. & Jongenburger, W. (1993) Evaluation of an electronic newspaper for the blind in the Netherlands — Intelligibility, acceptability, adequacy, and users' attitudes. Proc. ESCA Workshop on Speech and Language Technology for Disabled Persons, Stockholm. 195–198.

Van Bezooijen, R. & Pols, L. C. W. (1987) Evaluation of two synthesis-by-rule systems for Dutch. Proc. European Conference on Speech Technology, vol. 1. 179–183.

Van Bezooijen, R. & Pols, L. C. W. (1989) Evaluation of a sentence accentuation algorithm for a Dutch text-to-speech system. Proc. Eurospeech '89, vol. 1, Paris. 218–221.

Van Bezooijen, R. & Pols, L. C. W. (1990) Evaluating text-to-speech systems: Some methodological aspects. *Speech Communication* 9:263–270.

Van Bezooijen, R. & Pols, L. C. W. (1993) Evaluation of text-to-speech conversion for Dutch. In: Van Heuven, V. J. & Pols, L. C. W. (eds.) *Analysis and synthesis of speech: Strategic research towards high-quality text-to-speech conversion.* Berlin: Mouton de Gruyter. 339–360.

Van Bezooijen, R. & Van Hout, R. (1985) Accentedness ratings and phonological variables as measures of variation in pronunciation. *Language and Speech* 28:129–142.

Van Coile, B. (1989) The DEPES development system for text-to-speech synthesis. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. 250–253.

Van Compernolle, D., Smolders, J., Jaspers, P. & Hellemans, T. (1991) Speaker clustering for dialectic robustness in speaker independent recognition . *Proceedings of Eurospeech '91, Genova* 2:723–726.

Van Dommelen, W. A. (1993) Speaker height and weight identification: A re-evaluation of some old dates. *Journal of Phonetics* 21:337–341.

Van Hemert, J. P., Adriaens-Porzig, U. & Adriaens, L. M. H. (1987) Speech synthesis in the SPICOS-project. In: Tillmann, H. G. & Willée, G. (eds.) *Analyse und Synthese gesprochener Sprache: Vorträge im Rahmen der Jahrestagung 1987 der Gesellschaft für Linguistische Datenverarbeitung e.V., Bonn, 4–6 March.* Hildesheim: Olms. 34–39.

Van Heuven, V. J. & Pols, L. C. W., eds. (1993) *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation.* Berlin: Mouton de Gruyter.

Van Heuven, V. J. & Scharpff, P. J. (1991) Acceptability of several speech pausing strategies in low quality speech synthesis; interaction with intelligibility. Proc. 12th International Congress of Phonetic Sciences, Aix-en-Provence. 458–461.

Van Holsteijn, Y. (1993) TextScan: A preprocessing module for automatic text-to-speech conversion. In: Van Heuven, V. J. & Pols, L. C. W. (eds.) *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation.* Berlin: Mouton de Gruyter. 27–41.

Van Hout, R. (1989) De structuur van taalvariatie, een sociolinguistisch onderzoek naar het stadsdialect van Nijmegen. Doctoral dissertation. Nijmegen: University of Nijmegen.

Van Santen, J. P. H. (1992) Diagnostic perceptual experiments for text-to-speech system evaluation. Proceedings of the International Conference on Spoken Language Processing, ICSLP, vol. 1. 555–558.

Van Santen, J. P. H. (1993) Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language* 7:49–100.

Van Santen, J. P. H. (1994) Using statistics in text-to-speech system construction. Proc. ESCA/IEEE Workshop on Speech Synthesis, Mohonk NY. 240–243.

Van Son, N., Pols, L. C. W., Sandri, S. & Salza, P. L. (1988) First quality evaluation of a diphone-based speech synthesis system for Italian. Proc. 7th FASE/Speech '88 Symposium, Edinburgh, vol. 2. 429–436.

Vergeynst, N. A., Edwards, K., Foster, J. C. & Jack, M. A. (1993) Spoken dialogues for human-computer interaction over the telephone: Complexity measures. Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin, September. 1415–1418.

Voiers, W. D. (1983) Evaluating processed speech using the Diagnostic Rhyme Test. *Speech Technology* 1:338–352.

Voiers, W. D., Sharpley, A. D. & Hehmsoth, C. J. (1975) Research on diagnostic evaluation of speech intelligibility. Research Report AFCRL-72-0694. Bedford, Massachusetts: Air Force Cambridge Research Laboratories.

Vroomen, J., Collier, R. & Mozziconacci, S. (1993) Duration and intonation in emotional speech. Proceedings Eurospeech '93, Berlin, vol. 1. 577–580.

Wahlster, W. (1993) Verbmobil, translation of face-to-face dialogs. Proc. Eurospeech '93, opening and plenary sessions, Berlin. 29–38.

Waibel, A. (1988) Prosody and speech recognition. Research Notes in Artificial Intelligence. London: Pitman Publishing.

Waibel, A., Jain, A. N., McNair, A. E., Saito, H., Hauptmann, A. & Tebelskis, J. (1991) A speech-to-speech translation system using connectionist and symbolic processing strategies. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91. 793–796.

Waibel, A. & Lee, K.-F., eds. (1990) *Readings in speech recognition*. San Mateo, California: Morgan Kaufmann Publishers.

Walker, D., Zampolli, A. & Calzolari, N., eds. (1988) *Automating the lexicon: Research and practice in a multilingual environment*. Cambridge: Cambridge University Press.

Wall, L. & Schwartz, R. L. (1991) *Programming perl*. Sebastopol, CA: O'Reilly & Associates Inc..

Webber, B. L. & Nilsson, N. J., eds. (1981) *Readings in Artificial Intelligence*. Palo Alto, CA: Tioga.

Wells, J. C. (1987) Computer-coded phonetic transcription. *Journal of the International Phonetic Association* 17(2):94–114.

Wells, J. C. (1989) Computer-coded phonemic notation of individual languages of the European Community. *Journal of the International Phonetic Association* 19(1):31–54.

Wells, J. C. (1993a) An update on SAMPA. In: SAM-A (ed.) *SAM-A periodic progress report*. London: ESPRIT Project 6819 (SAM-A) Speech technology assessment in multilingual applications, Year 1, 1 April 1993–30 September 1993. 1–6. Document No: SAM-A/D1-Appendix A.

Wells, J. C. (1993b) Applying SAM-PA to Spanish, Portuguese, and Greek: A preliminary discussion document. In: SAM-A (ed.) *SAM-A periodic progress report*. London: ESPRIT Project 6819 (SAM-A) Speech technology assessment in multilingual applications, Year 1, 1 April 1993–30 September 1993. Document No: SAM-A/D1-Appendix B.

Whittaker, S. & Stenton, P. (1989) User studies and the design of natural language systems. Proceedings of the 4th conference of the European Chapter of the Association for Computational Linguistics, Manchester. 116–123.

Willems, N., Collier, R. & Hart, J. 't (1988) Synthesis scheme for British English intonation. *Journal of the Acoustical Society of America, JASA* 84:1250–1261.

Winer, B. J. (1971) *Statistical principles in experimental design*. Second edition. New York: McGraw-Hill.

Woods, W. A. & Zue, V. (1976) Dictionary expansion via phonological rules for a speech understanding system . IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Philadelphia. 561–564.

Wooffitt, R. C. & Fraser, N. M. (1992) We're off to ring the Wizard, the wonderful Wizard of Oz. In: Button, G. (ed.) *Technology in Working Order: Studies of work, interaction and technology*. London: Routeledge. 211–230.

Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C. P., Sloboda, T., Tomita, M., Tsutsumi, J., Waibel, N. & Ward, W. (1993) Recent advances in Janus: A speech translation system. Proc. of a Workshop: Human Language Technology, 21–24 March, Princeton, NJ. 211–216.

Yarowsky, D. (1991) Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. COLING 92.

Yarrington, D. & Foulds, R. (1993) Personalizing synthesized voices. Proc. ESCA Workshop on Speech and Language Technologies for Disabled Persons, Stockholm. 169–172.

Young, S. R., Hauptmann, A. G., Ward, W. H., Smith, E. T. & Werner, P. (1989) High level knowledge sources in usable speech recognition systems. *Communications of the ACM* 32(2):183–194. Also in: A. Waibel & K.-F. Lee, eds., (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 538–549.

Zue, V., Glass, J., Goodine, D., Hirschman, L., Leung, H., Phillips, M., Polifroni, J. & Seneff, S. (1991) The MIT ATIS system: Preliminary development, spontaneous speech data collection, and performance evaluation. Proceedings of the 2nd European Conference on Speech Communication and Technology, Genova, September. 537–540.

Appendix A

Computer readable alphabets

Appendix B

SAMPA computer readable phonetic alphabet

B.1 SAMPA

B.1.1 Introduction

SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machine-readable phonetic alphabet originally developed under the ESPRIT project 1541 (SAM) in 1987–89 by an international group of phoneticians and applied in the first instance to Danish, Dutch, English, French, German and Italian (SAM 1988, 1989). It has since been extended to other languages, including Norwegian, Swedish, Spanish, Portuguese and Greek.

Section A.1.2 covers the present status of SAMPA, and addresses the languages Danish, Dutch, English, French, German, Italian, Norwegian, Portuguese, Spanish, Swedish and Greek. Section A.1.3 discusses additional levels of annotation. A.1.4 addresses important issues to be considered in the relationship of SAMPA to other computer-coded phonetic transcription systems in use in the world. The IPA convention in Kiel, August 19–21, 1989, presented an opportunity to assess the situation.

B.1.1.1 Notation issues

As with any phonetic notation system, those who developed and applied SAMPA had to make decisions on issues of two types: transcription and coding (see discussion in Wells 1987). The first involves the selection of an appropriate phonetic symbol set; the second involves the allocation of an ASCII number to each symbol that we need, and therefore of a corresponding visual symbol chosen from the character set available on computers.

B.1.1.2 Transcription

Transcription involves many issues of principle over which phoneticians and linguists have debated for decades. These issues may be new, though, to many engineers and speech technologists. Among such issues are (i) whether the notation should be phonemic, or to some extent allophonic; and if phonemic, how the phoneme set is to be established; (ii) to what extent phonetic symbols should be required to have the same meaning across different languages; and (iii), the relation between the basic, lexical, pronunciation of a word and its actual pronunciation in context.

In principle, SAMPA provides for phonemic notation of languages. For example, the r-sounds of English *rip*, *trip*, and *drip* are all instances of the phoneme /r/, although different articulatory and acoustically (in voicing and in presence/absence of friction). These different allophones are predictable from the phonetic context: we can unambiguously write them all as /r/. The arguments for preferring phonemic notation to allophonic are (i) it is simpler while still being unambiguous; (ii) correct identification of allophones may be difficult for those without phonetic training; and (iii) too few codes are available in the range 32–127 to provide for all allophones.

In syllable-initial position, English /t/ is alveolar and aspirated; French /t/, dental and unaspirated; Swedish /t/, dental and aspirated. We ignore these comparative differences in our notation, writing all as /t/. SAMPA does not need to adopt distinct symbols to reflect these differences. (However, if and when SAMPA is applied to Hindi, for example, where these differences are phonemic, it would become necessary to notate them explicitly.)

In continuous speech the actual sounds used in pronouncing a word may well differ from the word's citation form (dictionary entry). A phonotypical transcription is one in which citation forms are modified in accordance with known phonetic rules of connected speech. For example, in a phonotypical transcription of English, final linking /r/ would be shown before a following vowel (*better ask*) but not before a consonant (*better go*); the lexical entry would be invariant. In an actual utterance the speaker might or might not conform to phonotypical expectations; an impressionistic transcription reflects a human (or mechanical) auditory or acoustic analysis of what was actually said. In the case at issue, /r/ would be shown if phonetically present in a given instance, not otherwise.

In practice, colleagues working on the various languages to which SAMPA has been applied have chosen to deviate in various respects from these principles. English has plosive /d/ and fricative /ð/ (SAMPA /D/) as distinct phonemes (*den*, *then*). In Spanish, they are undoubtedly allophones of the same phoneme, and could unambiguously both be written /d/; but for speech technology work our Spanish colleagues prefer to notate them distinctively, as “d” and “D” respectively. The r-sounds in French *rouge*, *lettre* are different from all the English r-sounds, being respectively a voiced and voiceless uvular fricative. It would seem unambiguous and logical to write them, too, as /r/. But our French colleagues have preferred to use the distinct uvular-r symbol, also provided in SAMPA, namely /R/.

Nevertheless I believe we should as far as possible discourage allophonic and comparative notation. Bulgarian has the simple 6-vowel system, IPA /i e a o u ə/. A colleague in Bulgaria has proposed that they be represented in SAMPA as /I, E, a, O, U, @/. About /a/ and /@/ (= IPA /ə/) we can agree. But the other symbols he proposes are inappropriately comparative. The Bulgarian vowels should appear in SAMPA as /i, e, a, o, u, @/.

B.1.1.3 Coding

SAMPA's coding principles involve restricting the available ASCII codes to the range 32–127. At the time SAMPA was formulated, many computers used only the 7-bit ASCII character set. With the spread of PCs and compatibles, the “extended ASCII” (8-bit) set has become familiar, allowing codes in the range 128–255. Has the decision to restrict SAMPA to the range 32–127 proved wise? Or should we now relax it?

In the (American) English extended ASCII character set used by PCs running MS-DOS, the range 128–255 is used to provide for the screen and printer a number of accented alphabetic letters, currency symbols, graphic symbols, and Greek and mathematical symbols. Those that are not available on the keyboard can be accessed by entering their ASCII number on the keypad while pressing the Alt key. Unfortunately, from the point of view of non-English-speaking Europeans, this extended ASCII fails to provide all the accented Latin letters needed for such languages as Portuguese, Icelandic, Czech, Polish and Hungarian. To remedy this shortcoming, a number of different “code pages” are now available, each providing a different set of characters in the 128–255 range. In the USA and the UK most PCs use code page 437 (International English), in Western Europe 850 (Multilingual Latin I), and in much of Eastern Europe 852 (Slavic Latin II).

Applications running under the popular front-end Windows use yet another character set, one known as "enhanced ANSI". This is identical with the ASCII set for 33–127; for 128–255 it offers its own specific choice of accented alphabetic and other characters, with codes different from ASCII.

The consequence is that in PC-compatible computing the code numbers in the range 128–255 (the "extended" characters) may currently have several different interpretations. Conversely, a given character may be coded in several different ways.

Consider the IPA symbols /æ/ and /ð/, both needed for the phonetic transcription of English. For reasons that seemed valid at the time (cf. Wells (1987: 95)), SAMPA assigned the former the code 123, which now appears on all Latin-alphabet PC screens as "{"; the latter was coded 68, "D".

Both "æ" and "ð" are now available on-screen for PCs running Windows. While "æ" is an ASCII character, with the extended code 145 (for those using code page 437 or 850), "ð" is not. But both are in the enhanced ANSI set, with codes 230 and 240 respectively. (Hence under Windows they can be accessed, if not on the keyboard, by keying Alt+0230 and Alt+0240; "æ" can also be accessed as Alt+145.)

However, a PC using code page 852 (Slavic) will display code 145 as an upper-case Polish L-with-acute-accent (Ĺ), 230 as "Š", and 240 as a dieresis (¨). With code page 860 (Portugal), 145 is "À", 230 "μ" and 240 "≡".

Recently a number of phonetic fonts have become available for use under Windows. These comprise only phonetic symbols (perhaps with a few punctuation signs). Unfortunately they disagree extensively on key assignment and coding. On my PC I now have three TrueType phonetic fonts provided by the Summer Institute of Linguistics and four others of whose origins, I regret to say, I am uncertain. These fonts agree with SAMPA (but not ANSI) in assigning "ð" to code 68/D; but for "æ" they assign codes and keystrokes 81/Q (SIL Doulos/Manuscript/Sophia IPA), 60/< (Times IPA New), 64/@ (Tech Phonetic), and 233 (IPA Roman 1, IPA Plus).

B.1.1.4 Further languages

A number of other EC languages have been examined in the light of the SAMPA recommendations, and a short summary of the possible solutions for their special features is given here. For more details, see J. Wells, "Computer-coded phonetic transcription", Journal of the International Phonetic Association 17, No. 2, pp. 94–114, and the SAM Definition Phase Final Report (ESPRIT project 1541), January 1988.

Most of the minority languages of Europe such as Basque, Breton, Catalan, and Frisian can be transcribed adequately at a phonemic level without the need to change the principles of the present recommendation. Irish and Scottish Gaelic, however, require a decision for coding the palatalised (or "slender") consonants and the "double" nasals and laterals. Scottish Gaelic also has a back unrounded vowel series which does not occur in other EC languages. Welsh requires a solution for the voiceless alveolar lateral, represented in the orthography as "<ll>".

We should like now to explore whether it would be suitable to extend SAMPA for application to other languages, including Chinese, and if so how.

The question of Chinese has arisen because of the prospect of a wider collaboration on speech research between University College London and the Chinese Academy of Sciences.

Chinese already has what appears to be a satisfactory machine-readable phonetic notation in the form of Pinyin, the romanisation that has for some years been standard in the People's Republic (though not in Taiwan). Pinyin is an ingenious quasi-phonemic notation. It includes a number of unconventional digraphs, together with unconventional uses of individual Latin letters. Thus *sh, ch*, and *zh* represent retroflex/postalveolar consonants of a type that would normally be written in SAMPA as [S, tS, dZ]. Pinyin *x, q, j* represent a corresponding series of alveolopalatal consonants, IPA [ɕ, tɕ, dʑ], for which SAMPA does not currently cater. Pinyin *c* represents [ts], *y* [j], and *ng* [ŋ]. The close front rounded vowel [y] is written *u* where there would be no confusion with [u], but *ü* where this confusion might arise. (This last Pinyin character is not actually machine-readable in our sense.)

Continuing to use Pinyin for Chinese but SAMPA for other languages would mean that characters such as "x, j" would have different meanings in different languages ("x" = alveolopalatal fricative, or velar fricative; "j" = alveolopalatal affricate, or palatal approximant). But this is perhaps no worse than the "comparative" differences already present in the interpretation of some symbols (see above). The Pinyin notation "i" already covers a remarkable range of allophonic possibilities (including an r-coloured back vowel in *shi* and a slightly fricative central vowel in *si*). Are Chinese speech technologists happy with this degree of phonemic abstraction?

Tone is shown in Pinyin (if indeed it is shown) by superscript accent marks, thus *mā, má, mǎ, mà*. These are not machine-readable in the SAMPA sense. The corresponding SAMPA tone-marks would be /"ma, 'ma, ``'ma, `ma/. However these SAMPA signs have not proved popular, and perhaps ought to be changed. For Chinese, we could perhaps consider instead the use of numerals, thus "ma1, ma2, ma3, ma4".

B.1.2 SAMPA: Present status

The following table presents the system agreed among the representatives of eight European countries engaged in European collaborative speech technology assessment research (SAM). It is currently being tested in the transcription and labelling of European multi-lingual databases.

