SHORT-TIME FEATURE VECTOR BASED PHONEMIC SPEECH RECOGNITION WITH THE AID OF LOCAL CONTEXT

Kari Torkkola

Helsinki University of Technology
Faculty of Information Technology
Department of Computer Science
Laboratory of Computer and Information Science

Thesis for the degree of Doctor of Technology to be presented with due permission for public examination and criticism in the Auditorium F1 of the Helsinki University of Technology on the 3rd of May 1991, at 12 o'clock noon.

Preface

The work presented in this thesis has been carried out at the Laboratory of Computer and Information Science in the Faculty of Information Technology of the Helsinki University of Technology.

I am greatly indebted to Professor Teuvo Kohonen for guidance and valuable contributions to my studies, as well as providing excellent facilities and financial support for this research. I wish to thank all of the personnel of the laboratory for creating a comfortable and inspiring atmosphere. I am especially grateful to my collaborators Jari Kangas, Sami Kaski, Mikko Kokkonen, Mikko Kurimo, Jorma Laaksonen, and Pekka Utela in the speech recognition group for useful discussions and cooperation at different phases of my work.

The financial support of the Jenny and Antti Wihuri Foundation, and the Foundation for the Promotion of Technology is gratefully acknowledged.

Espoo April 10, 1991

Kari Torkkola

Contents

List of Publications

This thesis consists of the following publications:

1. K.Torkkola, "Automatic Alignment of Speech with Phonetic Transcriptions in Real Time," Proceedings of the IEEE 1988 International Conference on Acoustics, Speech, and Signal Processing (ICASSP88), New York, USA, April 11-14, 1988, pp. 611-614.

2. K.Torkkola, T.Kohonen , "Correction of Quasiphoneme Strings by the Dynamically Expanding Context", Proceedings of the 9th International Conference on Pattern Recognition (9th ICPR), Rome, Italy, November 14-17, 1988, pp. 487-489.

3. K.Torkkola, K.Raivio, "Comparison of Symbolic and Connectionist Approaches to Eliminate Coarticulation Effects in Phonemic Speech Recognition", Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 89), Paris, France, September 26-28, 1989, Vol. 1, pp. 9-12.

4. M.Kokkonen, K.Torkkola, "Using Self-Organizing Maps and Multi-Layered Feed-Forward Nets to Obtain Phonemic Transcriptions of Spoken Utterances", Speech Communication, Vol. 9, No 5-6, December 1990, pp. 541-549.

5. K.Torkkola, "A Combination of Neural Network and Low Level AI-Techniques to Transcribe Speech into Phonemes", Proceedings of COGNITIVA-90, Madrid, Spain, November 20-23, 1990, vol. 2, pp. 637-644.

6. K.Torkkola, M.Kokkonen, "A Comparison of Two Methods to Transcribe Speech into Phonemes: A Rule-Based Method vs. Back-Propagation", Proceedings of 1990 International Conference on Spoken Language Processing (ICSLP90), November 18-22, 1990, Kobe, Japan, Vol. 1, pp. 673-676.

7. K.Torkkola, M.Kokkonen, "Using the Topology-Preserving Properties of SOFMs in Speech Recognition", to be published in Proceedings of the IEEE 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP91), May 14-17, 1991, Toronto, Canada.

8. K.Torkkola, J.Kangas, P.Utela, S.Kaski, M.Kokkonen, M.Kurimo and T.Kohonen, "Status Report of the Finnish Phonetic Typewriter Project", to be published in Proceedings of the International Conference on Artificial Neural Networks (ICANN-91), June 24-28, 1991, Espoo, Finland.

List of Abbreviations

IWR	Isolated word recognition
CWR	Connected word recognition
CSR	Continuous speech recognition
FFT	Fast Fourier transform
DTW	Dynamic time warping
HMM	Hidden Markov model
VQ	Vector quantization
ANN	Artificial neural network
SOM	Self-organizing map
LVQ	Learning vector quantization
MFN	Multi-layer feed-forward network
TDNN	Time-delay neural network
DEC	Dynamically expanding context
DFC	Dynamically focusing context

1 Introduction

Speech recognition both by man and machine can be presented as the following series of actions. The speaker has in mind an abstract message, that he or she first presents as a stream of some abstract units of the language. The smallest contrastive units in a language are generally agreed to be phonemes [78]. This stream is converted into a continuous acoustic signal by the speaker's articulators. The listener or an automatic speech recognizer is now faced with the task of decoding the acoustic signal to find out the original message.

The acoustic signal alone does not provide enough information to interpret any message. Humans take advantage of all their knowledge in speech perception. In fact, we constantly predict what is expected to be said, and what can be said in the context of the current interpreted message or discussion. The acoustic information is then continually matched against these expectations. Some specific sources of this knowledge are, for example, knowledge about phonemes appearing in the language, legal phoneme combinations, legal words, the syntax of the language, and the semantics of the message [119], [112], [80]. To attain human performance in speech recognition by machines all this should be taken into account. Experiments have been made with humans where one or more of these knowledge sources are missing [85], [79]. For example, humans have been compared to machines in the task of context-independent vowel recognition, where human performance fell short of that of machines [74]. The results indicate that knowledge outside the acoustic signal contributes a lot to the human recognition process. This suggests that in automatic speech recognition it is almost a necessity to exploit higher level linguistic knowledge. However, a prerequisite to successful application of this knowledge is that the acoustic signal is also taken advantage of in the best possible way. Neither linguistic knowledge, nor the acoustic signal alone can be used successfully without the other. There are also other means of communication that humans utilize in addition to the acoustic signal. Facial expressions and gestures carry a lot of significant information, and deaf people can even base their speech recognition solely on this. This visual data can even be utilized in automatic speech recognition [115].

This work concentrates on a subtask of speech recognition, namely phonemic speech recognition, that is, transcribing spoken utterances into phoneme sequences. In certain languages, such as Finnish, recognition of phonemes from the acoustic signal is a good basis for further processing (see Section 3). However, being abstractions, phonemes do not directly correspond to clearly distinct acoustic segments. The phonemes have varying lengths, they may consist of several partly overlapping acoustic segments, and their acoustic appearances vary according to their context, and also across speakers. Even the same speaker is not able to produce exactly the same acoustic signal for the same utterance. One of the most difficult problems in automatic speech recognition is posed by this acoustic variability. The task of transcribing speech into phonemes is thus anything but trivial.

The starting point of the work described in this thesis is a speech recognition system that is based on classifying short-time feature vectors into phonemic classes [43], [49], [50]. To improve the classification accuracy of individual short-time feature vectors, information about classifying their local context is extracted. Two different non-parametric methods are then used to re-classify this context information, one based on symbolic processing, and the other on artificial neural networks. This scheme can be viewed as a combination of linguistic and acoustic aspects. Linguistic information is considered by using phoneme classes as the basis for short-time classification. Acoustic information is incorporated by extracting the linguistic classification information from certain windows in the local context.

The outline of this thesis is as follows. Section 2 gives a brief overview of the properties of speech signals and different approaches to speech recognition. Section 3 presents the motivation to study phonemic speech recognition as well as a review of different approaches to it. Section 4 first describes the speech recognition system that is the starting point of this research. The methods developed in this thesis to take context into account in classification are briefly presented. Finally, Section 5 summarizes the publications, and Section 6 states the author's contribution to the work presented in the publications.

The most important goal in this work has been to maximize the accuracy of recognized phonemic transcriptions. The viewpoint taken is thus mainly engineering.

2 Automatic Speech Recognition

2.1 Speech Production and Acoustic-Phonetics

Human speech production can be viewed as a process where a sound source excites the vocal tract that acts as a filter. This sound source can be one of two types. In *voiced speech* air is released through the vocal chords that vibrate at the so called fundamental frequency of speech, usually denoted F0. The resulting periodic source then resonates in the vocal tract. In *unvoiced speech* this source is noise that is generated by forcing air through a narrow constriction. Sounds produced in this way are called fricatives.

As the sound source is filtered by the vocal tract, the spectra of voiced speech consist of harmonics of the F0. The peaks of the spectral envelope, corresponding to resonances in the vocal tract, are called formants. Their positions in frequency scale as well as their relative amplitudes are important in human speech perception, especially for the first three formants. The characteristics of speech signals often remain relatively fixed, or quasi-stationary over a short period of time (tens of milliseconds) but vary over intervals longer than distinct sounds (50-100 milliseconds). For more details on speech production and properties of speech see [18], [19], [78].

Phonemes can be classified in terms of voicing, manner of articulation, and place of articulation [26], [78]. The manner of articulation defines how the vocal tract restricts the airflow: 1) Occluding the airflow and then releasing it creates a stop consonant, 2) Lowering the velum causes nasal sounds, and 3) Other phonemes have vocal tract restrictions of varying degree. Place of articulation refers to the most narrow constriction, which determines the most prominent acoustic features of a phoneme.