SAMPA computer readable phoneme alphabet for European languages, with ASCII and IPA definitions (1990):

Table 1: Consonants

IPA	SAMPA	ASCII	Voicing	Place	Manner
b	b	98	voiced	bilabial	plosive
c	c	99	voiceless	palatal	plosive
ç	C	67	voiceless	palatal	fricative
d	d	100	voiced	dental/alveol.	plosive
ð	D	68	voiced	dental	fricative
f	f	102	voiceless	labiodental	fricative
g	g	103	voiced	velar	plosive
ɣ	G	71	voiced	velar	fricative
h	h	104	voiceless	glottal	fricative
j	j	106		palatal	approximant
k	k	107	voiceless	velar	plosive
l	l	108		dental/alveol.	lateral approximant
ʎ	L	76		palatal	lateral approximant
m	m	109		bilabial	nasal
n	n	110		dental/alveol.	nasal
ɲ	J	74		palatal	nasal
ŋ	N	78		velar	nasal
p	p	112	voiceless	bilabial	plosive
r	r	114		alveolar	trill
ʁ, ʀ	R	82		uvular	trill/fricative
s	s	115	voiceless	alveolar	fricative
ʃ	S	83	voiceless	postalveolar	fricative
t	t	116	voiceless	dental/alveol.	plosive
θ	T	84	voiceless	dental	fricative
v	v	118	voiced	labiodental	fricative
w	w	119		labial-velar	approximant
x	x	120	voiceless	velar	fricative
ɥ	H	72		labial-palat.	approximant
z	z	122	voiced	alveolar	fricative
ʒ	Z	90	voiced	postalveolar	fricative
ʔ	?	63		stød, glottal stop	

Table 2: Boundary and prosodic features

IPA	SAMPA	ASCII	Comment
ː	:	58	Length mark
ˈpapa	”papa	34	Primary stress (also in Accent I words in Norwegian and Swedish)
ˇpapa	””papa	34,34	Accent II words in Norwegian and Swedish
ˌpapa	%papa	37	Secondary stress
¯papa	-papa	45	Level tone (if followed by a tone group boundary)
	’papa	39	Rising tone
	`papa	96	Falling tone
	`’papa	96,39	Fall-rise
	’`papa	39,96	Rise-fall
no equiv.	$	36	Syllable boundary
no equiv.	|	124	Tone group boundary
no equiv.	–	45	Separator

Table 3: Vowels

IPA	SAMPA	ASCII	Open–Close	Front–Back	Rounded–Unrounded	Comment
a	a	97	open	front	unrounded	
ɑ	A	65	open	back	unrounded	
æ	{	123	near-open	front	unrounded	E bad
ɐ	6	54	near-open	central	unrounded	G Butter
ɒ	Q	81	open	back	rounded	
ɔ	O	79	open-mid	back	rounded	
e	e	101	close-mid	front	unrounded	
ɛ	E	69	open-mid	front	unrounded	
ə	@	64	mid	central	unrounded	Schwa
ɜ	3	51	mid	central	unrounded	
i	i	105	close	front	unrounded	
ɪ	I	73	near-close	front	unrounded	lax
o	o	111	close-mid	back	rounded	
ø	2	50	close-mid	front	rounded	
œ	9	57	open-mid	front	rounded	
ɶ	&	38	open	front	rounded	
u	u	117	close	back	rounded	
ʊ	U	85	near-close	back	rounded	lax
ʉ	}	125	close	central	rounded	
ʌ	V	86	open-mid	back	unrounded	
y	y	121	close	front	rounded	
ʏ	Y	89	near-close	front	rounded	lax

Table 4: Two character symbols

IPA	SAMPA	ASCII	Comment
	e~	69,126	Nasalised vowel
aɪ	aI	97,73	Diphthong
tʃ	tS	116,83	Affricate
ɖ	rd	114,100	Retroflex consonant
ɵ	u0	117,48	Mid-close rounded central vowel in Swedish
n̩	=n	61,110	Syllabic /n/
	E/	69,47	Indeterminacy (for /e/~/E/ in French)

Table 5: Currently under discussion

SAMPA	ASCII	Comment
*	42	Conjunctor
...	46,46,46	Silent pause
\	92	Phonetic case shift (e.g. *F might be used to signal a shift into French and \ would terminate the shift.

Table 6: Currently used in French work

SAMPA	ASCII	Comment
&		Phonological phrase
#	35	Word boundary
##	35,35	Absence of liaison
+	43	Morpheme boundary

B.1.3 The phonemic notation of individual languages

This section provides a brief outline of the phonemic distinctions in the languages of the eight countries engaged in the initial phase of the SAM project by providing example words for the use of each phonemic symbol. Information is also provided for the languages of Spanish, Portuguese and Greek considered additionally.

B.1.3.1 Danish

Consonants

The plosives are /p, b, t, d, k, g/:

Symbol	**Word**	**Transcription**
p	pande	"pan@
b	bande	"ban@
t	tand	tan?
d	dan	dan?
k	kalde	"kal@
g	galde	"gal@

The fricatives are /f, s/:

f	finde	”fen@
s	stand	sdan?

The approximants are /v, D, j, h/:

v	vinde	”ven@
D	bide	”bi:D@
j	Jul	ju:?l
h	hest	hEsd

The nasals are /m, n, N/:

m	mile	”mi:l@
n	ny	ny?
N	lunge	”loN@

The liquids are /l, R/:

l	land	lan?
R	ride	”Ri:D@

Stød is symbolised by “?” and may be found in syllables containing a long stressed vowel, or a short stressed vowel, or a short stressed vowel followed by a voiced consonant, e.g. *pæu* — /pE:?u/, *peu* — /pEu?/

Vowels

The vowel system chosen for broad phonetic transcription is: /i, e, E, a, A, y, 2, 9, u, o, O, @/, with all vowels except @ occurring with a length distinction: /i:, e:, E:, a:, A:, y:, 2:, 9:, u:, o:, O:/.

The unrounded front vowels are exemplified in the following:

Symbol	**Word**	**Transcription**
i:	mile	”mi:l@
i	ville	”vil@
e:	mele	”me:l@
e	visse	”ves@
E:	mæle	”mE:l@
E	tække	”tEk@
{:	male	”m{:l@
a	malle	”mal@

The central vowels are:

A:	parken	”PA:g=n	
A	pakken	pAg=n	
@			unstressed final *e*

The rounded front vowels are:

y:	hyle	"hy:l@
y	hylde	"hyl@
2:	køle	"k2:l@
2	kølle	"k2l@
9:	høne	"h9:n@
9	hønse	"h9ns@

The back vowels are:

u:	kule	"ku:l@
u	kulde	"kul@
o:	fole	"fo:l@
o	foto	"fodo
O:	måne	"mO:n@
O	munde	"mOn@
Q:	kåre	kQ:@
Q	kors	kQs

Diphthongs.

The falling diphthongs may be most economically analysed phonemically as vowel plus /j/, /v/, or /r/, but for the broad phonetic representation within SAMPA they are analysed as vowel plus /i/, /u/ or /Q/, for example:

Symbol	Word	Transcription	Comment
ai	sejle	"sail@	
Qi	søjle	"sQil@	
ui	huje	"hui@	(extremely rare)
i:i	pige	"pi:i@	
e:i	ege	"e:i@	conservative pronunciation
E:i	læge	"lE:i@	
iu	ivrig	"iuRi	
eu	peber	"peu@R	
Eu	evne	"Eun@	
i:u	hive	"hi:u	
e:u	leve	"he:u	
E:u	hæve	"hE:u	
au	havn	hau?n	
yu	syv	syu	
2u	døvstum	"d2usdom	
9u	støvle	"st9ul@	
iQ	kirke	"kiQk@	
EQ	bær	bEQ	
9Q	smør	sm9Q	

B.1.3.2 Dutch

Consonants

The plosives are /p, b, t, d, k/, (/g/):

Symbol	**Word**	**Transcription**	**Comment**
p	pak	pAk	
b	bak	bAk	
t	tak	tAk	
d	dak	dAk	
k	kap	kAp	
g	goal	go:l	(only foreign/loanwords)

The fricatives are /f, v, s, z, x, h/, (/G/):

f	fel	fEl
v	vel	vEl
s	sein	sEin
z	zijn	zEin
x	toch	tOx
G	goed	Gut (also: xut)
h	hand	hAnt
Z	bagage	bAga:Z(@)
S	show	So:u

The sonorants (nasals, liquids and glides) are /m, n, N, l, r, w, j/:

m	met	mEt
n	net	nEt
N	bang	bAN
l	land	lAnt
r	rand	rAnt
w	wit	wIt
j	ja	ja:

Vowels

The Dutch vowels fall into two classes, "checked" (not occurring in a stressed syllable without a following consonant) and "free".

The checked vowels are /I, E, A, O, Y, @/:

I	pit	pIt
E	pet	pEt
A	pat	pAt
O	pot	pOt
Y	put	pYt
@	gemakkelijk	G@"mAk@l@k

The free vowels comprise four monophthongs /i, y, u, a:/, three "potential diphthongs" /e:, 2:, o:/, and three "essential diphthongs", /Ei, 9y, Au/, exemplified as follows:

Symbol	Word	Transcription	Comment
i	vier	vir	
y	vuur	vyr	
u	voer	vur	
a:	naam	na:m	
e:	veer	ve:r	
2:	deur	d2:r	
o:	voor	vo:r	
Ei	fijn	fEin	
9y	huis	h9ys	
Au	goud	xAut	

There are also six vowel sequences which are sometimes described as diphthongs:

a:i	draai	"dra:i
o:i	mooi	"mo:i
ui	roeiboot	"ruibo:t
iu	nieuw	"niu
yu	duw	"dyu
e:u	sneeuw	"sne:u

Several marginal vowel phonemes are only found in loanwords:

E:	crème	krE:m
9:	freule	fr9:l@
O:	roze	rO:z@

B.1.3.3 English

Consonants

There are six plosives /p, b, t, d, k, g/:

Symbol	Word	Transcription
p	pin	pIn
b	bin	bIn
t	tin	tIn
d	din	dIn
k	kin	kIn
g	give	gIv

There are two phonemic affricates /tS/ and /dZ/:

tS	chin	tSIn
dZ	gin	dZIn

There are nine fricatives /f, v, T, D, s, z, S, Z, h/:

f	fin	fIn
v	vim	vIm
T	thin	TIn
D	this	DIs
s	sin	sIn
z	zing	zIN
S	shin	SIn
Z	measure	"meZ@
h	hit	hIt

The sonorants are three nasals /m, n, N/, two liquids /r, l/ and two sonorant glides /w, j/:

m	mock	mQk
n	knock	nQk
N	thing	TIN
r	wrong	rQN
l	long	lQN
w	wasp	wQsp
j	yacht	jQt

Vowels

The English vowels fall into two classes, traditionally known as "short" and "long" but better described as "checked" (not occurring in a stressed syllable without a following consonant) and "free".

The checked vowels are /I, e, {, Q, V, U/:

I	pit	pIt
e	pet	pet
{	pat	p{t
Q	pot	pQt
V	cut	kVt
U	put	pUt

There is a short central vowel, normally unstressed:

@	another	@nVD@

The free vowels comprise monophthongs and diphthongs, although no hard and fast line can be drawn between these categories. They can be placed in three groups according to their final quality /i:, eI, aI, OI/, /u:, @U, aU/, /3:, A:, O:, I@, e@, U@/. They are exemplified as follows:

i:	ease	i:z
eI	raise	reIz
aI	rise	raIz
OI	noise	nOIz
u:	lose	lu:z
@U	nose	n@Uz
aU	rouse	raUz
3:	furs	f3:z
A:	stars	stA:z
O:	cause	kO:z
e@	stairs	ste@z
U@	cures	kjU@z

The vowels /i:/ and /u:/ in unstressed syllables vary in their pronunciation between a close [i] and a more open [i] (close [u] — more open [u]). Therefore, it is suggested that /i/ and /u/ be used as indeterminacy symbols.

i	happy	h{pi
u	into	Intu

B.1.3.4 French

Consonants

There are six plosives /p, b, t, d, k, g/:

Symbol	**Word**	**Transcription**
p	pont	po~
b	bon	bo~
t	temps	ta~
d	dans	da~
k	quand	ka~
g	gant	ga~

There are seven fricatives /f, v, s, z, S, Z, j/. /j/ can be realised as a fricative or an approximant.

f	femme	fam
v	vent	va~
s	sans	sa~
z	zone	zon
S	champ	Sa~
Z	gens	Za~
j	ion	jo~

There are four nasals /m, n, J, N/, the last of which is only found in loanwords:

m	mont	mo~
n	nom	no~
J	oignon	oJo~
N	camping	ka~piN

There are two liquids /l, R/ and two vowel glides /w, H/ (besides /j/):

l	long	lo~
R	rond	Ro~
w	coin	kwe~
H	juin	ZHe~

Vowels

The vowel system comprises 12 oral vowels /i, e, E, a, A, O, o, u, y, 2, 9, @/, and 4 nasal vowels /e~, a~, o~, 9~/, exemplified as follows:

i	si	si
e	ses	se
E	seize	sEz
a	patte	pat
A	pâte	pAt
O	comme	kOm
o	gros	gro
u	doux	du
y	du	dy
2	deux	d2
9	neuf	n9f
@	justement	Zyst@ma~
e~	vin	ve~
a~	vent	va~
o~	bon	bo~
9~	brun	br9~

When they are functional, the load of the oppositions /a/–/A/, /e~/–/9~/, /e/–/E/, /o/–/O/, /2/–/9/ may be very low for certain speakers, and there is a tendency towards neutralisation. When they are not functional there is a strong tendency in unstressed syllables towards indetermination. "Indeterminacy" symbols have been agreed to cover occurrences of these phonemes or sounds

E/ = e or E
A/ = a or A
&/ = 2 or 9
O/ = o or O
U~/ = e~ or 9~

B.1.3.5 German

Consonants

There are six plosives /p, b, t, d, k, g/:

Symbol	Word	Transcription
p	Pein	paIn
b	Bein	baIn
t	Teich	taIC
d	Deich	daIC
k	Kunst	kUnst
g	Gunst	gUnst

There are four phonemic affricates /pf, ts, tS/ and /dZ/, which occur in a few loanwords:

pf	Pfahl	pfa:l
ts	Zahl	tsa:l
tS	deutsch	dOYtS
dZ	Dschungel	”dZUN@l

There are ten fricatives /f, v, s, z, S, Z, C, j, x, h/. /j/ is often realised as a vowel glide.

f	fast	fast
v	was	vas
s	Tasse	”tas@
z	Hase	”ha:z@
S	waschen	”vaS@n
Z	Genie	Ze”ni:
C	sicher	”zIC6
j	Jahr	ja:6
x	Buch	bu:x
h	Hand	hant

The sonorants are three nasals /m, n, N/, and two liquids /l, r/:

m	mein	maIn
n	nein	naIn
N	Ding	dIN
l	Lein	laIn
r	rein	raIn

Orthographic <r> is realised phonetically in a number of different ways:

1. As a dorso-uvular consonant — a voiced or voiceless fricative, approximant, trill or flap. This should be represented as ‘R’.
2. As an apico-alveolar consonant — a trill, tap, or flap. This may be represented as ‘r’ (e.g. <rein> — raIn).
3. As a vowel post-vocalically. This may be represented as ‘6’ (see below).

Vowels

The vowels fall into three groups, “checked” (short), “free” (long), and two short vowels that only occur in unstressed position. The checked vowels are /I, E, a, O, U, Y, 9/:

I	Sitz	zIts
E	Gesetz	g@"zEts
a	Satz	zats
O	Trotz	trOts
U	Schutz	SUts
Y	hübsch	hYpS
9	plötzlich	"pl9tslIC

There are 8 pure free vowels are /i:, e:, E:, a:, o:, u:, y:, 2:/ and three free diphthongs /aI, aU, OY/:

i:	Lied	li:t
e:	Beet	be:t
E:	spät	SpE:t
a:	Tat	ta:t
o:	rot	ro:t
u:	Blut	blu:t
y:	süss	zy:s
2:	blöd	bl2:t

aI	Eis	aIs
aU	Haus	haUs
OY	Kreuz	krOYts

The unstressed "schwa" vowel is:

@	bitte	"bIt@

The vowel realisation of <r>, represented as 6, fuses with schwa, but it also follows stressed vowels, resulting in additional centring diphthongs:

6	besser	"bEs6
i:6	Tier	ti:6
I6	Wirt	vI6t
y:6	Tür	ty:6
Y6	Türke	"tY6k@
e:6	schwer	Sve:6
E:6	Bär	bE:6
E6	Berg	bE6k
2:6	Föhr	f2:6
96	Wörter	"v96t6
a:6	Jahr	ja:6
a6	hart	ha6t
u:6	Kur	ku:6
U6	kurz	kU6ts
o:6	Ohr	o:6
O6	dort	dO6t

B.1.3.6 Greek

Consonants:

plosives

Symbol	Word	Transcription	Comments
p	*πίνω*	"pino	
b	*μπορώ*	bo"ro	
t	*τότε*	"tote	
d	*ντύνω*	"dino	
k	*κάνω*	"kano	
g	*γκρεμίζω*	gre"mizo	

affricates

ts	*τσέπη*	"tsepi	
dz	*τζαμπατζής*	dzaba"dzis	

fricatives

Symbol	Word	Transcription	Comments
f	*φέυγω*	"fevGo	
v	*βράδυ*	"vraDi	
T	*θέλω*	"Telo	
D	*δρόμος*	"Dromos	
s	*σπίτι*	"spiti	
z	*ζέστη*	"zesti	
x	*χρόνος*	"xronos	
G	*γάλα*	"Gala	

nasals

m	*μιλώ*	mi"lo	
n	*νερό*	ne"ro	
	αγγούρι	an"guri	([N] = allophone of /n/)

liquids

l	*λουλούδι*	lu"luDi	
r	*ρούχα*	"ruxa	

semivowel

j	*διαβάζω*	Dja"vazo	

(palatals)

c	*καί*	ce	
	κοιμάμαι	ci"mame	
	σκιάζω	"scazo	
gj	*άγγελος*	"angjelos	
	γκίνια	"gjinja	
	γκιουβέτσι	gju"vetsi	
C	*χέρι*	"Ceri	
	χειμώνας	Ci"monas	
	χιόνι	"Coni	
jj	*γέρος*	"jjeros	
	γύρος	"jjiros	
	γιός/υιός	jjos	

connected speech phenomena

δέν πειράζει	(Den pi"razi)	De bi"razi
o *γιός μου*	(o "jjos mu)	o "jjoz mu

B.1.3.7 Italian

Consonants

There are six single and six geminate plosives /p, b, t, d, k, g/, /pp, bb, tt, dd, kk, gg/ as follows:

Symbol	**Word**	**Transcription**
p	pane	"pane
b	banco	"banko
t	tana	"tana
d	danno	"danno
k	cane	"kane
g	gamba	"gamba
pp	coppa	"kOppa
bb	gobba	"gObba
tt	zitto	"tsitto
dd	cadde	"kadde
kk	nocca	"nOkka
gg	fugga	"fugga

There are four single and four geminate affricates /ts, dz, tS, dZ/, /tts, ddz, ttS, ddZ/:

ts	zitto	"tsitto
dz	zona	"dzOna
tS	cena	"tSena
dZ	gita	"dZita
tts	bozza	"bOttsa
ddz	mezzo	"mEddzo
ttS	braccio	"brattSo
ddZ	oggi	"OddZi

There are five single and four geminate fricatives /f, v, s, z, S/, /ff, vv, ss, SS/:

f	fame	"fame	
v	vano	"vano	
s	sano	"sano	
z	sbaglio	"zbaLLo	(only initial in cluster)
S	scendo	"Sendo	
ff	beffa	"bEffa	
vv	bevvi	"bevvi	
ss	cassa	"kassa	
SS	ascia	"aSSa	

There are three single and geminate nasals /m, n, J/, /mm, nn, JJ/, three single and three geminate liquids /r, l, L/, /rr, ll, LL/ and two semi-vowels /j, w/:

m	molla	"mOlla	
n	nocca	"nOkka	
J	gnocco	"JOkko	
mm	grammo	"grammo	
nn	panna	"panna	
JJ	aglio	"aJJo	
r	rete	"rete	
l	lama	"lama	
L	gli	"Li	(This is a function word; otherwise *gli* does not occur in word initial position.)
rr	ferro	"fErro	
ll	colla	"kOlla	
LL	foglia	"fOLLa	
j	ieri	"jeri	(initial and initial clusters)
w	uomo	"wOmo	(initial and initial clusters)

Vowels

The vowel system comprises seven vowels /i, e, E, a, O, o, u/:

i	mite	"mite
e	rete	"rete
E	meta	"mEta
a	rata	"rata
O	moto	"mOto
o	dove	"dove
u	muto	"muto

B.1.3.8 Norwegian

There are six plosives:

Symbol	**Word**	**Transcription**	**Comment**
p	hopp	hOp	
b	labb	lAb	
t	lat	lA:t	
d	ladd	lAd	
k	takk	tAk	
g	tagg	tAg	

There are six fricatives:

f	fin	fi:n	
v	vin	vi:n	
s	lass	las	
S	skyt	Sy:t	
C	kino	Ci:nu	(not syllable-final)
j	gi	ji:	
h	ha	hA	

There are five sonorant consonants (nasals, liquids, trills):

m	lam	lAm	
n	vann	vAn	
N	sang	sAN	
l	fall	fAl	
r	prøv	pr2:v	(not syllable-final)

Vowels

There are 9 long vowels:

i:	vin	vi:n
e:	sen	se:n
{:	vær	v{:r
A:	hat	hA:t
y:	lyn	ly:n
2:	søt	s2:t
O:	våt	vO:t
u:	bok	bu:k
}:	lun	l}:n

and nine short vowels:

i	vind	vin
e	send	sen
{	vært	v{rt
A	hatt	hAt
y	lynne	lyne
2	søtt	s2t
O	vått	vOt
u	bukk	buk
}	lund	l}n

There are seven diphthongs:

{i	vei	v{i
2y	høy	h2y
A}	sau	sA}
Ai	kai	kAi
Oy	konvoy	ku:nvOy
}i	hui	h}i
ui	hoi	hui

In addition there are important allophonic variants for which the transcription has been agreed:

rt	hardt	hArt	(retroflex t)
rd	verdi	v{rdi:	(retroflex d)
rl	ærlig	{:rlig	(retroflex l)
rn	garn	gA:rn	(retroflex n)
rL	blå	brLO:	(retroflex flap)

In cases where the dental consonants do not change into retroflexes, they are transcribed using the separator sign (ASCII 45), e.g. /r-t/, /r-d/:

r-d	verdig	v{:r-dt

B.1.3.9 Portuguese

Consonants:

plosives

Symbol	**Word**	**Transcription**	**Comments**
p	pai	pai	
b	barco	”barku	
t	tenho	”teJu	
d	doce	”dos@	
k	com	ko	
g	grande	”gr6nd@	

fricatives

f	falo	”falu	
v	verde	”verd@	
s	céu	sEw	
z	casa	”kaza	
S	chapéu	S6”pEw	
Z	jóia	”ZOj6	

nasals

m	mar	mar	
n	nada	”nad6	
J	vinho	”viJu	

liquids

l	lanche	”l6nS@	
L	trabalho	tr6”baLu	
r	caro	”karu	
R	rua	”Ru6	

Vowels and diphthongs

i	vinte	”vint@	”lapiS lápis
e	fazer	f6”zer	
E	belo	”bElu	
a	falo	”falu	
6	cama	”k6m6	m6”d6jr6 madeira
O	ontem	”Ont6 j	
o	lobo	”lobu	
u	jus	ZuS	fu”turu futuro
@	felizes	f@”liz@S	
i	fim	fi	
e	emprego	e ”pregu	(or = /em/)
6	irm	ir”m6	
o	bom	bo	
u	um	u	
aw	mau	maw	etc.: iw, ew, Ew, (ou)
aj	mais	majS	etc.: ej, Ej, Oj, oj,
6 j	têm	t6 j	etc.: e j , o j , u j

B.1.3.10 Spanish

Consonants:

plosives

Symbol	Word	Transcription	Comments
p	padre	"paDre	
b	vino	"bino	
t	tomo	"tomo	
d	donde	"donde	
g	gata	"gata	
affricates			
tS	mucho	"mutSo	
jj	hielo	"jjelo	
fricatives			
f	fácil	"faTil	
B	cabra	"kaBra	(= /b/)
T	cinco	"Tinko	
D	nada	"naDa	(= /d/)
s	sala	"sala	
x	mujer	mu"xer	
G	luego	"lweGo	(= /g/)
nasals			
m	mismo	"mismo	
n	nunca	"nunka	
J	año	"aJo	
liquids			
l	lejos	"lexos	
L	caballo	ka"baLo	(or as jj)
r	puro	"puro	
rr	torre	"torre	
semivowels			
j	rei	rrej	
	pie	pje	
w	deuda	"dewDa	
	muy	mwi	
Vowels:			
i	pico	"piko	
e	pero	"pero	
a	valle	"baLe	
o	toro	"toro	
u	duro	"duro	

B.1.3.11 Swedish

Consonants

There are six plosives:

Symbol	**Word**	**Transcription**	**Comments**
p	pil	pi:l	
b	bil	bi:l	
t	tal	tA:l	
d	dal	dA:l	
k	kal	kA:l	
g	gås	go:s	

There are six fricatives:

f	fil	fi:l	
v	vår	vo:r	
s	sil	si:l	
S	sjuk	S}:k	(front and back allophones)
h	hal	hA:l	
C	tjock	COk	(not syllable-final)

There are six sonorant consonants (nasals, liquids and semi-vowels):

m	mil	mi:l	
n	nål	no:l	
N	ring	rIN	(not syllable-initial)
r	ris	ri:s	
l	lös	l2:s	
j	jag	jA:g	

Vowels

There are nine long and nine short vowels.

Long vowels (followed by a short consonant):

Symbol	**Word**	**Transcription**	**Comment**
i:	vit	vi:t	
e:	vet	ve:t	
E:	säl	sE:l	
y:	syl	sy:l	
}:	hus	h}:s	
2:	föl	f2:l	
u:	sol	su:l	
o:	hål	ho:l	
A:	hal	hA:l	

Short vowels (followed by a long consonant):

I	vitt	vIt	
e	vett	vet	
E	rätt	rEt	
Y	bytt	bYt	
u0	buss	bu0s	
2	föll	f2l	
U	bott	bUt	
O	håll	hOl	
a	hall	hal	

There are also two pre-r-allophones (long and short) of /E/ and /2/ (see below).

The following important allophonic variants occur in Swedish which require separate symbolic representation:

{:	här	h{:r	pre-r allophones of E:
9:	för	f9:r	pre-r allophones of 2:
{	herr	h{r	pre-r allophones of E
9	förr	f9r	pre-r allophones of 2
@	pojken	pOjk@n	schwa vowel allophone
rt	hjort	jUrt	retroflex consonant, not initial
rd	bord	bu:rd	"
rn	barn	bA:rn	"
rs	fors	fOrs	"
rl	karl	kA:rl	"

B.1.4 Levels of annotation and extension of SAMPA

B.1.4.1 SAMPA as a phonemic system

The present SAMPA system, which was provisionally agreed at the end of the Extension Phase, is defined as a system for phonemic transcription and annotation. This means that the symbols are used according to the analysis of distinctive sound oppositions within each language. Thus, although their relation to sound category symbols of the International Phonetic Alphabet (IPA) is given, they are symbols of intra-language convention, and do not have an exact language-independent phonetic (auditory or acoustic) equivalence, nor do they represent a single sound within a language.

For example, the symbol /t/, used in the transcription of all 8 partner languages, could represent an unaspirated sound in French or Italian, a strongly aspirated sound in German or English, and an affricated sound in Danish. In English the /t/ can also stand for an unaspirated sound (following /s/) or the more usual aspirated sound. Vowel symbols often represent widely diverging sounds from one language to another; /{/ in Danish is very different from /{/ in English, for example.

This basically phonemic, or sound-system-orientated (systematic) function of SAMPA means that a general extension of the SAMPA coding system to allow fine phonetic differentiation of speech sounds is not possible. There are, however, examples in the SAMPA list of symbols which can be used to represent non-distinctive differences within a language, e.g. 'r' and 'R' for regionally dependent free variants, and some important allophonic variants are allowed for (e.g. in Swedish and Norwegian). Also, auditory transcription (French "notation") is meant to be a "broad phonetic" representation of the actual utterance, including elisions and assimilations (inasfar as these can be represented with the phonemically orientated SAMPA inventory) rather than the strictly phonemic string of the citation form.

One area in which an extension of SAMPA is possible, indeed probable, is prosody. Certain "Boundary and Prosodic Features" have been agreed preliminarily, but their use has only been illustrated in the English EUROM.0 transcriptions. The considerations of prosodic description in a multi-lingual context may well reveal the need to modify and extend SAMPA. The work on prosodic description may also conclude that a separate prosodic annotation tier is necessary.

B.1.4.2 Detailed phonetic or acoustic annotation

For finer segmental annotation of speech recordings, three basically different approaches are offered for discussion. All three approaches require a separate annotation tier, but the labels are temporally defined by the location of the phonemic segment boundaries (phonemic markers in the case of centre labelling).

1. The SAMPA symbols are given language-independent sound values (IPA equivalent values) and modified by means of agreed diacritic codes to reflect fine phonetic detail.

   Advantages: No new segmentation or marker placements would be required.

   Disadvantages:

   (a) Different symbols would sometimes be required at the phonemic and phonetic levels, particularly for vowels. E.g. Danish /{/ might have to be represented by phonetic [E]; English /{/ might have to be represented by phonetic [a] or even [A], depending on regional accent.

   (b) Diacritic symbols would have to be agreed for all partner languages.

   (c) ASCII coding on one keyboard would possibly not be sufficient for the necessary IPA symbols and diacritics, and there would be little or no mnemonic value in the choice of many symbols.

2. The SAMPA phonemic values are retained for each language, and the phonemic segment is subdivided into acoustically quasi-homogeneous elements. E.g. /k/ may contain a partially voiced closure, a clear burst, and a period of aspiration prior to the vowel onset. Note that this approach is an acoustic-event labelling and is used in a similar way at CERFIA, IES and UCL. The following characterisation retains the primary symbol as "pointer" to the phonemic identity of the utterance:

kv	=	Voiced portion of closure [1]
kc	=	Voiceless portion of closure
kp	=	k-burst
ka	=	k-aspiration

   Advantages: The acoustic realisation of each phonemic segment is defined in greater detail than is possible even in narrow phonetic transcription, where, for example, a partially voiced closure cannot be easily represented. [2]

   Disadvantages:

   (a) New segment markers have to be set.

   (b) The system can only apply to approaches that recognise the need to define segment boundaries (however arbitrary they may be theoretically).

   Note: It must be pointed out that the two-symbol representation given above is redundant, in that the acoustic-event categories are common to phoneme classes rather than individual phonemes; i.e. pc, tc, and kc would all be a period of voiceless closure and therefore not require the place specification. Also, if the phonemic category is specified in a different tier of annotation, it is recoverable, and may be used for a database search, e.g. with a view to developing a set of rules covering the possible "internal" structures of stretches of signal associated with a particular phoneme. At present, some partners need to retain the "phonemic pointer" in order to derive the phonemic label file from the lower level acoustic-event file.

3. A third approach, favoured by the linguistic group at ICP (Grenoble) recognises transitional phases between areas marked as optimally representative of a particular phoneme category. The finer labelling requires the delimitation of the (centre-marked) optimal area, thus also delimiting the area of co-articulation.

[2] Suggestions for the transcription of pathological speech made at the Kiel IPA Convention envisage this sort of detail, but they are still under discussion and have not yet been considered with respect to computer transcription.

Advantages: The theoretically doubtful "changeover point" from one "phoneme" segment to another is avoided, and areas of indeterminacy are identified.

Disadvantages: New markers have to be set.

Each of these approaches would provide an annotation which is closer to the (acoustic-) phonetic realisation of the utterance than the phonemic SAMPA labels. For the development of speech knowledge in general, and for the definition of rules describing the structure of continuous speech in particular, the use of a more detailed annotation is essential. It is the symbolic bridge between measurable acoustic parameters and abstract phonological categories. Which approach is selected for more detailed annotation within the SAM project depends on the use to which it will be put. Essentially, the closer a symbolic representation comes to significant acoustic events (whereby "significant" is an application-dependent term), the more useful it will be in speech-knowledge acquisition and rule development. Both synthesis and recognition assessment can only gain.

References

SAM 1988. ESPRIT Project 1541: Definition Phase Final Report "Multilingual Speech Input/Output Assessment Methodology and Standardisation". London: University College London.

SAM 1989. ESPRIT Project 1541: Extension Phase Final Report. London: University College London. VI.2: First appraisal of SAMPA.

Wells, J.C., 1987. "Computer-coded phonetic transcription", Journal of the International Phonetic Association 17:2, pp. 94–114.

Appendix C

SAM label file formats

C.1 SAM label file formats

C.1.1 Introduction

To facilitate multi-laboratory exploitation of speech database material, SAM has defined a standard, transparent specification of all files used in the acquisition of post-processing of speech recordings for assessment and research purposes. In addition, a standard SAM recording software package, EUROPEC, developed at ICP – Grenoble, uses a set of SAM standard speaker and prompt-file structures and produces speech and orthographic label files of standard format. This section gives details of formats and functions of files, and defines their function in the different stages of acquisition and processing.

C.1.2 Speech file and associated description file formats

It is now agreed as a standard for SAM speech databases, that a *speech file* contains only speech waveforms, and that an *associated description file* is generated at the recording session. Thus the files are matched, their names being identical, except for the last letter of the extension.

For example, if the speaker **AA** records the corpus number **BB** (list of six sentences in English), and the current available file number in the recording lab is **nnnn**, the files produced will be:

AABBnnnn.SES	sampled speech
(AABBnnnn.SEL	L for Laryngograph)
(AABBnnnn.SE2	for the second channel signal file)
AABBnnnn.SEO	associated description file generated automatically during recording.
	(O = orthographic time-aligned labelling)

The associated description file has standard *label-file format*, with a header and a body. (see below C.1.3.1 Header format for label files; C.1.3.2 for body of label file). It contains all the information usually needed by people working on the files without a database management system.

C.1.3 Label file format

A label file consists of a *header* and one or more *label bodies*. The header consists of the header keyword “LDH:” and a number of lines providing information about the labelling and the speech-file to which it applies. The number of lines in the header is not fixed. The header continues until the label body “LBD:” keyword appears.

The label body specifies the type and location within the speech-signal file of each segment that has been labelled. It continues until either another label body keyword “LBD:” or an end-of-label-file “ELF:”

Basic Label File Structure:

The label header	from	LHD: ...
		....
	to	LBD:
the label body	from	LBD: ...
		....
	to	ELF:
or a new label body	from	LBD: ...
		....
	to	ELF:

C.1.3.1 Label-file header format

```
LHD: header keyword + version (version V4.0, March 1991)
FIL: file type
TYP: specific file type
     (eg prompt orthographic, spoken orthographic, phonemic, prosodic etc)
DBN: database name
VOL: database volume ID
DIR: directory (for the source file)
SRC: source file name
CMT: comment
TXF: name of the textfile
     (comment on what this is supposed to contain - namely what was intended
     to be said, ie prompt file.txt or nothing (if it is a prompt or free
     speech label file then could be left blank) but could be a file containing
     instructions)
CMT: comment
SAM: sampling rate
BEG: labelled sequence start position
END: labelled sequence end position
RED: recording date
RET: recording time
REP: recording place
SNB: number of (8-bit) bytes per sample
SBF: sample byte order
SSB: number of significant bits per sample
RCC: recording conditions code (define a set of values, micro type, position..)
NCH: number of channels
SPI: speaker information: sex, age, native language
PCF: protocol file name (recording protocol used)
PCN: protocol number
CMT: comment
EXP: labelling expert
SYS: labelling system
DAT: date of completion of labelling
SPA: SAMPA version - note 2...and 3.... refer to different levels
CMT: comment
```

Note that the SAM labelling tool PTS requires all fields to have an entry otherwise it will not open the label file; "-" should be entered when the information is not available.

C.1.3.2 Label-file body format

Between the label-body keyword (LBD:) and the end-of-label-file keyword (ELF:) four categories of mnemonics can occur:

Label type:	LBR, LBO, LBB, LBP etc. (see below C.1.3.3)
EXT:	Line extension (In practice this is only used in orthographic recording
CMT:	followed by any comment needed
DSC:	Indication of discontinuous recording mode, i.e. that pauses between utterances have been automatically excised during recording (see C.1.3.4 a) below)

C.1.3.3 Label types

LBR: This label type occurs a special kind of label file which is created automatically during recording with EUROPEC (see 3 below). It contains sequence beginning (in samples), sequence end, input gain or recording, minimum sample value, maximum sample value, orthographic text. The text represents the prompt, not necessarily what the speaker actually uttered. In the case of two-channel recordings, each LBR label can be followed by an LB2: sequence (second channel speech file) or LBL: (laryngograph on 2nd channel) that contains the same information: beginning value (in samples), end value, input gain or recording, minimum sample value, maximum sample value. A mnemonic indicating other sensors (e.g. LBN (nasal), LBF (airflow), LBT (tongue contact) etc. can be defined if required. The purpose of this label file is to call up individual sections of larger files. The LBR labels represent "items" which can be specified when a signal file is opened.

LBO: Orthographic labels produced manually. These labels attempt to represent what the speaker actually produced, with indications of pauses, hesitations, repetitions etc. Often, an EXT: line and a CMT: line are needed.

LBB: Broad Phonetic labels, produced manually or by means of (semi-) automatic label alignment.

LBA: Acoustic-phonetic labels (sub-divisions of broad phonetic segments) produced manually or by means of semi-automatic labelling.

LBP: Prosodic labels, produced manually or by means of (semi-) automatic labelling.

C.1.3.4 Label File example

a) This is file DFS20014.SEO, created automatically during a single-channel recording of file DFS20014.SES.

```
LHD: V4.0
FIL: label
TYP: prompt orthographic (comment needed that mixtures are possible)
DBN: -
VOL: EUROM.1
DIR: ENGLISH
SRC: DFS20014.SES
TXF: S2.TXT
CMT: Information about the recording session
SAM: 16000
```

```
BEG: 0
END: 431872
RED: 07/11/89
RET: 15:10:33
REP: I.C.P. Grenoble (FR)
SNB: 2
SBF: 01
SSB: 16
RCC: 1
NCH: 1
SPI: M, 39, French
PCF: SENTEN.DES
PCN: 1
CMT: Information about the labelling session
EXP: -
SYS: -
DAT: -
SPA: -
CMT: Item: label start, end, input gain, min level, max level, string
LBD: -
LBR: 0, 55551, 0, -5128, 4775, Decimal numbers are an aid in
EXT: adding up.
DSC: -
LBR: 55552, 158975, 0, -7680,  8878, Monetary systems have
EXT: evolved to make use of this base ten notation.
DSC: -
LBR: 158976, 223743, 0, -7123,  7562, France became the first
EXT: decimal country in Europe.
DSC: -
LBR: 223744, 275199, 6, -12487, 13262, Germany's decision
EXT: followed eight years later.
DSC: -
LBR: 275200, 361983, 6, -11965, 12451, Scandinavian States and
EXT: Russia changed in eighteen seventy-five.
DSC: -
LBR: 361984, 431872, 6, -12902, 14320, Britain chose to have
EXT: decimal money only in nineteen seventy-one !
ELF: -
```

b) This is file ADS10010.SFO, the orthographic label file created automatically during the two channel recording of ADS10010.SFS

```
LHD: V4.0
FIL: label
TYP: orthographic
DBN: EUROM\_1
VOL: -
DIR: -
SRC: ADS10010.SFS
TXF: S1.TXT
CMT: Information about the recording session
SAM: 20000
BEG: 0
```

```
END: 193023
RED: 11/Apr/90
RET: 16:47:46
REP: ICP
SNB: 2
SBF: 01
SSB: 16
RCC: 2
NCH: 2
SPI: M, 49, French
PCF: PEQPHRAS.DES
CMT: Information about the labelling session
EXP: -
SYS: -
DAT: -
SPA: -
CMT: Item: label start, end, input gain, min level, max level, string
LBD: -
LBR: 0, 17663, 0, -12382, 14455, Maman a prpar une galette
EXT: pour jeudi ?
LB2: 0, 17663, 0, -12384, 14439
DSC:-
LBR: 17664, 41215, 0, -6954, 9023, Ceslves prendront
EXT: l'autocar tout  l'heure !
LB2: 17664, 41215, 0, -6958, 9024
DSC:-
LBR: 41216, 86527, 0, -11750, 15336, Parfois, monpicire
EXT: vend  crdit.
LB2: 41216, 86527, 0, -11756, 15348
DSC: -
LBR: 86528, 137983, 0, -11809, 11572, Personne n'a applaudi ce
EXT: beau discours ?
LB2: 86528, 137983, 0, -11816, 11560
DSC: -
LBR: 137984, 193023, 0, -15097, 18881, Je me demande pourquoi
EXT: on court sans cesse.
LB2: 137984, 193023, 0, -15102, 18880
ELF: -
```

C.1.4 Files used in a recording session using EUROPEC

The standard SAM recording software EUROPEC, is geared to producing signal and label files of standard format. In turn the software requires input data which is formatted according to the following specifications.

C.1.4.1 Corpus File: CORPUS.DBF

The Corpus File lists all the prompt files used in a particular set of recordings. This file is scanned whenever a file is specified for a recording "take". The prompt files, which have a 2- character name and extension .txt, are arranged in an alphabetical catalogue A–Z, each letter containing up to 36 prompt files (e.g. A0–A9 plus AA–AZ). Data entries for the prompt files are of two types:

1. those required for the prompting process, and

2. those required for the file header.