As opposed to, for example, image analysis, speech recognition has the advantage that some fairly well-defined levels of hierachy exist in language and consequently in speech. Starting from the top these can be enumerated as sentence, word, morpheme, syllable, phoneme, phone, and acoustic segment levels [26], [112], [78]. The borderline between linguistics and acoustics can be drawn between phoneme and phone levels; phoneme representing the lowest level of hierarchy in the language. A *phone* is defined as the acoustic manifestation of a phoneme [78]. Different manifestations, that is, different phones of a phoneme, are called *allophones* of the phoneme. A phone consists of one or more acoustic segments that may not be clearly separable from each other within a phone. Transitions from one phone to another are usually smooth and continuous due to inertia in human articulators. This continuity is so strong that the position of the articulators during the utterance of one phoneme affects their position during adjacent phonemes and even behind them. This phenomenon is called *coarticulation*. Often it is not possible to segment the speech signal into discrete phones with simple one-to-one correspondence. Speech thus exhibits a certain dualism — it is a sequence of discrete phonemes on the one hand and a continuous acoustic signal on the other hand.

2.2 Making Speech Recognition Approachable

Human performance in recognizing arbitrary speech is much ahead of the current state-of-the-art of automatic speech recognition technology [70], and it may not be equalled in a long time, if ever. The most striking difficulties in speech recognition can be summarized as follows [23]:

1. Acoustic features vary between speakers, and even for the same speaker, because the manner of speaking can never be kept constant. Different phonemes uttered by different speakers may have similar acoustic features.
2. Coarticulation causes variability in linguistic units (phonemes), which makes the acoustic features of phonemes dependent on their context. This also makes words spoken in a continuous manner different from those spoken in isolation.
3. Due to the continuous nature of the speech signal and due to coarticulation, it is difficult to segment the speech signal according to any linguistic units, such as phonemes, syllables, or words.
4. It is impossible to model and use all of the knowledge that humans take advantage of in speech perception. It is also difficult to model even a small fraction of it, for example, the syntax of a natural language.

However, by restricting the task very useful results have been obtained. There are three main dimensions along which the difficulty of the task may be reduced [85], [78].

1. Input variability can be limited by limiting the number of speakers. The system may be *speaker-dependent*, that is, it must be trained separately for each speaker. *Multi-speaker* systems are trained once for a group of speakers and thereafter require no further adaptation for those speakers. Finally, the most difficult task is *speaker-independent* speech recognition. It is impossible to achieve accurate and reliable speaker-independent phoneme classification based only on short-time feature vectors. Attention must also be paid to speaker's characteristics and/or phonetic environment [65]. However, short-time feature vectors can be classified speaker-dependently into phoneme classes as can be seen in [43] and in this thesis. On the other hand, a speaker-independent system may be more robust — being able to tolerate large deviations in speech. A system tuned for a specific speaker may become too sensitive even for small variations.
2. The style of speech can be restricted. In *isolated word recognition* (IWR) the speaker has to pause between utterances. The burden of detecting word limits is thus relieved from the recognizer. In *connected word recognition* (CWR) no pauses are required but the speaker has to pronounce and to stress all of the words clearly. *Continuous speech recognition* (CSR) is the least restricted style; usually this denotes normal conversational speech.

3. The number and confusability of the words accepted by the recognizer can be limited. The recognition task can be eased furthermore by restricting the syntax of the acceptable language so that there cannot be many different word candidates at any point in any sentence of the language. Incorporating semantic analysis can further reduce alternatives.

By comparing this list to the list of difficulties in speech recognition, we notice that there is at least one thing that remains even when all of the aforementioned restrictions are used. This is the coarticulation within isolated utterances. The work in this thesis is intended to compensate for those coarticulatory effects.

2.3 Different Approaches to Speech Recognition

2.3.1 Main Paradigms of Speech Recognition

The two main paradigms of automatic speech recognition are the knowledge-based approach and the pattern recognition approach [61], [78]. In the knowledge-based or expert-system approach, explicit rules are searched to explain the relationships between the speech signal and the corresponding linguistic representation. On the other hand, the pattern recognition approach, or statistical approach, views speech as a signal that can be studied by statistical analysis. These two approaches can also complement each other.