Order of data entries:

```
CMT:Optional entry providing information about the prompt file, typically the prompt file
name and a short description.
CCD:Corpus code  ( 2 letters (ex: P2) )
DBN:Database name (20 characters max)
CNM:Corpus name (40 characters max)
CTY:Item type (one letter)
NBI:Number of items in the file (int)
LAN:Language (char)
PCF:Protocol Description File  (XXXXXXXX.DES)
--------------------------------------------------------------
A
-
--------------------------------------------------
.........
--------------------------------------------------
P
-
CMT:       P1.TXTPrompt text for passage 1
CCD: P1
DBN: EUROM\_1
CNM: A reading passage (Esprit CD)
CTY: P
NBI: 1
LAN: E
PCF: PASSAGE.DES
--------------------------------------------------
Q
-
........
--------------------------------------------------
S
-
CMT:       S2.TXTPrompt text for Block 2 of 5 sentences
CCD: S2
DBN: EUROM\_1
CNM: Sentences (UCL)
CTY: S
NBI: 5
LAN: E
PCF: SENTEN.DES
--------------------------------------------------------
T
-
CMT:        T3.TXTPrompt text for Block 3 of 30 digit triples
CCD: T3
DBN: EUROM\_1
CNM: list 3D Digits Triples
CTY: N
NBI: 30
```

```
LAN: E
PCF: DIGIT.DES
---------------------------------------------------
```

C.1.4.2 Speaker File: SPEAKERS.DBF

This file contains details about the speaker which can be accessed via the speaker codes. To maximise the speaker number, arbitrary code allocation AA-ZZ (=680) should be used rather than the speaker initials used in the examples. In many countries, the SNM and SBN lines will be left blank for data-protection purposes.

```
        Description of speaker specificities
--------------------------------------------------------------
order :
SCD:Speaker code  (2 char)
SNM:Speaker name (75 char max)
SBN:Speaker birthname (75 char max)
SEX:Sex (one letter)
DOBDate of birth (year)
HET:Height (metres)
WET:Weight (kg)
NLN:Native language (75 char max)
ACC:Accent                "
ETH:Ethnic group          "
EDL:Education level       "
SMK:Smoking habit         "
PTH:Pathology             "
--------------------------------------------------------------
A
-
--------------------------------------------------
D
-
SCD: DJ
SNM: DURAND
SBN: Joseph
SEX: M
DOB: 1946
HET: 1,73
WET: 70
NLN: French
ACC: Ardche
ETH: white
EDL: -
SMK: heavy smoker
PTH: -
-----------------
SCD: DS
SNM: DUPUY
SBN: Simone
SEX: F
DOB: 1952
```

```
HET: 1,60
WET: 58
NLN: French
ACC: -
ETH: white
EDL: -
SMK: -
PTH: -
......
--------------------------------------------------
E
-
......
-
--------------------------------------------------
Z
-
--------------------------------------------------
```

C.1.4.3 Prompt file

The prompt file is given the name of the Corpus Code (CCD:) specified in the Corpus File. The name corresponds to its position in the Corpus File (X2 signifies that it is specified in 2nd position under letter x).

- Each prompt text file begins and ends with a specific mnemonic "SPR: - EPR:".
- Each item is preceded by the mnemonic "TXT:".
- Each line that is a continuation of the preceding one, begins with the mnemonic EXT.
- If the text file is to be used in continuous recording mode, it must specify the display time in seconds for each item (following the mnemonic "DLA:").

Examples of Prompt Text Files

S1.TXT (5 sentences) S1 is the corpus code

```
SPR: (start prompt file)
TXT: I have a problem with my water softener.
TXT: The water-level is too high and the overflow keep dripping.
TXT: Could you arrange to send an engineer on tuesday morning ?
TXT: It's the only day I can manage this week.
TXT: I'd be grateful if you could confirm the arrangement now.
EPR: (end prompt file)
```

S2.TXT (5 sentences) S2 is the corpus code

```
SPR:
TXT: Please put me through to the complaints department.
TXT: The repair to the water main outside my house was
EXT: unsuccessful, and my cellar's flooded.
TXT: Your Water Services Department was singularly unsympathetic.
```

```
TXT: All their repair teams are apparently booked out for the
EXT: next two weeks.
TXT: Am I supposed to use the cellar as a swimming pool till
EXT: then ?
EPR:
```

Note that if a sentence is more than one line long, there will be one extension line (field EXT).

P1.TXT (passage) P1 is the corpus code

```
SPR:
TXT: Please put me through to the complaints department. The
EXT: repair to the water main outside my house was unsuccessful,
EXT: and my cellar's flooded. Your Water Services Department was
EXT: singularly unsympathetic.  All their repair teams are
EXT: apparently booked out for the next two weeks.  Am I supposed
EXT: to use the cellar as a swimming pool till then ?
EPR: -
```

Note that all this passage will be displayed as one item (only one TXT field and several EXT fields).

ST.TXT (5 sentences) ST is the corpus code
To be recorded in "continuous mode" (= all pauses between sentences are retained).

```
SPR: -
TXT: Please put me through to the complaints department.
DLA: 3.0
TXT: The repair to the water main outside my house was
EXT: unsuccessful, and my cellar's flooded.
DLA: 5.0
TXT: Your Water Services Department was singularly unsympathetic.
DLA: 3.5
TXT: All their repair teams are apparently booked out for the
EXT: next two weeks.
DLA: 4.0
TXT: Am I supposed to use the cellar as a swimming pool till
EXT: then ?
DLA: 3.8
EPR: -
```

C.1.5 Protocol description file

In a recording session, the prompt presentation is driven by a specific file called: Protocol Description File (specified after the PCF: Corpus File entry). It specifies the various stages and messages of the prompt presentation .....

C.1.5.1 Codes used for presentation protocol (Protocol Description File)

Syntax is	S1:	<Field_ID>: <value>
or	S2:	<Field_ID>: <value>, <value>, <value>, ...

CLS:	ClearScreen and put cursor to 0,0
VON:	Video attribute ON: (S2) the following attributes are activated: center : text will be centered blink : blinking mode inverse : video inverse mode underline : text will be underlined bright : text will be bright
VOF:	Video attribute OFF (S2)
BEL:	ring bell
MSG:	message to be displayed according to valid video attributes
NWL:	newline on screen
PAU:	n (integer) = duration of the pause in seconds
TRN:	set training mode, no recording
CMT:	comment
RWD:	rewind external file (corpus text file)
GET:	n (integer) n: number of items to be read in the external file (corpus text file) then displayed and recorded (if recording mode is activated) or not (if training mode is activated)
DTA:	(string) Data. The item to be displayed (and recorded or not) is given here.
EXT:	extension to previous line.
SKP:	n (integer) n: number of lines to be skipped in external file (corpus text file)
LOP:	-n, m loop in the present file n: number of one-line-back in this file m: number of iterations
REC:	recording mode activation
BCK:	n (integer) n: number of one-line-back in the external file (not yet implemented)
RET:	waiting for a CR (carriage return)

NOTE: Using the "GET: 3" command, will produce the following effect:

Using Manual or Automatic mode:	getting the three next items in the text file, displaying them and recording them as a single item.
Using Continuous mode:	getting the three next items in the text file, displaying them one after the other, and recording them without any discontinuity.

C.1.5.2 Example of Protocol Description File

This is SENTEN.DES (Protocol Description File for a corpus like S1 or S2 and given after PCF: in the Corpus File entry)

```
SPF: start of protocol file
CLS: -
VON: inverse, center
BEL: -
NWL: 2
MSG: ENGLISH SENTENCES
NWL: 2
```

```
MSG: I.C.P. GRENOBLE
VOF: inverse
PAU: 2
CLS: -
NWL: 2
MSG: Let's practise a little :
NWL: 2
MSG: please read the following
NWL: 2
MSG: two sentences
PAU: 3
CLS: -
TRN: -
CMT: DISPLAY BEGINNING - TRAINING PHASE
RWD: -
GET:* 1
SKP:* 2
LOP: -2, 1
CMT: DISPLAY END - TRAINING PHASE
RWD: -
CLS: -
NWL: 2
MSG: Now you're going to be recorded
NWL: 2
MSG: this session consists of five sentences
NWL: 2
MSG: read each of them
NWL: 2
MSG: while they are being displayed
PAU: 4
CLS: -
NWL: 8
MSG: CAUTION ! SESSION BEGINNING
VON: blink
NWL: 1
MSG: Hit Space or Click Left Button
RET: -
VOF: blink
CMT: DISPLAY BEGINNING
REC: -
CLS: -
GET:* 1
LOP:* -1, 4
CLS: -
NWL: 2
CMT: END DISPLAY
CLS: -
NWL: 8
VOF: center
VON: blink
MSG:   STOP
VOF: blink
MSG:    Recording
```

```
PAU: 1
VOF: inverse
VON: center
CLS: -
BEL: -
NWL: 4
MSG: It's all over for this file
NWL: 1
MSG: THANKS
NWL: 8
EPF: end of protocol
```

NOTE: For the ST corpus, using continuous mode, the protocol file is nearly the same, but:

- No LOP instruction is allowed.
- There is one GET instruction for all the items to be recorded in sequence together.

C.1.6 Recording conditions file

A Recording Conditions File is selected (and can be modified) prior to each recording session. Some of the entries are accessed for completion of the label-file header, others are for general reference. A number of files specifying different recording conditions can be stored and selected when requested. It describes a referenced set of recording conditions.

Example: 1.RCD
(1 is recording conditions code 1, implying that this file is the first of a number of defined conditions.)

```
SCD:                   start of conditions
RCC:    1              recording condition code
VER:    V3.0           label version
VOL:                   needed for label file
DIR:                   needed for label file
SNB:    2              sample byte number
SBF:    01             sample byte order
SSB:    16             sample significant  bits
NCH:    1              number of channels
LGG:    0              laryngograph used or not
PCN:    1              protocol code number
SAM:    20000          sampling frequency (20kHz is SAM standard for EUROM.1ff)
MIN:                   micro name
MIT:                   micro type
MIP:                   micro position
MID:                   micro distance
NOB:                   noise bandwidth
NOL:                   noise level
CHB:                   channel bandwidth
CHN:                   channel noise
FLT:                   additional filter
PRO:                   external processing
ENV:                   environment
ECD:                   end of condition
```

C.1.7 PRIVATE.PEC file

This file is local to the SAM workstation. It provides information used for the recording process (OROS board type and address) and for completion of the label-file header, as well as registering the number of files recorded on the workstation.

Example:

```
BTY: 2                              Oros board type (1: AU21  2: AU22)
ADR: 784                            Oros board address
LOC: I.C.P. Grenoble (FR)           Recording place
NUM: 0000                           First file number
NSC: 1                              Number of monitors
```

Appendix D

SAM recording protocols

D.1 SAM recording protocols

D.1.1 Definition of terms

SAMPLE A digital representation of the analogue waveform at any one instant in time

SAMPLING RATE The (normally constant) rate at which SAMPLEs are collected and transferred to a storage medium.

SAMPLING INTERVAL The time between successive SAMPLES - fixed for any one SAMPLING RATE while the sampling process is not stopped.

DISCONTINUITY The point in a stream or stored set of SAMPLES where the sampling process has been stopped for longer than one SAMPLING INTERVAL.
(Explanation - it is equivalent to the splice edit point on the tape of an analogue tape recorder.)

SEGMENT A set of sequential SAMPLES held on the storage medium which does not contain a DISCONTINUITY. A SEGMENT will be bounded by two DISCONTINUITIES
(Explanation - this means that during the period over which a SEGMENT has been collected the process of sampling and storing has not been stopped.)

SPEECH FILE A set of sequential SAMPLEs held on the storage medium. The SPEECH FILE will contain one or more SEGMENTS.
(Explanation - the sampling process may have been stopped, then started again during the period over which a SPEECH FILE has been collected.)

TAKE The time over which a subject is invited to complete the production of a specific set of utterances. The TAKE will be recorded in one or more SEGMENTS. A SPEECH FILE will often (but not always) contain one TAKE.
(Explanation - the start of a single TAKE will often be identified by the issuing of a set of instructions to the subject and/or the inputting into a prompting system of the utterance set ID. It will normally be terminated by the prompting system finishing an utterance list and further operator-subject dialogue. There will normally be no operator-subject dialogue within a TAKE.

SESSION The period of time from when a specific subject arrives at the recording location to when he leaves. During this time one or more TAKEs will be recorded.

SPEECH CORPUS A set of SPEECH FILES.

SPEECH DATABASE One or more SPEECH CORPORA complete with the information to allow the CORPORA to be usefully accessed. This information will usually be in the form of time-aligned annotation, speaker descriptions, recording condition files, etc.

D.1.2 Classification of general strategies for recording and prompting

When designing a SPEECH DATABASE a number of decisions have to be made concerning the size and content of utterance lists, size of TAKES, length of SESSIONS, methods of prompting and recording, and dealing with errors. It is assumed in the following that a digital storage method is to be used, although many of the points in the following would apply in a slightly modified form for an analogue recording system.

There follows a classification and discussion of some of the design areas.

D.1.2.1 Recording mode

The RECORDING MODE defines the exact method of controlling transfer of speech samples to the storage medium. It is defined in terms of the TAKE, SEGMENTS and SPEECH FILES. It is concerned with the stopping and starting of the sampling process and the way the recording system treats any possible speaking errors.

MODE 1 A TAKE is recorded in one complete SEGMENT, errors are not excised.
(Explanation - this means that the sampling and transfer process is started at the beginning of the TAKE, all the acoustic signal, including breath noise, speaking errors are recorded and the sampling and transfer process is only stopped at the end of the TAKE.)

MODE 2 A TAKE is recorded in one complete SEGMENT, except when a speaking error is detected during the TAKE, and then that error is excised from the storage medium. If there have been no speaking errors there will be no DISCONTINUITIES in the SPEECH FILE. For each section of speech removed due to a speaking error there will be one DISCONTINUITY.

MODE 3 A TAKE is recorded in a number of SEGMENTS. There will be one SEGMENT per utterance, and speaking errors will be included in the SPEECH FILE.
(Explanation - in general this means that an endpoint detector is used to start and stop the sampling and transfer process. Apart from the 'back-off' time no inter-utterance acoustic signal will be stored in the SPEECH FILE.)

MODE 4 A TAKE is recorded in a number of SEGMENTS. Speaking errors will be excised from the storage medium, and so there will be one SEGMENT per required utterance. This MODE is the same as MODE 3 except that the resulting file will contain no errors.

D.1.3 Prompting style

The PROMPTING STYLE primarily defines how a prompting system reacts when speaking errors have been detected during a TAKE and how the subject is instructed to behave after the production of speaking errors. The following classifications assume an automatic prompting system, but non-automatic systems can be classified in the same way. For each style there are three different timing strategies.

STYLE 1 ABORT TAKE and re-record. The subject is instructed that if he realises that they have made a speaking error then they should indicate the fact to the operator. When the operator is aware of a speaking error the prompting and recording system are stopped by a suitable 'escape' mechanism. This situation is indicated to the subject and the prompting system is started to re-record that TAKE.

STYLE 2 BACKUP-ON-FLY n utterances. If and when a speaking error is detected during a TAKE, the prompting system is backed up to a point before the error and continues from there. What happens to the recording will depend on the MODE being employed.

STYLE 3 TACK-ON-END correction. If and when a speaking error is detected during a TAKE the prompting system carries on without stopping until the last required utterance has been produced and then the subject is prompted for one or more utterances to correct for the speaking errors. The recording action depends on the MODE employed.

STYLE 4 NO PROMPTING REACTION. The prompting system carries on producing prompts in the pre-determined order. The subject may or may not be asked to react by correcting the error in some way.

D.1.3.1 Timing strategy

REGULAR The timing of the prompt is independent of the rate at which the subject is actually speaking. In its simplest form each prompt is displayed at regular predefined intervals which will depend on the type of utterance that is to be produced. Alternatively the interval may vary from prompt to prompt depending on the expected (not the actual) time to speak the prompted text, i.e. dependent on the number of words/phonemes in a sentence.

ENDPOINT The timing of the prompt is totally controlled by the production of utterances. The display of each new prompt is controlled by the detection of the endpoint of the last utterance.

MIXED The timing of the prompt is controlled by a logical combination of a predetermined interval and the endpoint of an utterance. The display of each new prompt is triggered by whichever is later of the predetermined interval or the endpoint. This means that if a subject is slow the prompting system will slow down, but there is a maximum rate of prompting even for the fastest speaker.

D.1.4 Recording protocol

D.1.4.1 Microphone

A single B&K half inch pressure microphone number 4155 will be used connected to a 2230 level meter, used as a microphone amplifier. If a recording site proposes to use another microphone or preamplifier the alternative choice of equipment should be ratified by other members of the project. IES has suggested a 4165 microphone, 2660 pre-amp and 2636 amplifier. Bochum University recommend the ADG C-414 microphone in omnidirectional mode connected to the John Hardy M-1 mic preamp.

The audio output of the level meter will be connected to both the line input of the OROS AU21 board and the left channel of the digital audio safety backup recorder. The method of splitting the signal will be made in such a way to ensure no degradation of the audio signal particularly in terms of loading, balanced/unbalanced connections and hum loops. Bochum University recommend the Brooke Siren Sytems microphone splitter.

The microphone will be positioned for 90 degree incidence, 50cm from lips 15 degrees off axis. (TNO configuration)

Tests by NPL have shown that this distance is acceptable in terms of speech signal to ambient/monitor noise.

The microphone position relative to the head/lips will be set and maintained by a headrest (or the not so consistent method of requesting the subject to maintain head position).

D.1.4.2 Other sensors

A Laryngograph will be used for some speaker/material combinations as specified in material choice section. The neck sensors will be positioned in accordance with UCL instructions, namely on either side of the point of the thyroid cartilage. The Lx output of the Laryngograph will be connected to the right-hand channel of the digital audio (safety backup) recorder.

D.1.4.3 Speech data capture

Single channel, digital, direct recording to disc will be used for the microphone signal. Digital audio tape recorder safety backup will be made of microphone channel at all times. Second channel recording direct to disc and/or digital audio tape recording of the Laryngograph signal will be made as required.

The mechanism for direct digital recording to disc will be by use of the OROS AU21 board for single channel or an OROS AU22 for two-channel recordings on the SESAM workstation with EUROPEC software. The sampling rate will be 20KHz. and the standard OROS digital filtering used with an oversampling factor of 4. The nominal resolution of this system is 16 bits. Care should be taken to ensure that OROS board is placed in a slot in SESAM which minimises PC electrical interference.

The 'standard' analogue filter on the OROS board is a 20 KHz low-pass filter with 0.3 dB ripple in the passband, 160 dB/octave slope and 80dB rejection. This is followed by a digital filter implemented on the TMS320C25 having the -3dB point at 8KHz. Any DC offset is automatically removed by the OROS/EUROPEC software and the gain of the line input will be set within EUROPEC on a 'per-speaker' basis.

The safety back-up (and second sensor recordings if an OROS AU22 is not used) will be made on a digital audio tape system such as the SONY Video PCM or DAT systems, having a specification with regard to bandwidth, signal-to-noise ratio and wow and flutter in excess of that of the OROS AU21 board. The recorder shall be left recording for the WHOLE of a session and not be stopped or rewound at all during the session. This removes the chance of over-recording wanted portions, reduces the operator effort will allow later analysis of a number of features not captured on the direct to disc recordings, such as number of retakes, motivation level of subject, coughs etc.

D.1.4.4 Recording environment

The recordings will take place in an anechoic room.

The subject will not be provided with an electrical sidetone path for their own voice.

The lighting and temperature levels will be such as to cause no stress to the subject and care must be taken to ensure that there is no distracting light reflection off the prompting screen. Care must also be taken to ensure that the anechoic nature of the recording is not compromised by allowing significant acoustic reflections to be picked up on the microphone. The prompting monitor should be slightly angled and there should be no tables or other sound reflecting surfaces between the subject and the microphone (or tables to be covered with sound absorbing material).

D.1.4.5 Recording mode and prompting style

It has been agreed that the IASM-A CORPUS should contain no speaking errors. Further the inter-utterance pauses should be captured in their entirety and there should be no DISCONTINUITIES within a TAKE.

> PROMPTING STYLE 1 WILL BE USED TOGETHER WITH RECORDING MODE 1. THERE SHALL BE ONLY ONE TAKE PER FILE.

The recording system will be integrated with the prompting system and this is to be as automatic as possible (see EUROPEC Section III below). It is clearly possible for the prompting system to start the recording system before the first prompt. In the general case of RECORDING MODE 1 as discussed in the earlier part of this document, the end of the final utterance is not defined. However, it has been agreed that a base level annotation will be automatically produced during the TAKE. This will consist of the prompted orthographic text along with the endpoints of each utterance. As each endpoint is known to the prompting/recording system, then this information can be used to stop the recording process at the end of the final utterance.

> A TIME ALIGNED ORTHOGRAPHIC ANNOTATION WILL BE PRODUCED AUTOMATICALLY DURING THE TAKE — THE RECORDING PROCESS WILL BE STOPPED AUTOMATICALLY AFTER THE FINAL UTTERANCE.

However, if in the light of experience, it is found that significant SESSION time is wasted in re-recording TAKES or subjects are being excessively stressed by the 'ABORT TAKE on error' style then the next most appropriate technique would be to 'BACKUP-ON-FLY' and excise any errors from the TAKE. This technique will require an addition to the previously agreed standard of annotation file, namely a label to indicate the position of the DISCONTINUITY so caused.

> PROMPTING STYLE 2 AND RECORDING MODE 2 WILL BE USED ONLY EXCEPTIONALLY AND WITH THE AGREEMENT OF OTHER MEMBERS OF THE CONSORTIUM — A DISCONTINUITY MARKER TO BE ADDED TO THE ANNOTATION FILE. ONE TAKE PER FILE.