Since speech is comprised of a hierarchy of linguistic units, each of the paradigms can be used in two directions. The hierarchy can be followed bottom-up or top-down. The bottom-up approach starts from the acoustic signal converting it to parameters, features, phones, words, and sentences. The top-down direction tries analysis by synthesis, generating possible utterances and comparing them to the input at some or all levels. Many existing systems employ both approaches in a hybrid way: data reduction by bottom-up approach and hypothesis generation by top-down approach [95]. Of course, properties of speech are exploited as a part of a general framework in all approaches. Features known to be relevant for speech perception are extracted from the signal. The quasi-stationary nature of the speech signal is taken into account in models of the speech sub-units.

2.3.2 Selection of Sub-Units

One more main taxonomy can be presented: whether the recognizer treats each utterance as a complete unit, or first analyzes the utterance in terms of sub-units. In the former and simpler approach that is generally used in isolated word recognition (IWR), the recognizer has a limited collection of reference templates or models to which the input is compared. The closest template is then decided to be the one uttered. In the latter approach, a lexicon of some sort of sub-units of speech or language is defined. Entries of the lexicon are first recognized in a bottom-up style. From a stream or a lattice of these sub-units, higher level linguistic units are constructed according to some imposed restrictions. To this end we must know some syntactic rules to arrange the units into words and/or sentences. Semantic relations between the units must also be known if the meaning is analyzed. The most appropriate sub-unit, or whether to use sub-units at all depends on the type of recognition to be performed and the task.

In IWR the problem of coarticulation within an utterance can be avoided by treating each utterance as a whole. The reference templates can be stored as sequences of some feature vectors. The main problem in comparing them to the corresponding representation of the input lies in temporal alignment. This can be accomplished effectively by dynamic time warping (DTW) that combines alignment and vector-to-vector distance computation through dynamic programming [92]. The other main approach to isolated utterance recognition is based on constructing statistical models to describe the time evolution of the speech signal as a sequence of feature vectors. These are most often based on representing the utterance as a first-order Markov process. Each utterance is represented as a state transition network. The utterance model that has the maximum likelihood of generating the observed sequence of feature vectors is selected. Efficient algorithms exist to train and utilize these so-called hidden Markov models (HMM) [83]. The main limits to IWR are posed by the size of the vocabulary. This is due to the effort required by each new speaker to produce the templates for each word of the vocabulary. In addition, the words have to be carefully selected not to be easily confusable with each other. Sub-unit based recognition is then the only possibility for more ambitious tasks.

In CWR the sub-units may also be words. Several variations of the DTW-scheme have been developed for CWR. These include two-level DTW [91] and level-building DTW [75]. The disadvantage is that between-word coarticulation is difficult to handle by using word templates. This can be overcome by employing sub-units below words with the network approach that is described below in connection to CSR.

Examples of sub-units below word level might be syllables, diphones, or phonemes [56]. The greatest advantage in using phonemes as the basic unit is the small number of them. For example, in the original Finnish language (without loan words) there are only 21 phonemes, whereas in English the number is approximately 40 [78]. The greatest disadvantage is that the phoneme is an idealized abstraction, and thus its acoustic appearance shows extremely great variance. The approaches to phoneme recognition and difficulties therein are described in Section 3.

2.3.3 Taking Linguistic Knowledge into Consideration

There are basically three different approaches to take high-level constraints (syntax, semantics) into account in CSR [23]:

1. Different levels of knowledge are distributed into *hierarchical* modules. Lower level modules produce candidates, and upper level modules produce evaluations or predictions limiting the number of possible candidates. Examples of this approach are [116] and [28].
2. The *blackboard* model differs from the previous approach in that a database common to all modules is used. Each module can gain access unsynchronously to the blackboard to verify hypotheses present there or to insert new hypotheses to be verified by other knowledge modules [16], [2].
3. In the *network* approach, knowledge about syntactic and semantic constraints on allowable word sequences may be coded as a network, whose states are the words of the vocabulary. Transitions between states are allowed only if the resulting sequence of words produces a legal sentence of the language. The states in the network may also represent phones, acoustic segments, or feature vectors. The most common networks in the CSR-literature are HMMs, see for example [4] and [57].

A thorough overview of syntactic methods used in CSR can be found in [107]. Good presentations about the whole field of speech recognition are given in [1], [23], [85], and [78].

3 Phonemic Speech Recognition

3.1 Is Separate Phoneme Recognition Useful?

Word-based approaches to speech recognition are difficult in many languages, because the words may have numerous inflected forms. To obtain sufficient coverage for dictation tasks in, for example, office environments, all these forms should be included in the lexicon as distinct entries. This would expand its size too much. Therefore, in these languages the first level of a speech recognizer should distinguish units smaller than words, such as syllables, diphones, or phonemes.

The language used in this work is Finnish. Some important characteristics of this language are: 1) it contains 21 phonemes in its original form, 2) the written and spoken forms have almost one-to-one correspondence, that is, the written text symbols (graphemes) and phonemes correspond to each other almost identically, 3) the words are highly inflected, for example, a single verb can have over 1000 different inflected forms in which the baseform of the word can also change. Compared to the English language, the number of phonemes in Finnish is quite small. Furthermore, a large proportion of the phonemes are quasi-stationary (they have a stationary part), and thus are relatively easy to recognize [48]. The choice of phonemes as the first recognition level is therefore natural. The methods described in this work are not limited to Finnish. Their applicability to Japanese has also been demonstrated [50].

Trying first to recognize phonemes seems to be well suited to the speaker-dependent case as can be seen in the results of this work. The speaker-independent case is more problematic due to larger variability. It has even been argued that speaker-independent phonemic speech recognition is impossible because no phonemic invariance exists in this case. The message is structured into several interpretation levels that complement each other. The signal does not always carry information necessary to decode low-level elements such as phonemes [65]. Of course, this applies also to the speaker-dependent case, but the variability is much lower.

Clearly, separate phoneme recognition cannot be a basis for a complete system unless its accuracy were very close to one hundred per cent. Where could we then use a separate phonemic recognition stage? In IWR, HMMs and DTW have proven to be extremely powerful tools. An intermediate phonemic recognition stage seems to be unnecessary, possibly introducing extra errors. However, phonemes can be used as the sub-unit of recognition in IWR and CSR even though no intermediate phonemic transcription is produced [56], [58]. It has been shown that at least in the speaker-dependent case, IWR can be based on a separate phonemic recognition stage when the accuracy of produced phonemic transcriptions is high enough [48]. This approach has also been in use in speaker-independent CSR [63], [77]. The advantage of an intermediate phonemic level is that the rest of the processing can be performed on a symbolic level, thus reducing the required bandwidth.

To answer the question in the heading: it depends on the case. When the number of phonemes in the language is small, and the variability can be reduced, for example, by speaker-dependent recognition, a separate phonemic recognition stage is clearly an attractive and feasible alternative even as an intermediate phase in IWR or CWR. Reliable phoneme recognition requires careful articulation and it is thus most feasible for dictation, not for conversational style speech. On the other hand, when the number of phonetic elements to be recognized is large, and/or speaker-independent recognition should be implemented, better results have been achieved by using no such fixed intermediate descriptions ([63] compared to [58]). An alternative to a fixed intermediate representation is to relay more information to later processing stages by means of a lattice or a network of phoneme alternatives [9], [117].

3.2 Pattern Recognition Methodology

Phoneme recognition can be approached by the traditional vector space pattern recognition methods for which a multitude of different approaches have been described in the literature [21], [105], [8], [96]. The three important points to consider before applying these approaches are:

1. What are the classes? Should we take phonemes or phones as such, should some of the phonemes be combined, or should some phoneme classes be split into subclasses? This depends again on the task. For example, if there are phoneme classes whose combination does not confuse the words in the vocabulary they can be combined to simplify the classification problem. Since many phonemes are composed of very different acoustic segments it is not sensible to try to classify these segments by the same phoneme label, but rather to create new classes according to different kind of segments. Plosives are a clear example of this – having a closure part of silence followed by a burst.
2. What are the features to extract? Some alternatives are treated in depth in [120], [78], [23]. The features should, of course, be chosen to separate the selected classes in the best possible way. Features should be extracted from a period in which the speech signal remains fairly constant. This period is usually considered to be 5 – 30 milliseconds. Feature selection can be approached from two points of view:
   (a) Speech acoustics. The starting point of this approach is that the most important features lie in the spectral envelope of the vocal tract response. This can be attained through short-time spectral analysis [84], or by the linear prediction analysis [69]. The spectral analysis can be computed using the fast Fourier transform or by a filterbank. Cepstral analysis (that can be interpreted as computing the spectrum of the logarithmic power spectrum), has been quite popular lately, since it allows the spectral envelope to be conveniently separated from the exitation signal (the fundamental frequency F0) [78], [23]. Cepstral coefficients also provide a very compact representation of the spectral envelope [14].
   (b) Speech perception. Feature selection is approached psychoacoustically. This approach is based on the assumption that speech communication has been adapted to (or even optimized for) the properties of the human auditory system. Systems have been built that reflect human response to acoustic input [94], [31], [35]. These features seem to exhibit greater tolerance to variations of input speech. Some details, such as the human frequency resolution (mel-scale), have also been adopted in conventional spectral or cepstral analysis [14].