It has been suggested that a regular timing strategy puts too much stress on some speakers and that there is evidence of excessive speaking rate when endpoint timing (i.e. speaker driven prompting rate) is used.

> MIXED TIMING STRATEGY WILL BE USED

The exact value for tP (the minimum prompting interval in this strategy) depends on the display time (DLA: [time in seconds] in the prompt file) plus the stop back-off time.

D.1.4.6 Recording control

The subject will be prompted by an automatic system — EUROPEC running on the SESAM workstation positioned outside the anechoic room. The EUROPEC system will be controlled by a recording manager (operator), viewing the standard SESAM monitor. The subject will use a second monitor placed in the anechoic chamber, taking its video input from a 'T' connection from the SESAM video output. The size and positioning of the monitor will allow the subject to see the prompted text and level meter without stress. The positioning of the monitor with respect to the microphone will ensure that there is no detectable noise pickup, either acoustically or electrically (particularly at line and frame frequencies.) Bochum University suggest that by using EGA or VGA graphics mode the line frequency is above 20 KHz and therefore unlikely to be a source of interference.

The speaker to microphone distance, azimuth and elevation will be controlled by ensuring minimal head movement commensurate with minimum subject stress. A headrest is highly recommended.

The subject will have, for the material/country combinations specified in the recording material section, an acoustical prompt generated by a second SESAM fed by the RS232 output of EUROPEC.

The operator will continuously listen to the subject at all times via the recording chain. This will allow subject-operator communication and enable the operator to do a 100% check on the speech material content. Bochum University suggest that 2 operators should be employed simultaneously to reduce fatigue. The headphone output of the PCM/DAT digital audio safety backup machine would be a suitable source for audio monitoring.

The operator will speak to the subject in the anechoic room via a one way intercom/loudspeaker system, which will be switched off during takes. The gain of the intercom system shall be set to a level at the ear representing speech 1 metre from the lips of an average talker. The setting of this level is important as it will tend to effect the speaking level of the subject — and should be consistent across recording sites.

Speaking effort will be controlled for some speaker/material combinations as specified in the reference material section. This will be done by the use of a speech level meter displayed on prompting monitor by EUROPEC. There will be three modes of operation:

1. The level meter gain set to nominal, after the speaker has stabilised in level the recording gain will be adjusted so that the normal peak level of the speech reaches a reference point, 12 dB below peak. The speaker is controlled for consistent speaking level during the subsequent take by being asked to keep to reference point as much as possible.
2. The level meter gain is set 6dB lower, the speaker keeps to reference point and so will speak 6dB louder.
3. The level meter gain is set 6dB higher than nominal, the speaker keeps to the reference point and so will speak 6dB more quietly.

Speaking style will be controlled by:

1. For some specified material the prompting system will show sets of sentences as a single item, and for other material they will be shown as separate items.
2. For some country/material combinations the speaker will be additionally prompted by an audio stimulus. The audio will be generated as a result of an RS232 serial output from EUROPEC/SESAM.

Speaking rate will not be controlled except that the maximum rate of presentation of prompted items will be at a rate set within EUROPEC dependent on the material. This rate will be defined within the choice of material section.

There will be no attempt to produce stress in the subject. The total amount of material recorded in each session will be restricted to 8 minutes representing approximately a 30 minute elapsed time session. This limit will minimise stress on the subject and represents the maximum time that can be recorded on a 20 Megabyte disc.

The recording manager will ensure that the subject is in a relaxed condition.

There will be no attempt to control any speaking effect due to the time of day of the session.

D.1.4.7 Recording procedure

At each session the speaker code of the subject will be confirmed. (Initially this will be done by checking a printed list of speaker details — an automatic system will subsequently be used.)

Session specific information will be entered via the EUROPEC system — some manually and some from the conditions file.

At each session the recording manager will provide an outline of the procedure for that session. This briefing should be in written or recorded form so as to achieve consistency between subjects.

At each session in the anechoic room the subject will be asked to practise speaking typical material for that session, with the level meter switched off. A note will be made of the peak level during the practice and the gain of the OROS AU21 will be set such that this peak level is 12–15 dB below maximum possible recording level. A large number of experienced recording sites have rejected the original suggestion of 6 dB head room. They feel that 12 to 15 dBs would be safer and reduce the possibility of takes having to be re-recorded due to overload, this in turn will reduce the stress on inexperienced speakers. The gain setting will be automatically recorded for inclusion in the session dependent recording conditions file.

If the session is to involve the use of control by level meter the subject should practise controlling their level with respect to the meter, for all (3) level meter gain settings.

The recording conditions/material etc. for each take will be entered into EUROPEC — this will cause the correct file names to be generated.

At each take the subject will, where appropriate, be given the opportunity to preview the prompt material. This is particularly needed for continuous speech material. In some cases, e.g. for particular forms of numbers, orthographically ambiguous nonsense material etc., additional instructions need to be given, and the subject given time for familiarisation with the task before the session.

The recording manager will monitor the speech production with reference to the prompted text. If there are any errors the take will be stopped and the subject asked to start the relevant take again. The disc file will be discarded and the same serial number reused for the re-recording. The digital audio tape recorder safety backup will not be stopped due to re-recording of errors — only at the end of a session. If after the integrity check a take has to be re-recorded then the digital audio safety backup recorder should be restarted.

D.1.4.8 Integrity checks

At the end of each session the quality of the recordings should be assessed by a 10% review, and the integrity of the item end-point labelling should be 100% checked by using the VERIPEC module within the EUROPEC software.

D.1.4.9 Backup procedures

At the end of each session all files produced should be transferred to Exabyte 8mm tape system or via Ethernet to local mainframe for backup. The available disc space on the SESAM workstation is approximately 20Mbytes, representing 8 minutes of speech — this would therefore set a convenient and reasonable maximum to the recording session and represents approximately a 30–40 minutes elapsed time per session. All files produced during the session should also be backed up at the same time — the data is a valuable resource and files are easily lost. The digital audio safety backup tapes produced for each session shall be kept indefinitely as an archive of the session. It is expected that 4 to 6 sessions can be accommodated per tape, and care must be exercised not to over-record a previous session.

D.1.4.10 Retrieval procedures

If a speech file is lost AND its computer backup is also lost (OR the backup was never made!!!) then the digital tape recording safety backup will have to be used. This raises a number of questions. If the recordings are transferred to computer form in the digital domain, then an OROS AI-PCM type interface will be required — but this does not allow a resampling rate of 20KHz–19.8KHz is closest possible. This retrieval method is only possible if the recording was made on a PCM or DAT machine where the pre-emphasis was switched off. The alternative is to copy in the analogue domain, the problems here are:

1. There is normally a degradation of frequency and phase responses and signal-to-noise ratio is reduced (the latter effect is possibly masked by the reported poor signal to noise of OROS A/D board).
2. The gain of the original signal is relatively poorly defined because the gain on the PCM unit is infinitely variable and there is no automatic method of noting the gain set for the microphone channel for each speaker. The calibration signal should be used for this purpose.

D.1.4.11 Calibration

The recording chain will be calibrated in two ways:

1. A B&K 1KHz 4230 calibrator is placed over the microphone.
2. The injection of a 4:7 Mark-Space ratio square wave of 20Hz and 100 Millivolts pk-pk. electrical signal into the microphone body. This is achieved by the use of:
   (a) a signal generator (the circuit of which can be obtained from UCL);
   (b) a B&K Input Adaptor, JJ 2614, which replaces the microphone capsule;
   (c) a capacitor of a value equal to that of the capsule element replacing the 1nF capacitor in the B&K adaptor.

A recording of both calibrator signals shall be made every 20 sessions, providing confirmation of gain calibration, and recording path frequency response, phase response and noise figure. The length of recording of each calibrator shall be 10 seconds plus 10 seconds of no signal condition. It would be sensible to check the signal-to-noise ratio using these files at the site to confirm the correct working of the system.

The Laryngograph channel shall also be calibrated, using the UCL 4-7 mark-to-space ratio rectangular wave calibrator.

D.1.4.12 Inter site consistency and recording procedure verification

A pilot set of recordings will be made at each site consisting of 1 speaker speaking 10% of all the sentence and number material and all CVCs. This pilot material from each site will be transferred using floppy discs, one each for calibration, sentences, CVCs and numbers. The material will consist of the speech files together with corpus description files, labelling files, speaker file, text files and conditions file.

The floppy discs will be sent to NPL where the quality and consistency will be assessed by a panel of 'experts'. Modifications and clarifications to this set of protocols may then be made.

D.1.4.13 Collation of recordings

All the recordings at one site will be collated and checked. Checks will be made for:

- Quality of speech material and labelling on a 1% basis.
- Consistency of file names.
- Number of files of each type.

The central collation point will configure the material in the correct directory structure for the production of one or more sets of CD-ROM.

Appendix E

SAM software tools

The following is a summary of the software packages which have been developed within the SAM project. With the exception of PTS (see 4) and SAMSPEX (see 10), all software is freely available to other ESPRIT projects, but a charge will be made for media, postage and packing.

E.1 DKISALA (Interactive Semi-Automatic Labelling Software)

Version: 1.1, March 1992

Hardware / OS requirements: Standard SESAM workstation

Description:

- The DKISALA software tool for semi-automatic labelling of speech corpora runs on the SESAM workstation. It is possible to listen to the speech signal during annotation if a non-SAM-standard PC speech processing DSP32C board is installed in the workstation (The AT&T development board or the Loughborough board PC system board based on the AT&T DSP32C signal processor. Details of configuration and board addresses can be given by the developing laboratory).
- The DKISALA system uses a trained and calibrated Self-Organising Neural Network (SONN) to convert speech frame cepstrum coefficients into a set of continuously valued Acoustic-Phonetic Features, which are further transformed into a smaller set of Principal Components. The Principal Components are used to model individual allophones in a multivariate Gaussian probability density function, and these are processed by a Viterbi Search and Level-Building algorithm, which is constrained by the independently given string of phoneme symbols corresponding to the speech signal being labelled.
- The SONN is trained on large speech corpora, which have been manually labelled prior to training. The presently used training data comprises EUROM.0 speech material from three speakers for the language under analysis.
- The DKISALA system is presently working as a preliminary version of an interactive system. Results have shown that certain sound-class transitions are positioned very accurately and reliably, while some classes are regularly inaccurate.
- The interactive component of the SALA system has been introduced to provide a means of preventing a cumulative error accruing as a result of a number of such sounds occurring within a short space. The automatic process stops when such a sound occurs. In such cases, the speech signal, the corresponding acoustic-phonetic features and the spectrogram are displayed on the graphic screen, and by using the mouse, the information on the screen and a listening facility, the user can propose the positioning of the

specific boundary transition to the system, which then takes control and reruns the label procedure. This procedure is repeated until the entire speech corpus is labelled.

Developing lab:

Speech Technology Centre
Institute of Electronic Systems
Aalborg University
Fredrik Bajers Vej 7
DK-9220 Aalborg, Denmark

Tel: (45) 98 42 11 x 4866

Fax: (45) 98 15 67 40

Contact: Paul Dalsgaard

E.2 ELSA (ESPRIT Labelling System Assessment software)

Version: 2.4, December 1990

Hardware / OS requirements: SESAM workstation or Unix workstation with X11 (e.g. SUN)

Description:

- The aim of this software is to provide the laboratories involved in the SAM project with a common and useful tool to assess semi-automatic labellers. It deals with the comparison between manual labelling of speech files (the reference) and semi-automatic labelling of the same files (both in the PTS format with SAMPA labels).
- Both files must contain the same phonetic strings. Therefore, the comparison can only be drawn in terms of boundaries location or centre location shifts. Included in this package are standard statistic tools, such as histograms, signed means and the best 10 and worst 10 boundary locations. It is possible to analyse either fine errors or gross errors and results can be written in a text file for analysis or printing. Confidence intervals are computed for a confidence level of 95%.

Developing lab:

CRIN-CNRS & INRIA laboratories
B.P. 239
54506 Vandoeuvre-les-nancy CEDEX
France

Tel: (+33) 83.91.20.00

Fax: (+33) 83.41.30.79

Email: {bourjot,boyer,fohr}@loria.crin.fr

Contact: Christine Bourjot, Anne Boyer, Dominique Fohr

E.3 EUROPEC (European Program d'Enregistrement de Corpus)

Version: 4.11, September l991

Hardware / OS requirements: Standard SESAM workstation, with AU21/22 Oros board

Description:

- This is a corpus recording software which facilitates the recording of speech databases. It ensures constant recording conditions for various speakers and/or corpora. The recording session is driven by a protocol file (user-made) which describes the way in which items (constituents of the corpus) will be displayed on the screen.
- Features and functionalities:
  - "Standard SAM format" file production
  - One or two-channels
  - Recording and listening facilities
  - Three recording modes (manual, automatic, continuous)
  - Parameter adjustment (sampling rate, input and output gain, etc.)
  - Sound level meter
  - Free/operator controlled session
  - Error treatment
  - Easy data verification
  - File safeguards
  - Easy data retrieval and modification
  - Simplified recording session
  - User's guide, and many example files. User customisable in any language

Developing lab:

I.C.P. (Institut de la Communication Parlee)
46 Avenue Félix Viallet
38031 GRENOBLE Cedex
France

Tel: (+33) 76 57 45 33

Fax: (+33) 76 57 47 10

Contact: Jerôme Zeiliger, Jean François SERIGNAT

E.4 PTS (Progiciel de Traitement de Signal)

Version:

- 4.40p (protected mode) March 1992
- 4.40r (real mode) March 1992

Hardware / OS requirements: standard SESAM workstation with AU21/22 Oros board

Protected mode version: at least 1 Mbyte of extended memory is required

Console: Microsoft mouse, EGA or VGA display

Description:

The tool consists of multi-window graphic software for speech:

- speech signal display; several formats including SAM standards;
- listening; total, selected records, between cursors, etc.;
- signal editing: copy, cut, move, etc.;
- speech processing: spectrogram, spectrum, etc.;

- manual labelling, semi-automatic label file correction;
- measurements;
- networking facilities (file transfers, remote login, etc.);
- user compliant by means of configuration files.

Developing lab:
I.C.P. (Institut de la Communication Parlée)

46 Avenue Félix Viallet
38031 GRENOBLE Cedex
France

Tel: (+33) 76 57 45 39/33

Fax: (+33) 76 57 47 10

Contact: Jean-Claude Caërou, Jean-Marc Dolmazon

NB: This software package has different distribution regulations. The user should first sign a confidentiality agreement due to the use of two licenses:

- Rational System's DOS/16M extending memory for protected mode operations;
- Metawindow Plus graphic library.

E.5 RESAM

Version: 1.1, November 1991

Hardware / OS requirements: Low Level SESAM Workstation

Description:

- The RESAM software package contains a program for changing sample rates of signal files, RESAMP, and a program for adding two signal files, ADD.
- The RESAMP program changes a signal's sample rate from a sample frequency Fa to a sample frequency Fb by up-sampling from Fa to the least common multiple of both Fa and Fb, and subsequently down-sampling to Fb. When up-sampling a signal k times, each sample is multiplied by adding (k-1) zeros. The up-sampled signal is interpolated using a low-pass filter and down-sampled m times by extracting each m-th sample.
- For example, when resampling noise signals (derived from Noise-Rom 0) from 20 kHz (up 4 times, down 5 times) into 16 Khz (Eurom 0 standard), a lowpass fir filter with a cutoff frequency of 16/2=8 Khz at a sample rate of 20*4=80 Khz has been defined (0.1 fs).
- Resampling a signal from 20 kHz into 16 kHz using the 1024 taps FIR filter will take about one hour per megabyte on a 286-12 mHz with co-processor. A program for designing 'windowed FIR filters' named WFIRDESN is also supplied. More details on the use of these programs can be found in the readme file on the distribution diskette.
- The program ADD allows to add signals with a specified gain factor and an offset value, which both can be determined with the SLM program. Also for the first signal, a time-offset measured in samples can be specified. This feature allows for a random start point in the noise files.

- Each i-th output-sample is calculated according to the following expression: Out[i] = Gain0 * (Input0[i + SampleOffset0] - Offset0) + Gain1 * (Input1[i] - Offset1) No indication is given if the 16 bits dynamic range is exceeded.

Developing lab:

TNO Institute for Perception
P.O. Box 23
3769 ZG Soesterberg
Netherlands

Tel: (31) 34 63 56 230

Fax: (31) 34 63 53 977

Contact: Jan A. Verhave

E.6 SAMITPRO (SAM Iterative Proportional Fitting)

Version: 1.1, March 1992

Hardware / OS requirements: Low Level SESAM Workstation

Description:

- A software tool for analysis of interaction between recognition performance and test data characteristics. Examples of data characteristics are speaker characteristics such as sex, age, dialectal categories, educational level, weight, etc.
- Other types of characteristics could be recogniser type, testing-site, microphone, testing environment, etc.
- All these examples are that they can easily be represented by subsets/sessions within the SAM_SCOR software, so that test performance can be automatically derived for each subset/session by means of SAM_SCOR.
- Subsets/sessions are to be defined in the SAM_SCOR control file which can easily be set up using the SAM tool RISE.
- Input to SAMITPRO is provided by applying the SAM_SCOR to a series of result files from recogniser experiment(s), so that the performance is sampled in each cell of the input dimension.
- SAM_SCOR outputs the performance for each subset/session in a DBMS file, which can subsequently be read by SAMITPRO.
- SAMITPRO places the recognition performance in cells in a multi-dimensional table. Each dimension represents one characteristic, i.e. cells are described by a number of categories that are mutually exclusive. After filling up the table, hypotheses concerning the interactions of data characteristics (dimensions) can be evaluated by means of log-linear model fitting.

From this analysis conclusions can be drawn concerning the correlation between speaker characteristics and recognition performance: for instance whether some characteristics have an impact on performance or whether a complex interaction exists between several characteristics in combination.

Further work aims at extending SAMITPRO to provide the capability of correlating recognition performance with speaker dependent speech inherent parameters derived by SAM_SPEX.

Developing lab:

Speech Technology Centre
Institute of Electronic Systems
University of Aalborg
DK-9220 Aalborg, Denmark

Tel: (45) 98 15 85 22 - 48 75

Fax: (45) 98 15 67 40

E-mail: oa@stc.auc.dk Fredrik Bajers Vej 7

Contact: Ove Andersen

E.7 SAM_REC0 Isolated Word Recogniser

Version: 1.0 November l990 (corrections released January l991)

Hardware / OS requirements: Standard SESAM workstation, equipped with Oros AU21 (or AU 22) DSP board and the library DISK_TRA of the recording software that accompanies OROS DSP boards

Description:

- The software implements an isolated word recogniser.
- A word is recorded and endpoint detection is then carried out by analysing an energy contour and the zero crossing rate.
- The endpointed word is preemphasised and windowed by a Hamming window.
- Discrete Fourier transforms are calculated over the word.
- The transform samples are filtered by a filter bank of triangular filters, linearly distributed in frequency up to 1 kHz and logarithmically distributed above 1 kHz.
- From the filter outputs cepstrum coefficients are calculated.
- The time warping uses eight cepstrum coefficients per frame, allows local time distortions between 50% and 200% and is based on a symmetric slope constraint with P=1.
- The word chosen is the word among the reference templates with the smallest distance to the recorded word.

Developing lab:

Telia Research
Systems Research, Spoken Language Processing
Rudsjoterrassen 2
S-136 80 Haninge
Sweden

Tel: (+46) 8 707 55 72

Fax: (+46) 8 707 55 96

E-mail:gunnar.hult@haninge.trab.se

Contact: Gunnar Hult, Room A309

E.8 SAM_SCOR (Sam Input Assessment Scoring Software)

Version: 3.1, 3. March 1992

Hardware / OS requirements : Low Level SESAM workstation

Description:

- SAM_SCOR is a software tool for scoring results from automated testing of recognition devices attached to the SESAM workstation.
- SAM_SCOR aims to cover the main classes of recognisers, from small vocabulary speaker-dependent isolated-word systems to medium/large vocabulary speaker independent continuous speech recognisers.
- For isolated word recognisers, standard statistics comprise correct recognition, miss, substitution, correct rejection and false alarm scores. Confusion matrices and similarity matrices are listed and the confidence for the conducted experiments is estimated. Special attention is given to the scoring of recogniser testing in which CVC type words are applied.
- Other features are:
  - a detailed listing of the different error conditions during testing (miss causes)
  - intergroup and intragroup scoring
  - the option of using the McNemar test for comparing recognisers that have been tested on the same database
- For connected/continuous speech recognisers, SAM_SCOR acts as a preprocessor for running the NIST string alignment and scoring program (Release 3.0, modified to run on SESAM, and distributed along with SAM_SCOR with the kind cooperation of David Pallett, NIST).
- Apart from the result files from testing a recogniser, SAM_SCOR needs as input a scoring control file in SAM file format, specifying the general conditions and type of scoring to be carried out. This file may easily be produced manually (in a text editor) or be generated by applying the SAM software package RISE. The output of SAM_SCOR is both a standard documentation file, which is easy to interpret, and a DBMS file in SAM file format, which relates recogniser performance measures to speech material and speaker characteristics.

Developing lab:

Speech Technology Centre
Institute of Electronic Systems
Fredrik Bajers Vej 7
DK-9220 Aalborg, Denmark

Tel: (45) 98 15 85 22 - 48 67

Fax: (45) 98 15 67 40

E-mail: bli@stc.auc.dk

Contact: Borge Lindberg

E.9 SAM_SLM

Version: 2.0, 24 November 1991

Hardware / OS requirements: Low Level SESAM Workstation

Description:
- The Speech Level Measuring program, SAM_SLM, has been developed in order to determine the speech level of a signal file. The method implemented for measurement of the signal level is based on a level distribution histogram of a squaring and low-pass filtering of the signal waveform, and the RMS value is calculated from this histogram. In addition the program permits the user to determine the DC offset of the signal file, too.
- The program is self explaining, ie. invoking the program without parameters results a display of usage. Even though SLM stands for Speech Level Measurement the program can be used to determine the levels of noise files as well.

Developing labs:

1. TNO Institute for Perception
   P.O. Box 23, 3769 ZG Soesterberg
   Netherlands
   Tel: (31) 34 63 56 230
   Fax: (31) 34 63 53 977
2. Jydsk Telefon
   Sletvej 30
   DK-8310 Tranbjerg
   Denmark
   Tel: (45) 86 29 33 66
   Fax: (45) 86 29 90 68

Contact: Jeroen G. van Velden, Sven Danielsen

E.10 SAM_SPEX (Speech Parameter Extractor)

Version: 2.0, 25. Feb. 1992

Hardware / OS requirements: Low Level SESAM Workstation

Description:
SAM_SPEX provides an analytic basis for the definition of recogniser performance. The SAM project has adopted six speaker dependent speech parameter defined within the British Alvey project, MMI 132 STA. SAM_SPEX version 2.0 implements the extraction of these parameters from isolated words. The parameters measured are:

- Speaking Rate: The duration of each token is calculated using an energy-based endpoint algorithm (time resolution 1 msec) and related to the mean duration for all tokens of the same utterance.
- Energy: The ratio of the mean energy and the peak energy of the utterance.
- Pitch Frequency: The mean and variance of the pitch frequency estimated within the voiced parts of the endpointed utterance using a standard SIFT (Simple Inverse Filtering Tracking) algorithm.
- Voice Quality: The ratio of the energy between 0–2 kHz and 0–fs/2 Hz (fs = sampling frequency).
- Vocal Tract Area Distance: Calculated in two steps:
  1. The average area vector is computed.

2. The vocal tract area distance is computed as the Euclidean distance between the average vocal tract area vector and the average of all vocal tract area vectors for tokens of the same utterance.

- Pattern Congruence: Calculated in two steps:
  1. The output values of an eighteen channels filter bank are computed every 10msec of a token.
  2. A dynamic time warping algorithm is used to calculate the distance between the token and all other tokens of the same utterance.

  At present the definition of the parameters is under review and an improved definition is in preparation.

Developing lab:

Jydsk Telefon
Sletvej 30
DK-8310 Tranbjerg – J
Denmark

Tel: (45) 86 29 33 66

Fax: (45) 86 29 90 68

Contact: Sven Danielsen

NB: This software is currently only available to SAM partners as the definition of the parameters is under review and an improved definition is in preparation.

E.11 SAMTRA (SAM TRanscription Analysis)

Version: 1.0, June l991

Hardware / OS requirements: Standard SESAM workstation

Description:

- The software package SAMTRA normalises and verifies transcriptions in SAMPA notation and computes phoneme and diphoneme distribution statistics. The data flow between the three modules of SAMTRA is via the intermediate files SAMTRA.NOR and SAMTRA.STA (which are accessible for further developments). The modules perform the following actions:
  1. Module 1: Transcription verification
  2. Module 2: Computation of phoneme and diphoneme distribution
  3. Module 3: Flexible printing of statistics
- Procedure:
  1. SAMTRA reads two input files, a SAMPA symbol file and a SAMPA transcription file, and tries to write a normalised transcription to the file SAMPTA.NOR.
  2. If there are errors in the transcription file, SAMTRA prints them out on the screen and to the file SAMTRA.NOR.
  3. If the normalisation procedure succeeds, SAMTRA reads the normalised transcription file SAMTRA.NOR and computes the phoneme and diphoneme distributions, and writes them to SAMTRA.STA.
- The text output of SAMTRA is dependent on the file SAMTRA.SET (which overrides the default printout constants) and command line switches.

Developing lab:

Faculty of Linguistics and Literary Studies
Universität Bielefeld
P 100131
D-33501 Bielefeld

Tel: (49) 521 106 3510

Fax: (49) 521 106 6008

Email: gibbon@asl.uni-bielefeld.de

Contact: Dafydd Gibbon

E.12 SOAP (Speech Output Assessment Package)

Version: 4.0, March l992

Hardware / OS requirements: Standard SESAM workstation, with AU21/22 OROS board

Description:

- SOAP is a PC-based system for outputting speech and collecting the responses of a listener.
- The protocol (in terms of audio/visual cues, the file/path/part of speech files, time-out error strategy, etc.) can be defined by the user.
- The listener can enter responses via the keyboard or using a mouse/screen interface with a user-defined screen display.
- The software comes with demonstrations of SAM standard assessment protocols for testing at four levels: segmental, suprasegmental/SUS, prosodic and quality.
- Various utilities allow automatic randomisation of test sequences, the interpretation and scoring of test responses and automatic rule-based generation of segmental test files.

Developing lab:

Dept of Phonetics and Linguistics
7406 University College London
Gower Street
London WC1E 6BT

Tel: (+44) 71 387

Fax: (+44) 71 383 0752

Contact: Kate Jones

E.13 PTM (Parametric Test Manager)

Version: 1.0, March l994

Hardware / OS requirements: Extended SESAM workstation:
(Windows operating system and a MCI-compatible sound card, 800x600 pixel graphics system). There are no special memory or hard disk requirements. The hard disk capacity depends on the amount of speech files which the user wants to use. The PTM software needs 4.5Mb for installation where 2.4Mb are used by the sample test files.

Description:

The new Speech Output software (SO software) of the SAM-A Project was designed to provide a tool which covers most aspects of speech assessment. This tool is called Parametric Test Manager (PTM). The PTM system consists of different databases and programs. Hence PTM is not a single application, but a software package.

The PTM includes the following features: (1) resource data management, (2) test management modules and (3) demonstration test. The resource data management includes the facilities to allocate stimuli, their acoustical realisations and subject data. From these data test corpora or subject groups can be created.

The new Speech-Output software does not only deal with data representation, but also with execution of a test. The software provides facilitites to support people who want to carry out a certain test. A test execution can be divided into three phases: (1) test definition and preparation, (2) stimulus presentation and (3) scoring and documentation. The PT design includes applications to perform typical tasks as data acquisition, test list generation and session supervision.

Developing lab:

Ruhr-Universität Bochum (RUB), GERMANY
Lehrstuhl für allg. Elektrotechnik und Akustik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

Tel: (+49) 234 700 6652

Fax: (+49) 234 7094 165

Email: hegehof@aea.ruhr-uni-bochum.de

Contact: Thomas Hegehofer

Appendix F

EUROPEC recording tool

EUROPEC : [EUROpean Programme d'Enregistrement de Corpus]

License : non commercial

Author : J. Zeiliger

Task :

- Record and playback
- Protocol-driven recording session.
- Visual and acoustic prompts.
- Start and endpointing of utterances
- Signal and label files production

Description : This is a corpus prompting/recording software making easier the recording of speech databases. It ensures constant recording conditions for various speakers and/or corpus. The recording session is driven by a protocol file (user-made) which describes the way items (components of the corpus) will be displayed on the screen.

Components :

- "Standard SAM format" files production.
- One or two-channels recording and listening facilities.
- Three recording modes (manual, automatic, continuous).
- Parameter adjustment (sampling rate, input and output gain, etc.).
- Sound levelmeter.
- Free / operator controlled session.
- Error treatment.
- Acoustic prompt.
- Easy data verification.
- Safeguarding files.
- Easy data retrieval and manipulation.
- Simplified session.
- User's Guide and many example files.
- User-customizable in any language.

Modularity :

Extensibility :

Size : Version 4.11: 170 kbytes.
(data: examples of speaker, corpus, protocol and label files provided.)

Implementation : Microsoft C (v5.1)
TMS code for the I/O board

Platform : SESAM workstation (PC-DOS based), with OROS AU21/22 I/O board.

Languages : All.

Retargetability :

Orthography : Labelling via ASCII.

Examples : The program is used in laboratories all over Europe.

Status :

- robust (used in over 30 labs).
- stable
- Windows release under consideration.

Documentation : User's guide (30 Pages) SAM-ICP-O45 , March 31st, 1991

Upgrades : No foreseen upgrade. Windows implementation under consideration.

SourceCode : Provided to members of ESPRIT projects only

Consulting :

Format : 1.4 Mb HD floppy disk

Price : One floppy and mailing costs.

Restrictions :

- Research purpose only.
- Citation

Contact :

Address : Institut de la Communication Parlee, Grenoble, France.

The EUROPEC prompting/recording tool has been clearly designed to make easier the recordings in large recording campaigns for the constitution of speech databases. However, it provides also simple and basic functionalities. Some particular points must be developed here:

Data authentification

The EUROPEC software ensures automatic and consistant filenaming of the files it produces. The uniqueness of a filename during a recording campaign is ensured by an incremental serial number. For each recording session, the components: speaker, corpus to be recorded, recording conditions, protocol to be used, have to be registered in the relevant DB files.

Protocol of corpus

The protocol-based approach uses a protocol file that contains instructions to be executed by the program during the recording session. These instructions address the prompting of items on the screen and the design of the session. They define how and when items will be displayed, and allow training phases before the recordings. Here are some examples of the relevant codes implemented:

- CLS: CLearScreen and put cursor to 0,0
- VON: Video attribute ON: center, blink, inverse, underline, bright...
- VOF: Video attribute OFF
- BEL: ring bell
- MSG: message to display according to valid video attributes
- NWL: newline on screen
- PAU: n (integer) = duration of the pause in seconds
- TRN: set training mode, no recording
- CMT: comment
- GET: n (integer)
  n: number of items to read in the corpus text file, then to display and record (if recording mode is activated) or not (if training mode is activated)
- LOP: -n, m loop in the protocol file
  n: number of one-line-back in the present file m: number of iterations
- REC: recording mode activation
- RET: waiting for a RC (return)

In a research environment a prompting/recording tool can easily be extended or retargeted with usefull functionnalities to cover other experiments and researchers needs. For instance, local variants of the EUROPEC software have been successfully produced to perform adequate experiments in the following domains:

Perception: The acoustic prompt can be used to perform perception tests. Stimuli are played to the listener that is then presented menu options to indicate his answer. The answers are stored in the label file (results of the test).

Prosody: On-line listening facility is provided when the recordings aim at matching a particular ('spontaneous') prosodic profile. The session is a sequence of record, listen/accept/reject (menu options), record, listen/.. phases.

Acoustic: The acoustic prompt can be used to play blank noise during articulatory positioning then simultaneous I/O is used to record the production of the vocal tract excitated by a blank noise.

All these examples demonstrate that prompting, recording and playback facilities can appear in different ways in various experiments. The protocol-based approach as initiated in EUROPEC is adapted to the definition of the experiment session by the user. This approach must be clearly enhanced: a recording tool should be better considered as a software platform providing basic functionnalities as signal input/output, acoustic or text prompts, storage of labels or results, mouse driven menu. The modularity of these functions would allow the user to design his experiment and write the corresponding protocol file in a high level language; the protocol being then interpreted by the program all along the session.

Appendix G

EUROM-1 database overview

The first SAM database, EUROM 0, was the precursor to the much more substantial EUROM-1 recording undertaken in SAM. This database has been widely distributed on a single CD-ROM and contains five hours of speech material recorded with 16 kHz 16 bit sampling using a condenser microphone in anechoic rooms from four single accent speakers in each of five languages (English, French, Dutch, Italian and Danish). NATO single and triple digit sequences are recorded using only the speech signal, and a continuous speech passage, with a common numeric theme across languages, is also recorded here using two channels — with both speech and laryngographic inputs. This CD-ROM has been used extensively in the SAM project and is specified as a reference for calibration of the speech input assessment tools.

A subsequent major data collection activity has resulted in the collection of a very substantial amount of data which is unique in the size and the breadth of its coverage of different European languages. The EUROM 1 database contains more than twelve hours of data for each of the eleven European languages covered: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish. The material is of high acoustic quality, and was selected specifically for use in the assessment of speech technology devices.

The control software used in making the recordings provided for orthographic labelling of the data and alignment of the text and signal portions at the level of the prompt units. Phonotypical transcriptions have been made separately for all languages and broad phonetic labelling using SAMPA (SAM Phonetic Alphabet) has been applied to some parts of the database. Language subsets of EUROM 1 are now available on CD-ROM, and the provision of EU funding is planned to ensure availability of all recorded material of this important reference resource. A minimum of three CD-ROMs are planned for each language.

Very careful consideration was given to the homogeneity of the data across languages. This was achieved by the use of identical recording protocols, which were specified earlier in the project and applied using standard software tools, and by a careful definition of the speech content, such that each language was represented in the same way wherever possible. The speech (and calibration) recordings were made in acoustically treated rooms using calibrated condenser microphones and, in addition to the acoustic signal, larynx activity was recorded simultaneously, using a laryngograph, for samples of the speech in each language. The use of anechoic condenser microphone recordings permit the subsequent imposition of post-production effects. The recordings were made using the SAM agreed standard of 20 kHz 16 bit sampling to ensure optimal signal representation; inter-utterance acoustic background signals were also preserved. The protocols defined for collection of database materials have been developed to provide guidelines on recording procedures and quality criteria for use in the wider European Speech Community. These are provided in full in Appendices B and C.

The database material specification for each language is as follows:

- C(C)VC(V) material in isolation and in context, in the range 60–100 items per language

- 100 selected numbers from 0–9999, providing complete coverage of the phonotactic possibilities of the language number system
- 40 short passages comprised of five thematically linked sentences
- 50 sentences composed to compensate for phonemic frequency imbalance in the passages
- 5 pairs of context words for use with C(C)VC(V) material

The database has been designed with a hierarchical structure to maximise its usefulness both for training and testing different types of speech technology device and for more basic research including inter-language comparisons. In each language, material was recorded by 60 subjects, 30 female and 30 male, each of whom recorded 100 numbers, 3 passages and 5 sentences. Of these a "few talker" subset of 5 females and 5 males made extended recordings: isolated C(C)VC(V) items, 500 numbers, 15 passages, 25 sentences. A further "very few talker" subset of one female and one male, selected from the 10, additionally recorded the contextualised C(C)VC(V)s and the 5 context words, using both acoustic and laryngographic signals. A total of 660 speakers thus recorded over 130 hours of data, making this a very substantial multi-lingual resource with many different applications.

Appendix H

Polyphone project overview

Use of the Dutch Polyphone Corpus for Application Development

T.I. Boogaart[1]& L. Bos[2]& L. Boves[1,2,3]

[1]PTT Research, [2]SPEX, [3]Nijmegen University

Paper presented at IVTTA 1994 Japan.

Abstract

This paper gives a short description of the creation of the Dutch POLYPHONE corpus. It then proceeds to show how that corpus has already been put to use in the development of a number of practical applications and how it is used for research purposes.

Applications that are touched upon include a fully automatic Train Time Table Information system and the automation of collect call and phone card calls.

The paper concludes with remarks on and recommendations for future corpus development.

H.1 Introduction

From a technological and algorithmic point of view speech recognition in the Public Switched Telephone Network (PSTN) has reached a quality that is sufficient for a range of commercially attractive applications, at least in the U.S.A. and Japan, where large populations use the same language and recognisers can be trained for the native language. Although the recognition technology is basically language independent, similar applications in most European countries lag behind Japan and the U.S. This technology gap has no simple explanation; it is due to a complex of factors, some of which are cultural in nature, while others are organisational or financial.

It is probably true that Europe has a cultural attitude that is less open towards new technologies than the U.S.A. and Japan. However, all European language areas have less inhabitants than either the U.S.A. or Japan. In terms of native speakers German is the biggest European language, with some 100 million speakers, if one is willing to include the German part of Switzerland and Austria, even if the varieties of the language spoken in these countries differ from standard German at least as much as American and British English. Compared with 245 million Americans who speak English and 125 million Japanese, it is evident that return on investment in speech technology will be smaller in Europe.

Number of speakers is not the only factor determining traffic volume for a given application; cultural factors also play a role. The number of calls placed from residential areas in the U.S.A. is substantially higher than in most European countries, because the telephone is used in a different way. Even within Europe the differences are large. In the Netherlands the transition from local calls to a single national premium rate number for public transport information caused virtually no protests from the public, whereas the German Railway Company feels that such an action would cause an uproar in their country. Perhaps the fact that the transition was accompanied by a substantial improvement of the service (more information — also on busses and trams — and more accurate information and shorter waiting times) has helped to keep protests down. But the fact that premium rate services are becoming widely accepted in the Netherlands has also been of importance. Moreover, toll free numbers as a weapon in the battle for the consumer plays virtually no role in Europe. Thus, the fact that a certain application has proved to be a commercial success in the U.S.A. cannot be simply generalised to the European market.

In a situation where the commercial viability of speech technology applications is not always evident, cost and ease of application development become automatically of paramount importance. Managers lose interest in new technologies that require substantial capital investments if the return is insecure. They also quickly lose interest if application development takes longer than a few months.

H.2 Reusable resources

Both investment capital and development time can be reduced substantially if reusable tools and resources become available. Speech corpora to develop and train recognition systems are the most expensive and time consuming resource needed to build applications that use speech recognition in the PSTN. Moreover, annotated speech corpora are also indispensable for running objective performance tests of off-the-shelf technology that has proven adequate performance in some languages, but not yet in the language of the country of interest. Because large commercial interests are at stake with the introduction of automated services using ASR, one can understand that marketing managers are reluctant to accept new technologies as long as there is no definite proof of their adequacy for the home market.

In the fringe of ICSLP-'92, the international Cooperating Committee on Speech Databases and Assessment (COCOSDA) defined guidelines for a corpus that should be able to solve at least some of the problems slowing down the uptake of speech technology applications. Comparable corpora, now known under the name POLYPHONE, should be recorded for as many different languages as possible. In the course of 1993 these guidelines were made more concrete by the Linguistic Data Consortium (LDC), which specified a recording protocol for American English and Spanish [1,2]. Taking those protocols as a point of departure, PTT Research and the Speech Processing Expertise Centre SPEX set out to record the Dutch POLYPHONE corpus.

H.3 The Dutch POLYPHONE corpus

H.3.1 Recording workstation

The recording workstation used for POLYPHONE was based on an Aculab telephone interface, a Rhetorex Voice Card and driver software, Show-'n-Tel application development software, and a 16 port operational license, in an OS/2 PC. Each item recorded was stored in a separate file; all files were copied to a Unix network for transliteration and archiving.

The recording platform was set up to record and store the speech signals in A-law format. Because the Dutch PSTN is completely digital, the acoustic quality of the recordings is determined by the characteristics of the caller's local loop and the background noise in the caller's location.

H.3.2 Speaker selection

Prospective callers received a personalised letter. Originally, we aimed at collecting 5000 speakers, uniformly divided over a large number of cells, defined according to four criteria, viz. (1) geographical region, (2) socioeconomic status, (3) sex, and (4) age. It should be emphasised that the uniform sampling of the cells was mainly motivated by scientific arguments: in order to find the funds for creating the corpus it was necessary to make it attractive for a wide range of linguistic research, including sociolinguistics and dialectology. Perhaps part of the speakers in our corpus will not be heavy users of the automated services that can be developed by means of the Dutch POLYPHONE corpus. However, we trust that a wide coverage of language and speech behavior will lead to applications that are more robust than what could have been obtained with recognisers trained with much more restricted speech material.

Geographical region, operationalised as the province in which the speaker lives, is the best practically feasible approximation to regional accent and dialect background. By sampling provinces, we sidestep the unsolved problems of how many different regional accents should be distinguished and how these should be defined. Due to the very uneven distribution of the population over provinces it appeared to be practically impossible to get equal numbers of speakers from each province [3,4].