Feature selection can be further divided according to whether acoustic or phonetic features are used. All of the aforementioned methods produce acoustic features. Some examples of phonetic features are the fundamental frequency, the degree of voicing, and formant positions and amplitudes.

Many phonemes, especially in conversational style speech, do not have stationary parts. That is, they are a dynamic phenomenon. This temporal change can be included in features by incorporating some sort of estimates of time derivatives of the features [22].

3. What kind of a classifier should be chosen? A multitude of classical approaches can be found from the previously mentioned pattern recognition references. In general, it can be said that non-parametric classifiers perform best, since the classes seldom follow any nice parametric distributions, such as the Gaussian distribution.

   Recently, there has been a resurgence of interest in artificial neural network (ANN) research [42], [32]. ANNs consist of simple connected elements inspired by biological neurons. The decision-making is distributed in a collective and cooperative way. Many ANN architectures can be used for pattern classification [81]. Two of those most commonly used in speech recognition are multi-layer feed-forward networks (MFN) [90] and Learning Vector Quantization (LVQ) [41]. A review of these and other ANN alternatives in speech recognition is presented in [67].

   In general, non-parametric classifiers can be implemented quite "effortlessly" by using ANNs. We do not have to choose a classifier whose structure matches the problem at hand, but the structure can be implicitly learned through the adaptive training process of ANNs. Furthermore, some ANNs are computationally less expensive than many classical non-parametric classifiers such as k-nearest-neighbor or Parzen-window classifiers [96].

An academic approach to phonemic speech recognition can be taken by classifying previously segmented portions of a speech signal with the knowledge that these portions correspond to certain phonemes. This classification problem is reviewed in Section 3.3. From the engineering point of view these locations have to be found before classification. This combined problem of detection and classification is described in Section 3.4.

3.3 Short-Time Classification

The difficulty of short-time phonemic classification depends mostly on speaker-dependence/independence and the number of classes used. The language thus partly determines the complexity of the task. The simplest decisions could be made between voiced and non-voiced speech or between some broad categories, such as vowels, plosives, liquids, glides, nasals, and fricatives. In the work presented in this thesis, 21 phonemic classes of the original Finnish are used. In English, as many as 60 phone classes (for the 40 phonemes) can be defined [54], although the definitions of phoneme or phone classes may be subject to a certain controversy. For example, in short-time classification it is not reasonable to have two different classes according to different durations of a phoneme. Likewise in short-time classification it is not sensible to have diphthongs as separate classes, but rather as concatenations of two different phones. These topics may also depend on the language.

Classification of given phoneme positions can be divided roughly into two categories:

1. Instantaneous classification (for example, stationary region of a phoneme) using single or several concatenated short-time feature vectors.
2. Classification of a sequence of short-time feature vectors (for example, the whole phoneme duration).

3.3.1 Classifying Feature Vectors

The first category falls within the traditional vector-space pattern recognition framework for which many different approaches exist. They can be divided according to the three questions presented in Section 3.2. There seems to be no general agreement about the best possible features for speech. It seems that the task as well as the classifier contribute to feature selection. In some experiments auditory models have been the best ones [35], while in other experiments they have not [87]. However, some details duplicated from the human auditory system such as the mel-scale prevail in most systems.

The current trend in short-time classification seems to be that parametric, such as Gaussian classifiers [38], have been abandoned in favor of adaptive nonparametric classifiers, especially ANNs. One example of adaptive nonparametric classifiers is the Learning Subspace Method that has been applied succesfully in recognition of Finnish phonemes [48]. Evaluations of ANN-based classifiers in phoneme recognition can be found in [33] and [45] in which the ANNs have been found to be at least equal, and often slightly better than conventional classifiers. MFNs have been a particularly popular approach lately. A multitude of references can be found about applying them to short-time vector classification, e.g., [66], [110], [6] and [60]. References to LVQ can be found in [41], [45], [71], [114]. Some examples of other ANN approaches are [82] and [86].

Of course, better results can be expected if the whole phoneme duration, not just one short-time feature vector can be taken into account in classification. However, this leads to problems in the vector-space approach. It is not possible to concatenate all feature vectors from one phoneme area into a longer vector. The lengths of phonemes vary and the corresponding parts should always be time-aligned. To a certain extent, concatenation introduces improvement [71], but when the lengths of shortest phonemes are exceeded, the additional irrelevant data may eventually decrease performance [99].

3.3.2 Time-Series Classification

The classical approach to sequence classification in speech recognition is to use hidden Markov models [83]. Each phoneme or class to be recognized is represented by a separate model usually made up of from three to five states [4], [57]. The principle is exactly the same as used in isolated word recognition: the model having the maximum likelihood of generating the observed sequence is selected. Continuous density models that produce feature vectors directly [3], or discrete models producing somehow vector quantized (VQ) feature vectors can be used [57]. The drawback of the former approach is that a parametric distribution must be assumed for the feature vectors. On the other hand, the latter approach loses some information through vector quantization. In the latter case LVQ can be used to construct a vector quantizer that emphasizes phoneme class differences [36], [39]. Durations can also be incorporated into the HMMs to further improve the accuracy [89], [28]. Recently, many ANN classifiers have been introduced to the sequence classification problem. These can include some kind of internal delays to achieve shift-tolerance in time. Examples of this class of structures are time-delay

MFNs (or TDNNs) [109] and shift-tolerant LVQ [71]. A combination of TDNNs and LVQ has been presented in [10]. Another possibility is to have recurrent connections in the network to incorporate the time dimension [7], [17], [87].

Expert systems have also been designed for phoneme classification [113], [111]. These systems typically inspect phonetic features such as formants or so called distinctive features [26], [78]. These denote a small set of usually binary characteristics such as voiced, labial-alveolar-velar, nasal, fricative, etc.

No clear benchmarks exist to compare the different approaches, and comparison may not even be sensible or possible, because of the differences in problems for which the methods have been developed. However, one possibile benchmark might be a publicly available database, such as the DARPA TIMIT acoustic-phonetic database, that has been designed for speaker independent recognition [54], [24]. By comparing the published results on this database it seems that the best phonemic recognition results in classifying pre-segmented sections have been achieved using either recurrent MFNs or TDNNs [30], [60], [17]. In some other comparisons TDNNs and LVQ have been found to be about equal to best continuous density HMMs with duration modelling [76].

3.4 Transcribing Spoken Utterances into Phoneme Sequences

Phonemic recognition gets more difficult when the positions of phones have also to be found. In this case spoken utterances are converted directly into phonemic transcriptions without prior information about the message. On the other hand, information about the legal phoneme sequences in the (restricted) language can be easily taken into account. This information may actually simplify the task by limiting the number of valid phoneme class alternatives. The basic approaches to this problem can be enumerated as follows:

1. Phone spotting (segmentation) is separated from classification. The advantage is efficiency, since phones consist of several frames, from which feature vectors are computed. Labeling often requires examination of dynamic features over the whole phone duration. On the other hand, there exists no method to reliably accomplish the segmentation. Many ways that use acoustic features are described in the literature for the task, for example, stationarity of spectral features [64], temporal decomposition [5], [106], or similarity of consecutive frames [25]. Expert systems have also been built for segmentation [118], [11], [72], [20], [52]. These systems usually concentrate on phonetic features such as formants and their trajectories. An interesting approach combining ANN-based statistical and expert-system approaches is presented in [13] and [12]. Multiple resolution segmentation could also be conveyed to latter processing stages in order not to lose any important information [25], [117]. After segmentation any method described in Section 3.3 can be used in identification. An extensive review of different speech segmentation methods can be found in [108].
2. Each speech frame can be classified independently before segmentation producing a so called quasiphoneme sequence (if phoneme-like labels are used). Segmentation can then be based on the quasiphoneme sequence solely, or on some kind of distance measures to class prototypes used in the classification procedure. An example of the former is given in [27], where entropy of the quasiphoneme sequence is used as the basis of segmentation. The speech recognition system that has been the starting point in this thesis also uses this approach (see Section 4.1). Here the segmentation

(decoding) is based on context-independent durational rules [46]. An example of the distance-based approch to segmentation is given in [48].