Socioeconomic status is difficult to define, and even more difficult to assess reliably from what respondents are willing to say. We decided to approximate status on the basis of the education level of the respondents. We distinguished three levels, viz. (1) only primary school, (2) secondary school and (3) college/university. Using hindsight, this division was somewhat unfortunate: in formal terms almost every person younger than about 60 has been to school until at least the age of 16, so only a very small proportion of the population falls into the first category. Thus, it is not surprising that we were able to recruit very few speakers who said that they had no more than elementary school. The numbers in the remaining two classes are approximately equal.

We distinguish four age classes, i.e., under 20, between 21 and 40, between 41 and 60, and 61 and older. Information about age is acquired by asking the respondents for their year of birth. Since we set a minimum age of 16 for participation, the under 20 group is much smaller than the other groups. The group of 61 and older is also underrepresented. The group between 20 and 40 is about 50% larger than the group between 40 and 60.

H.3.3 The speech material

The speech material recorded in the POLYPHONE project consists of 32 read items, 14 extemporaneous answers to printed questions, and 4 extemporaneous answers to questions not printed on the response sheet.

The material *to be read* consists of the following items:

- 12 numbers (digit strings, natural numbers, guilder amounts, postal codes)
- 2 city names
- 4 application words (e.g. *start, accounts payable, tomorrow, etc.*)
- 3 spelled words, designed in such a way that all letters of the alphabet occur approximately with equal frequency
- 1 date and 1 time
- 4 sentences with an application word
- 5 phonetically rich sentences, designed in such a way that each speaker would realise all phonemes of Dutch at least once.

The following list of *printed* questions is asked:

- Is Dutch your native language?
- Did you ever live in another country than the Netherlands?
- Would you volunteer to participate in another study like this one?
- What is your last name?
- What is your house number?
- What is the name of the street you are living?
- What is your postal code?
- In which city do you live?
- In which cities did you grow up?
- Are you a man or a woman?
- What is your age?
- Which code represents your education level?
  (1 = primary school, 2 = high school, 3 = college/university)
- Please, say a familiar phone number.
- Please, give your comments about this recording session.

The following *unprinted* questions are asked:

- Please, spell your name.
- Are you calling from your home phone?
- Are you using a cordless telephone?
- What time is it now?

H.3.4 Postprocessing

Postprocessing was done at the Dept. of Language & Speech, Nijmegen University, using software running on a PC under MS-Windows, equipped with a Pro-Audio board. Whenever the answer is predictable (i.e., in all cases where the caller is supposed to read preprinted material), the expected answer is displayed on the screen.

Postprocessing consists of four steps, viz. (1) word-by-word transliteration of all items, (2) transliteration of extra sounds and noises, (3) collecting demographic data, (4) assessing the quality of all items. The students who carried out the work were instructed to do the tasks in exactly this order. On average, a recording session took slightly less than 20 minutes to process.

We have seen that the sentences contain substantially more disfluencies than the other items.

H.4 Use of POLYPHONE in application development

At present, a number of application development projects are under way in the Netherlands that employ ASR techniques. In this section we describe the way in which POLYPHONE helped to enable these projects.

H.4.1 Train time table information

In a collaborative project between PTT Research, Philips Research Aachen and the Netherlands Organization for Scientific Research NWO we are working on the development of a Dutch version of the Train Time Table Information service that is already available for German and that is described in another paper in these Proceedings. The system we have in mind can best be characterised as a guided mixed-initiative dialogue: the system will ask specific questions, like *From where to where do you want to travel?*, but it will allow the user to give under- and overinformative answers. If an answer is underinformative, the system will ask explicit questions to elicit the missing information. When an answer is overinformative, for instance when the caller adds the desired arrival time to departure and destination station, the system will try to process that additional information too. If time and date information is not offered spontaneously, again the system will ask explicit questions to obtain it.

Clearly, a Train Time Table Information System is an application intended to be used by the general public. Moreover, most users will call the system only occasionally, so that one cannot rely on users getting acquainted with the peculiarities of the service. Pilot experiments carried out by the *Nederlandse Spoorwegen*, the Dutch Railway Company, have shown that the part of the public who need Time Table information are not able nor willing to deal with a menu-based interface. Thus, there seems to be no alternative for starting an automated service with a mixed-initiative dialogue system.

To implement a Dutch version of the Train Time Table Information System a number of steps must be taken:

- Train a phoneme based continuous speech recogniser for Dutch, designed to handle telephone speech.
- Design the phonemic recognition lexicon for the application.
- Design a language model suitable for use in this application.

The POLYPHONE corpus has been instrumental in all these steps.

H.4.1.1 Training of the recogniser

For the development of the phoneme based recogniser use has been made of the phonetically rich sentences in POLYPHONE. Following the approach that has proved successful for German we have started with a recogniser based on context independent phone models. The recogniser has been trained assuming that the automatic grapheme-to-phoneme transcription of the transliteration data is correct. That assumption is probably wrong to some extent: Dutch has quite some pronunciation variation at the phonemic level. At the time of this writing we are using the POLYPHONE recordings for an empirical investigation of the range of that variation. Up to now, researchers had to be content with rather subjective ideas about this crucial issue.

H.4.1.2 Building a phonemic lexicon

An essential part of the recognition engine in a Train Time Table system is a lexicon comprising phonemic representations of station names. Here too, there is nonnegligible pronunciation variation. Since all station names have been read by at least five speakers, we can use the POLYPHONE recordings to make an inventory of the pronunciations. This is especially relevant for the names of the smaller stations, since pronunciation variants for larger stations can also be collected by other means.

H.4.1.3 A model for *yes/no* expressions

Virtually every information dialogue contains *yes/no* questions. In previous applications of ASR in telephone information systems for the general public it has appeared that there is quite some variation in the way people answer these questions. Since POLYPHONE contains four *yes/no* questions, all to be answered spontaneously, we have a substantial amount of data to build a model of the answers.

The analysis of the answers that we have performed so far confirms the existence of substantial variation; yet, it appears that the very large majority of the expressions adhere to a simple schema, so that it is easy to build a model. We have seen a large difference between the two items for which we expected affirmative responses: Almost 93% of the subjects used a single word (e.g. *ja, jawel, jazeker*) to confirm the assumption that Dutch was their native language; the proportion of one word confirmations dropped to 75% for the question whether the caller was willing to participate in another recording session. Very few callers said "no", but the way in which they expressed their confirmation was much more varied.

83% of the subjects used a single word (e.g. *nee, neen*) to deny that they ever lived abroad for an extended period of time. Most of the people who used more complicated expressions did so to tell us in what foreign countries they had lived. 80% of the callers used a single word to deny that they were using a cordless phone; over 13% of the callers said they were using a cordless phone.

A detailed analysis of the more verbose answers showed that only a very small proportion of the affirmative answers contained *no*-words and that the same is true for negative answers and *yes*-words.

Another observation worth mentioning is that politeness forms like *yes, sir; no ma'am* were virtually absent. This may be due to the fact that the *yes/no* questions were located in the last part of the recording session, when the callers should be fully aware that they were talking to a computer. However, it is also possible that what we see reflects the growing casualness in the Dutch society, where 'speaking with two words' is quickly becoming the exception rather than the rule.

All these observations confirm our expectation that the NLP module in our system should be able to handle the large majority of the *yes/no* expressions that will be used by the callers. In confirmation sub-dialogues in an information system (e.g. after the caller has given departure and/or destination station) the language model expects an affirmative expression, but negations may occur due to errors of the recogniser. The POLYPHONE corpus contains a number of examples of negations where confirmations were expected. We are working on a closer analysis of these cases, to find out whether they contain systematic syntactic structures that could help in making the language model more specific.

H.4.1.4 Time and date expressions

In previous experiments with information and reservation systems it has appeared that — quite surprisingly — linguists do not have accurate models of the way in which people express times and dates. The POLYPHONE corpus contains a large number of these expressions. Currently we are analysing the syntax of these expressions in order to build models for use in the Train Time Table Information system. Unfortunately, the expressions used by the POLYPHONE speakers are to a large extent determined by the way in which the items were printed on the response sheets. No spontaneous expressions of dates or times were obtained. This will make it very difficult to derive reliable estimates of the relative frequency with which individual expressions will occur.

H.4.2 Phone card services

PTT Telecom has started a campaign to promote Operator Services, like Collect Call and Phone Card calls. The PTT Telecom Phone Card is marketed under the name of *Scopecard*. As in other countries, operator services are expensive, mainly because of the high costs of the personnel. Thus, it is obvious that Telecom is looking for ways in which these services can be automated.

One simple way of automating Phone Card calls is to connect customers with an IVR platform that handles the recognition of the card number, the PIN code and the number to be dialled via DTMF; this is how a large proportion of domestic calls are automated. However, few operators offer the capability to connect toll-free lines to an IVR application in another country. Also, in many of the countries where PTT Telecom customers spend their vacation rotary dial phones still form the large majority. Thus, large scale automation of card services implies the deployment of ASR. The same goes for automatic collect calls, of course.

H.4.2.1 Automating card services

For applications like automated card services it is not enough to have a recogniser that can handle isolated or connected digits *in abstracto*. For the real-world performance of the application a solid knowledge of the way in which customers pronounce card and phone numbers is at least as important, since that knowledge can be exploited in designing application specific language models. POLYPHONE has provided us with a rich source of information about the way in which Dutch people express these numbers.

Two items in POLYPHONE that are related to telephone numbers were analysed. The first pertains to numbers read from the response sheet. All these numbers were printed in the same format, i.e., area code, dash, subscriber number (e.g. 020–5252183). The second item consists of answers to the question *Please, say a familiar telephone number.* In discussing the results we will use the term *digit* for the words *zero, one, ..., nine*; the term *number* will denote numbers between 10 and 99.

Presently, the Dutch PTT's number plan has two groups of area codes, one comprising three digits (like 020 in the example above) and one comprising five digits (e.g. 08894). Subscriber numbers can have four to seven digits. Because transliteration does not include intonation markers, it is not possible to discriminate between three and five digit area codes. We doubt whether the transliterators would have been able to parse all answers to the request to give a familiar number correctly.

The format of the read numbers is quite different from the format of spontaneously produced familiar numbers: in read numbers the proportion of *digits* is much larger than in spontaneously pronounced numbers. It is also worth mentioning that 18% of the read and 23% of the spontaneous numbers contain extra sounds, far more often preceding the number than following it.

POLYPHONE provides similar information about the way in which long card numbers and shorter PINs are pronounced. At the time of this writing these expressions are still under analysis.

H.4.3 Speaker verification

Advanced telephone applications will inevitably grow from pure information systems to mixed information and transaction systems. Security and fraud prevention then become major issues. Card Services of Dutch PTT is investigating the possibility of using speaker verification as one means of fraud combatement. In its original specification the POLYPHONE corpus is not suited for research into speaker recognition. Nijmegen University, in collaboration with the Dutch National Forensic Science Laboratory, have made additional recordings of 100 speakers who have called eight times, using different handsets. Half of the speakers is recruited from the Nijmegen area, the other half from the Hague, in order to minimise dialect differences. Also, the speakers form 50 pairs of brothers or father-son, in order to allow us to investigate whether speaker recognition techniques can be fooled by close relatives.

H.5 Requirements for future corpora

At this point in time it is too early to give very specific recommendations for future speech corpora. Yet, a number of things can be said. In future corpus projects an attempt should be made to obtain spontaneous expressions of time and date, in addition to the read expressions in POLYPHONE.

Speaker selection and recruitment is still a difficult issue. In the Dutch POLYPHONE project much time, effort and money was spent in order to get a maximally uniform sampling of a large number of cells. To a considerable extent, these efforts have been to no big avail. The major reason to strive towards uniform sampling was scientific: we wanted the corpus to be as attractive as possible for linguists and dialectologists, of course without interfering with the requirements of speech technology. The latter requirements are ill-defined. It is quite likely that applications like Train Time Table Information must deal with the public at large, including low income groups whose speech may differ from the general standard. More research is needed to clarify this issue.

This research was supported by the Foundation for Speech Technology, which is funded by the Dutch National Program for the Advancement of Information Technology (SPIN).

References

[1] Godfrey, J., Graff, D. & Martin, A. (1994) "Public databases for speaker recognition and verification." *Proc. ESCA Workshop Automatic Speaker Recognition, Identification and Verification,*, Martigny, 5–7 April 1994, pp. 39–42.

[2] Bernstein, J., Taussig, K. & Godfrey, J. (1994) "Macrophone: An American English Telephone Speech Corpus for the Polyphone Project." *Proc. ICASSP-'94,* Adelaide 19–22 April 1994, pp. I-81–I-83.

[3] Boves, L., Boogaart,T. & Bos, L. (1994) "Design and recording of large data bases for speaker verification and identification." *Proc. ESCA Workshop Automatic Speaker Recognition, Identification and Verification,*, Martigny, 5–7 April 1994, pp. 43–46.

[4] Damhuis, M., Boogaart, T., in 't Veld, C., Versteijlen, M., Schelvis, W., Bos, L. & Boves, L. (1994) "Creation and analysis of the Dutch Polyphone Corpus." *Proc. ICSLP-'94, paper 30.5.*

Appendix I

Overview of speech corpora

[1]

I.1 Overview

Spoken language is central to human communication and has significant links to both national identity and individual existence. With the increase in availability and capabilities of computing resources, there has been and will continue to be a large expansion in computer-based language technologies. These technologies include speech recognition and synthesis, vocal access to information retrieval systems, speech understanding (or spoken language) systems and spoken language translation. Central to progess made in spoken language technologies lie large corpora of speech with associated text, transcriptions, and lexica.

The structure of spoken language is shaped by many factors, including the phonological, syntactic and prosodic structure of the language being spoken, the acoustic environment in which it is produced, and by the communication channel. The speech signal is produced differently by each speaker, each with a unique vocal tract which assigns its own signature to the signal. Speakers have different dialects, accents and speaking rates, and their speech patterns are influenced by their emotional and physical state, and the context in which they are speaking (e.g., reading aloud, in conversation, giving a lecture) and the acoustic environment. Due to the many sources of variability in the speech signal, a great deal of speech data are needed to model different speech characteristics, and in particular, different dialects and accents.

Recent activities, such as the creation of the Linguistic Data Consortium (LDC) and the Center for Spoken Language Understanding at the Oregon Graduate Institute (OGI)[2] in the U.S. and the LRE Relator project in Europe, national efforts in Japan, Australia and China, as well as the international Coordinating Committee for Speech Databases and Assessment (COCOSDA), point out the growing worldwide awareness of the need for and importance of large, publicly available common corpora for the development and evaluation of language technologies, particularly speech recognition and spoken language understanding, as well as for the development and assessment of speech synthesisers. These corpora allow scientists to study, understand, and model the different sources of variability, and to develop, evaluate and compare speech technologies on a common basis.

[1]CAVEAT:
The present report has been achieved under the account of the Relator Project(LRE-62056). It is the fruit of the work of the speech partners involved in this project. They have tried to accurately represent the existing resources in Europe, after having contacted colleagues and compiled various sources of information. But they are sure that there are emission and errors in this status report and they apologise in advance and welcome corrections/additions/comments.An extended version of the survey is available as an annex of the Relator WP1 deliverable.

[2]OGI has collected an 11-language telephone-based corpus which has been used for common evaluation of language identification algorithms. This corpus is currently available through the LDC.

Corpora collection in Europe is the result of both National efforts and efforts sponsored by the European Community. Several ESPRIT projects have attempted to create comparable multilingual speech corpora in some or all of the official European languages. The first multilingual speech collection action in Europe was in 1989, consisting of comparable speech material recorded in five languages: Danish, Dutch, English, French, Italian. The entire corpus, now known as EUROM0 includes 8 languages: Danish, Dutch, English, French, German, Italian, Norwegian, Swedish. Other corpora resulting from CEC projects include: SAM/SAM-A EUROM1 (11 languages: Danish, Dutch, English, French, German, Greek, Itaian, Norwegian, Portuquese, Spanish, Swedish), ARS (Adverse Recognition system: Italian, English? **langs**), POLYGLOT (7 language IWSR database: Dutch, English, French, German, Greek, and Spanish, 5 language TTS database: Dutch, English, French, German, and Greek), ROARS (Robust Analytical Recognition System: Spanish, ?? **langs**), SPELL (Interactive System for Spoken European Language Training — French, Italian and English), SUNDIAL (Spoken language queries in the travel domain for English, French, Italian and German), SUNSTAR (Integration and Design of Speech Understanding Interfaces; English, German, Danish, Spanish and Italian), and ACCOR (cross-language acoustic-articulatory correlations: Catalan, English, French, German, Irish Gaelic, Italian and Swedish).

What follows is a brief status of linguistic resources for the European Countries, as well as a summary of some of the corpora resulting from European Community projects. This list results from the summary of the survey conducted in Europe, which complete version is presented in annex **??**.

I.2 Criteria for assessment of the situation of Spoken Language Resources

As our main concern is language, the situation of spoken language resources is assessed on a per language basis. This does not correspond exactly to the situation on a per country basis, as several countries may contribute resources to the development of the same language (for example, Belgium, Switzerland and France all have French linguistic resources), or one country may contribute resources for several languages (Switzerland may contribute to French or German, and the UK has produced the Oxford Acoustic Phonetic database which contains spoken data for several European languages).

While ideally all the European languages should be adequately represented, it is clear that from a commercial standpoint the importance of a language depends on the potential market demand. While in the European market all European languages are of potential interest, large companies tend to prefer English, German, French, Italian and Spanish. This leaves an even stronger need for multilingual EEC initiatives in this domain which can pay attention to the under-represented European languages, as in at least the immediate future such needs cannot not be expected to be filled by industrial demand.

To aid in assessing the current situation in Europe, we provide guidelines on characterizing the spoken language resources, and then summarise the main speech resources in Europe and identify the respective actors in their production. A more complete survey on a per country basis can be found in annex **??**.

I.2.1 Types and specificities of corpora

There are as many types of corpora as relevant factors which can be used to define them: speakers, texts, speech type, recording conditions, tasks and so on. Among this wide range of possible corpora, we may characterise them according to their intended use:

1. **Experimental research:** These corpora are widely used for speech technology development and assessment. Much of the basic material was collected several years ago, and more recent technology requires more advanced materials.

   1.1 Basic material: Numbers, Words, Sentences, Logatoms

   Number of speakers: medium (100–500)
   Several repetitions.

   1.2 Advanced material: Continuous speech, Passages, situated dialogue.

   Number of speakers: small to medium (10–200)
   Recently the trend has been to increase the number of speakers in such corpora

   1.3 Specific databases: multi-sensors corpora (Lx), articulatory, acoustic, video databases.

   Number of speakers: small
   These corpora tend to be relatively expensive to collect and may require sophisticated recording facilities and sensors, as well as specialised operators.

2. **General-purpose Telephone corpora:** These used for speech recognition and coding over the telephone. These type of corpora are relatively easy to obtain (the speaker only needs to call a specified telephone number) and relatively cost effective. However, with advances in communication technology, some of the problems currently posed by the limited bandwidth and noisy communication channel of today's telephones can be expected to disappear.

   Material: word lists, numbers, spelled names
   Number of speakers: large (several hundreds to several thousands)

3. **Application-oriented corpora:** For specific tasks and/or environments (many of which involve the telephone network). By essence of their application specificity, many of these corpora are not easily reused for other applications.

   Domains: Information retrieval — travel inquiry (train and flight information and reservation), leisure activities, telephone services.
   Vocal dictation — medical, legal and insurance areas. Adverse environments: Car database, Handicap applications.
   Number of speakers: variable (small to medium)

The two extremes in corpus type are on one side very specific corpora for fundamental research, which may require complex recording conditions with multi-channel recordings, and a low number of speakers, and at the other application-specific corpora which may be recorded over the telephone with a large number of speakers. In addition to the recorded speech signal, we must highlight the importance and effort required to ensure that the appropriate associated information is provided. This associated information depends heavily on the type of corpus, but at a minimum must include revelant speaker information, transcriptions (at a miminum an orthographic transliteration), prompt material in the case of read-speech corpora, lexica, noise or channel characteristics and details of the recording configuration.

I.2.2 Actors in speech resource production

a. **National and/or academic initiatives:** In every country where there is a history of speech research, academics (universities, speech research labs) produce the databases they need. However, these corpora tend to be specific to the needs of the producer and rest the property of the producer.

b. EEC initiatives: EEC projects are a catalyst of production. Both for basic research and application-oriented databases, they are a way of developing links between academics and industry. Corpora also tend to remain the property of the consortium. As the conortium members are both academic/research and industrial, the needs cover all areas.

c. Telecommunication/telephone sectors: Major telephone operators historically have been interested in speech technology, and most have their own research centers which collect the necessary corpora for their research activities. Most of these corpora are not publicly available.

d. Industry: Companies developing or integrating speech products need application-oriented databases. This is true both at national and international level, where foreign languages represent viable market opportunities. The data that can be provided by industry is varied, but for the most part unknown, other than that resulting from EEC initiatives.

The more of these actors present in a given country, the more we tend to find a developed speech community both with existing linguistic resources, but also with a strong demand for additional resources. As the speech community grows and the number of speech-based products extended, the amount of needed resources also grows.

```
a) need 1.1, 1.2, 1.3        provide 1.1, 1.2 with a), b), c)
                             provide 1.3  with a)

b) needs variable            provide 1.1, 1.2 with a), d)
                             provide 2 with c)
                             provide 3 with d), a)

c) need 1.1, 2               provide 1.1 with a), b)
                             provide 2 with b)

d) need 1.1, 1.2, 2, 3       provide 1), 2) with b)
```

So far, in reviewing the already existing resources, the presence of traditional actors, and the on-going projects, we assess the current situation as follows. Our starting consideration is that the under-represented European languages will need at minimum the resources that better-represented languages already have (at least the basic resources), and that well-represented languages will need still more resources. These needed resources will come from ongoing projects, and further needs can be foreseen through interviews with relevent actors in the speech research community and industry (ISC).

I.3 Summary of the current situation on a per language basis

I.3.1 DANISH (Denmark)

I.3.1.1 Existing databases and their actors

- Fundamental research: EEC initiatives produced basic/advanced material: EUROM1, EUROM0, SUNSTAR.
- Application-oriented: National initiative produced recorded/transcribed dialogues in the domain of flight ticket reservation and information.

I.3.1.2 Ongoing projects & new initiatives

- Experimental research: National initiative for advanced material (dialogue).
- Application-oriented: ONOMASTICA (EEC).
- Telephone: SPEECHDAT (Jydsk Telefon will collect a Polyphone type Danish Corpus).

I.3.1.3 Conclusion

There is a national academic effort. Telecom operator has a history of being involved in EEC projects and in the domain. Activities are carried out within National or EEC funded projects. The basic material exists (EUROM1) but is still not available. No other known industrial initiatives.

I.3.2 DUTCH (The Netherlands)

I.3.2.1 Existing databases and their actors

- Experimental research
  — academics: GRONINGEN Corpus (basic research), BLOEMENDAL corpus, DEMSI corpus, Collections of speech recorded on analogue tapes. In coordination with the Reusable resources task group of Elsnet, the Groningen corpus has been produced on CDROM for public distribution.
  — CEC: POLYGLOT, EUROM0, EUROM1, NOISE-ROM-0
- Application-oriented
- Telephone: Dutch POLYPHONE
- Lexica: CLEX (Dutch/German/English, distributed by SPEX and LDC)

I.3.2.2 Ongoing projects & new initiatives

- Experimental research: "Dutch National Corpus" (similar to BNC, 10 M words spoken material to be collected)
- Application-oriented: ONOMASTICA (EEC)
- Telephone
  — POLYPHONE: Dutch corpus of similar design, sponsored by SPEX and PTT Research.
  — SPEECHDAT (EEC)
  — NWO Priority Program: information system about public transportation via the telephone.
  — EUROCOCOSDA, COCOSDA
  — Eurescom related Telecom cooperation.

I.3.2.3 Conclusion

The speech community in Holland is very active. Active participation in ESCA (president is L. Pols), Elsnet, Eagles, Eurococosda, Cocosda. Strong cooperation between academics and telecom. Strong national effort. Industrial involvement unknown.
SPEX, the Dutch Speech Processing Expertise Centre located in Leidschendam, is interested to get involved, as the Dutch node, in a European infrastructure for database collection and exploitation. It distributes some corpora. EUROM1 to be made available on CDROM.

I.3.3 ENGLISH (United Kingdom)

I.3.3.1 Existing databases and their actors

- Experimental research
  — Academics: SCRIBE, HCRC Map Task Corpus: British 'Normal Speech Corpus (advanced), Oxford Acoustic Phonetic database (8 languages).
  — EEC: EUROM1, EUROM0, POLYGLOT
- Application-oriented
- Telephone: Bramshill corpus of British English Home Office Telephone recordings: (12 CDs) held by the LDC.

I.3.3.2 Ongoing projects & new initiatives

English is represented in the EUROCOCOSDA (TED) , EAGLES, and ELSNET initiatives.

- Experimental research: National BE-WSJ0CAM, British National Corpus (BNC)
- Application-oriented: SQALE, ONOMASTICA (EEC)
- Telephone: SPEECHDAT (POLYPHONE corpus to be collected by GEC Marconi)

I.3.3.3 Conclusion

Very active speech community in a variety of areas from fundamental research to technology development and applications. Many universities, companies, national telecom involved in the domain. Strong national effort, with active participation in EEC projects. Basic material seems to be available. EUROM1 to be made available as multi-lingual resource (organised by A. Fourcin, UCL). However at the current time, the only resources whose dissemination are well-defined are those managed in cooperation with LDC. Other cases are by local arrangement.

I.3.4 FRENCH (France, Belgium, Switzerland)

I.3.4.1 Existing databases and their actors

- Fundamental research
  — National initiatives produced basic/advanced material (BDSONS, BREF, BDBRUIT, ICY, SPOT)
  — EEC actions produced basic (EUROM0, EUROM1, POLYGLOT), specific (ACCOR) material.
- Application-oriented: There are indices of cooperation between academics (PSH/DISPE) and industry with the sponsorship of EEC (FREETEL, SPELL).
- telephone / telecom: Speaker verification corpus (Switzerland)
- Lexicon: BDLEX

I.3.4.2 Ongoing projects & new initiatives

French is represented in the main European (and world-wide) current initiatives such as COCOSDA, EAGLES, EUROCOCOSDA, RELATOR, SQALE, ONOMASTICA, FREETEL and (through Switzerland) Polyphone, TED-Martigny. Some of these initiatives deal with linguistic resource standardisation, dissemination and production. The very next resources to be produced in French will be under the sponsorship of

— the French-speaking network: FRANCIL (Réseau Francophone de l'Ingénierie de la Langue): speech synthesis, vocal dictation, person-machine dialogue
— EEC: Polyphone type French corpus by Philips (SPEECHDAT), timetable enquires over the telephone (RAILTEL).

I.3.4.3 Conclusion

- Situation
  The French speech community is quite active. All types of actors are present with many universities, academic institutions, and both public and private research laboratories are involved in national and EEC projects, in France but also in Switzerland and Belgium. The recent creation of FRANCIL demonstrates this activity. Major telephone, telecom companies operate in the field and the many companies demonstrate the interest in speech technology.

- Needs
  — telephone databases (SPEECHDAT)
  — application-ready corpora (over-the-phone, spontaneous, different conditions...)
  — advanced corpora (dialogue, dictation) (Aupelf)

- Dissemination
  There are some attempt at making databases available on the market as commercial products, by the producer itself or through a licensed company. Most of the resources produced by academics are available free to other academics but distribution to industrials is still a problem and, when possible, is carried out on a case-by-case basis. As most academics have no way to handle properly commercial issues, the French organisation GRD-PRC is very interested in setting up a national repository that would take care of the dissemination of both written and spoken language resources. There is a strong desire for a European center for resource distribution, that would take care of issues such as IPR and licensing agreements.

I.3.5 GERMAN (Germany)

I.3.5.1 Existing databases and their actors

- Experimental research
  — Academics: basic (PHONDAT1), German Pronunciation Dictionary.
  — EEC: EUROM1, EUROM0, POLYGLOT

- Application-oriented
  — Academics: PHONDAT2, ERBA (train)
  — Industry: SUNSTAR (EEC)

- Telephone: Siemens Telephone Database

I.3.5.2 Ongoing projects & new initiatives

Various national, industrial, and EEC initiatives and cooperations, as:

- Experimental research
  — VERBMOBIL (basic, advanced) by industry and universities
  — VERBMOBIL-PHONDAT (advanced) by academics
  — TEDspeeches, TEDlaryngo (advanced), TEDphone by Eurococosda
  — Articulation of German Vowels (advanced)
  — Publicly Spoken German, very large database of german utterances
- Application-oriented
  — ONOMASTICA (EEC)
  — Siemens "1000 read sentences" by company/university
- Telephone
  — TEDphone (Eurococosda)
  — SPEECHDAT (EEC) Polyphone type German Corpus by Siemens
  — Stemmer Telephone Database

I.3.5.3 Conclusion

- Situation: German is very active. All actors are present. Many universities, many large companies, the national telecom, are involved in the domain.
- Needs: Basic research material is more or less available, but making EUROMS available could be important as part of a really multilingual corpus. Application-oriented corpora are requested by industrials.
- Dissemination: Spoken language resources in German are either free for any use, available for research only, available for project partners only (EEC projects), or of unknown availability. So it is clear that in Germany, commercialisation of linguistic resources is an important issue.

I.3.6 GREEK (Greece)

I.3.6.1 Existing databases and their actors

- Experimental research: Basic material
  — EEC: EUROM1, POLYGLOT
  — Academics: isolated speech 2000 words (alphadigits).

I.3.6.2 Ongoing projects & new initiatives

EEC: ONOMASTICA

I.3.6.3 Conclusion

Considering that Greece is a bit late for what concerns telephone equipment, speech technology is not the today crucial concern in Greece. So the investment of telephone companies is low, and there is no promising national initiative. The main basic research corpora have been produced due to EEC contracts, and it is therefore very important to achieve the work so far done (e.g. Basic material to be made available (EUROM1)).

I.3.7 ITALIAN (Italy)

I.3.7.1 Existing databases and their actors

- Experimental research
  — EEC: EUROM0, EUROM1
  — National: IRST Acoustic Phonetic corpus of 3000 sentences, AIDA (CVCV words, digits).
- Application-oriented: ARS1000 database, SUNSTAR database (EEC)
- Telephone: COLLECT: telephone customers utterances

I.3.7.2 Ongoing projects & new initiatives

Italy is member of the COCOSDA, EUROCOCOSDA initiatives.

- Application-oriented: ONOMASTICA (EEC)
- Telephone
  — POLYPHONE: CSELT is collecting a large telephone speech corpus including a complete POLYPHONE data set.
  — SPEECHDAT

I.3.7.3 Conclusion

- Situation: All actors present. Italian Speech community is active. Academics, telecom operator, companies do cooperate.
- Needs: Mainly telephone (Speechdat) and application-oriented corpora. Advanced material for basic research is also needed.
- Dissemination: Basic research material(EUROM1) was produced on CDROM and is being marketed by CSELT; the availability of other corpora is not-well defined, but the general tendancy is to produce the corpora on CDs with the possibility of being made available on a case-by-case exchange basis.

I.3.8 NORWEGIAN (Norway)

I.3.8.1 Existing databases and their actors

- Basic material: has been produced due to national (NTH) initiative and EEC initiative (EUROMs)
- Telephone corpora: are produced in cooperation by national telecom often in cooperation with Denmark.

I.3.8.2 Ongoing projects & new initiatives

- Experimental research
- Application-oriented
- Telephone
  — EEC

— ONOMASTICA- Norwegian Telecom, SINTEF
— COST project 232 - NTH multi-accented English speech corpus collected over dialed-up international telephone lines
— COST project 249 -NTH, Norwegian Telecom

— National: Continuous speech recognition over the telephone line.

I.3.8.3 Conclusion

Research institutes, National Telecom, telephone operator are present. EUROM1 to be made available as basic material. General lack of basic research corpora.

I.3.9 PORTUGUESE (Portugal)

I.3.9.1 Existing databases and their actors

- Fundamental research: EUROM1 (CEC)
- Application-oriented: SUNSTAR (Portuguese accent for English words)

I.3.9.2 Ongoing projects & new initiatives

Portugal is represented in some European current initiatives such as ONOMASTICA, RELATOR, ELSNET, SPEECHDAT and COST.

I.3.9.3 Conclusion

- Situation
  The Portuguese speech community is relatively small, and its main actors have been so far Universities and academic research institutions. There is very limited funding for basic research in this area. Telecom companies have only very recently shown some interest in speech technology. There are extremely few databases for European Portuguese, although EUROM1 will soon be made available as basic material.
- Needs
  — fundamental research (namelly in terms of corpora with phonetic and prosodic labelling)
  — telephone databases (SPEECHDAT)
  — application-specific corpora
  — advanced corpora (dictation, dialogue)

I.3.10 SPANISH (Spain)

I.3.10.1 Existing databases and their actors

- Experimental research
  — EEC: EUROM1
  — National

  Corpus oral de referencia del espanol contemporaneo (Advanced)
  ALBAYZIN (Exp. and Appl.)
  Automatic Speech Recognition PA85/86
  Construccion de Sistemas de Reconocimiento de Habla Mediante Tecnicas De Aprendizaje Automatico TIC-448/89

— Industry TANGORA (IBM, also for other European languages (French, Italian??), but not available)

- Application-oriented ROARS, POLYGLOT, SUNSTAR (EEC)
- Telephone

I.3.10.2 Ongoing projects & new initiatives

- Experimental research — National: ALBAYZIN project, child language project (advanced)
- Application-oriented ONOMASTICA (EEC)
- Telephone SPEECHDAT (EEC): polyphone corpus to be collected by Vocalis.

I.3.10.3 Conclusion

All actors present. Many universities, research institutes public or privates, national telecom operator involved in the domain. Growing national effort, accompanied by a survey of textual and spoken corpora in Spanish. No corpus is currently available on CDROM. Availability and cost information not well-defined.

I.3.11 SWEDISH (Sweden)

I.3.11.1 Existing databases and their actors

- Experimental research

  EEC initiatives produced basic material: EUROM0, EUROM1
  National initiative produced basic material:Swedish sentence material.

- Application-oriented National initiative produced CAR database.

I.3.11.2 Ongoing projects & new initiatives

- Experimental research

  WAXHOLM dialogue project (national)
  ONOMASTICA (EEC)
  A prosodic Swedish database project is under discussion.

I.3.11.3 Conclusion

Some research institutes are involved in speech technology. Effort for basic Swedish resources was supported the Swedish government to participate in EEC projects and by other national efforts. There is now national effort towards application-oriented and advanced material. But the basic material (EUROM1) is still to be made available.
Status of telephone-based corpora unknown.

I.4 General conclusions

- Among existing and prospective resources, usability and reusability of speech databases produced through EEC projects should be enhanced. These resources are most of the time still unavailable several years after the end of the projects, because the time-intensive process leading from the raw data to marketable data was not initially included (or funded) in the projects (the marketable product is not a deliverable of the project). The distribution of these databases can be stimulated by the existence of an European Center for the Distribution of Language Resources who can provide funding for reorganisation and documentation of the already existing data and its production on CDROM.

- Existing databases whose diffusion is confidential should be highlighted. These databases are either not well-known because they remain in the scope of the project for which they have been designed, and/or have been kept confidential by their designers who don't want to disseminate them. It is obvious that for many industrials, linguistic resources are not yet considered as pre-competitive resources, but as strategic ones. Any future European Center of LR production and distribution will have to manage the gap between the views and methods of industrials and academics: long-term widely available resources for academics vs. short-term strategic resources for industrials.

- There is a clear lack of speech databases suitable for multilingual evaluation. At the curent time, only EUROM has a Europe-wide dimension. Large, multi-lingual corpora are needed, and added value (full annotation and labelling) must be included in the design of new databases. It should be noted that the most widely used multilingual corpora are currently being recorded by OGI in the U.S. This corpus is being used for development and testing of automatic language identification techniques, which have a wide variety of practical applications.

- Pronouncing dictionaries and spoken language lexicons are very limited and should be developed for speech output systems assessment. Large phonetic lexicons are needed for use in speech recognition. One particular source that should be exploited is the outcome of the Onomastic project, and in particular the legal details for distribution of the lexica need to be worked out.

- A corpus of at least 100 hours of high quality speech should exist in each language providing coverage of many words and phonetic contexts, with data from a reasonably large number of speakers (200–500). Most languages are far from that (for example at 20 Khz, 100 hours of speech correspond at 15 Gbytes, or about 25 to 30 CDROMs). In addition, this basic corpus will need to be supplemented by application-oriented corpora (such as for dictation (medical, legal, insurance domains), office systems, speaker-verification/identification, topic spotting, information retreival, etc.). Newspaper-based corpora are a potential multilingual source that should be encouraged, as these basic corpora provide at the same time speech and language modeling data, that can be used for dictation as well as other applications such as topic spotting. Multi-sensor and multimodal corpora are also needed for more basic research.

- For short-term commercial use there is a need for large telephone-based corpora, with speech from many (several thousands of) speakers covering a wide range of dialect, age, and socio-economic backgrounds. These corpora should be extensive enough to permit the design of speaker-independent, vocabulary-independent speech recognisers that can serve as the source for a variety of applications.

Our conclusions are in agreement with those in the recent Eagles report on Spoken Language Systems. It is clear that the need for adequate resources is a prime concern for many actors in the field of language technology. The following excerpts come from a draft EAGLES report on Spoken

Language Systems (For EAGLES Restricted use). We highlight three areas of needs: speech corpora, lexicons and assessment of speech output.

- Speech corpora size (in terms of speakers), p. 10–11

  - few speakers (< 5 speakers):for development of speech synthesis systems (dictionaries of phonetic elements). Advanced research. Multi-channel recordings (Electroglottogram, subglottal pressure...)
  - medium (5–50 speakers): experimental research. Number of speakers and repetitions large enough for statistical processing / or for broad coverage of phenomena.
  - large (> 50 speakers): Train and test of speaker-independent recognition systems.

- Spoken language lexicons (p.67)
  "Large-scale spoken language lexical resources, from reference sources of standard, stylistic and regional pronunciations through vocabularies which are characteristic of spoken language, are required for current research and development with both statistical and knowledge-based technologies.()
  These spoken language lexical resources in the form of actual lexical databases and tools for constructing them, are sadly lacking."

- Assessment of speech output systems (p.102)
  "A short-term recommendation is to develop multi-lingual machine readable pronouncing dictionaries at the single word level which list permissible variations ()."

I.5 Production costs

Speech database is the only type of speech resource we may be able to estimate the costs of production, based on concrete production cases. That is rough speech database. The cost of any Added Value (annotation, transcription, phonetic or prosodic labeling ...) will depend on the manual/semi-automatic/automatic way to proceed and on the corresponding tools available or to be developed. And this refer to Tools in general, whose development costs are those of the software industry (it is essentially a matter of specifications and corrresponding man-months). The production of spoken language lexicons and pronunciation dictionnaries should be comparable to speech databases in term of production costs. An estimate of this cost can hopefully be provided by the ongoing Onamastica project, and perhaps by SPEX from their experience with CLEX.

A current estimation of the minimum cost for database production is 1 Ecu per utterance. But this is the lowest case which may be applicable to large telephone corpora, for which there is no need to move people (even not to pay them), and for which the initial equipment investment is relatively minor. However, these corpora are by nature of telephone speech quality and may not be useful for basic research and technology development. But many other factors can come to increase the cost. At the other extreme, multi-channel recordings in a specific and controled environment with representative selected speakers from all over a country, may be requested for advanced research material. The cost in this case, estimated from completed and available databases, may reach 10 Ecus per utterance (e.g. ten times more than for the telephone corpus.) In fact, more specific corpora, including a variety of articulatory sensors can be substantially more expensive.

So we can argue that the cost scale for the type of widely applicable, general purpose speech data is from *1 to 10 Ecus per utterance*, depending on various factors as:

- Type of utterance: word, sentence, passages, dialog ... the cost is higher and higher in term of recording time, error recovery, storage capacity.

- Speech quality: sampling frequency and quantification (number of bits per sample) are relevant factors. ¿From telephonometry to high quality standard audio files, the same utterance will require 5 times storage capacity (e.g. more CDROMs for the same database). The use of lossless compression techniques (T. Robinson, CUED) can reduce the needed storage by a factor of 2. The Relator project has supported the adaptation of the UNIX compression algorithm to DOS for use on PCs.

- Recording protocol: Multi-channel recordings require n times the storage space. For example, if both speech signal and Lx signal are recorded together, for one utterance the volume is doubled. The more the protocol is complex (multi-sensors, specific equipment) the more the production process is time-expensive and the pressing process is space- and hence time-consuming (number of CDs, effort to prepare them..). Due to the limited bandwidth of certain channels, the actual storage requirements may be less, as lower sampling rates can be used.

- Speakers selection: from people calling a toll-free number from their own phone, to speakers that must be present in a specific recording site (including transport, hotel, food expanses and eventual payments), there is a wide range of situations and corresponding costs.