3. Combined segmentation and labeling. In this approach the speech frames are usually not labeled at all. Phone-based HMMs can be used in parallel by allowing a state transition from one model to another. Viterbi decoding then finds the most likely state sequence through the phoneme models, and thus produces a phonemic transcription of the utterance. The HMMs can be of continuous density type [62], [68], or discrete if feature vectors are vector quantized [15].

   Phonemic vector quantization, which is equivalent to classifying frames into phonemic classes, also produces a quasiphoneme sequence. HMMs can again be used to decode them [99]. This combination could fall under the previous category as well. TDNNs have also been used in phone spotting, but with quite high insertion error rates [73].

The TIMIT acoustic-phonetic database mentioned in the previous section could act as a basis to compare speaker-independent methods of transcribing speech into phones. The best results reported in the literature are close to 70% correct phones using various methods [88], [59], [62].

4 Introduction to Phonemic Speech Recognition Methods Studied in This Thesis

4.1 The Starting Point

The starting point of the work in this thesis is the phonemic speech recognition system developed at Helsinki University of Technology. This system has been presented in detail in references [49], [50], and [43]. A short description will be given below.

The system is based on adaptive vector quantization methods of Kohonen [47], especially on the LVQ. The aim of vector quantization here is not to represent speech feature vectors with the smallest possible distortion, but with the best possible separation between phoneme classes. The LVQ accomplishes this by modifying the model vectors adaptively so that the borders between phoneme classes (midpoints between model vectors of different classes) approximate Bayes' decision surfaces. Good initial values for the LVQ are obtained by using the Self-Organizing Maps (SOMs) [47] to cluster the initial training data. These maps have also the property of preserving the topologigal ordering of the input data [44], but this property is not taken advantage of, except for display purposes [46].

The feature vectors used in the system are 15-component approximations of the short time power spectra of the speech signal. These are based on 256-point FFT, computed every centisecond from a window of two centiseconds. LVQ trained only with stationary speech samples is first used to vector quantize feature vectors derived from the entire speech signal. As a result, the speech signal will be vector quantized into a string of phoneme labels at centisecond intervals, which is called the *quasiphoneme* string. The quasiphoneme strings must then be transformed into phonemic transcriptions. This transformation, or phonemic decoding, has been done using simple context-free transformation rules that take into account only the durations of individual phonemes, not the effect of the contexts of the phonemes on them [46].

Some phoneme classes, for example, the unvoiced plosives /k/, /p/, and /t/ cannot be classified based on instantaneous features. In this case the plosives were treated as a single class. A separate "transient" LVQ using information from plosive bursts is then used to classify them more accurately, and to assign corresponding labels to the final phonemic transcription [37].

A major problem clearly visible in the phonemic transcriptions derived by applying durational rules is caused by coarticulation effects. For example, diphthong /au/ is often recognized as /aou/, /o/ being the realization of a smooth transition from one vowel to another. These kinds of effects are systematic to a large extent. To handle this problem, a rule-based correction method called the Dynamically Expanding Context (DEC) has been introduced by Kohonen [40]. This method takes context into account on phoneme level. However, the major part of this thesis work is to show that the coarticulation problem can be handled before phonemic decoding, on the quasiphoneme level.

We now briefly describe the original phoneme-level DEC-approach suggested by Kohonen (Section 4.2), and then proceed to show how the DEC-method can be modified to better suit to quasiphoneme data (Section 4.3). Alternative correction methods based on MFNs are also described both for phoneme and quasiphoneme levels (Sections 4.2 and 4.3.3).

4.2 Taking Coarticulation into Account after Phonemic Decoding

The problem for which the Dynamically Expanding Context was developed was to implement a mapping between symbolic data streams. This mapping should be learnt from examples. The central idea is to derive unique context-dependent symbol to symbol(s) transformation rules from two streams of symbols: the source stream, and the desired target stream. In the speech recognition application described here the source stream is the phonemic transcription produced by the LVQ-based system, and the target stream is the correct transcription, or the orthographic form of the utterance [40]. The two streams are first aligned symbol by symbol. Initial context-free transformation rules are formed by segmenting the aligned streams by symbols. A segment of the source stream acts as the condition part of a rule, and the corresponding segment of the target string constitutes the production part of the rule.

In case of a *conflict*, that is, when there are two rules among the training examples with equal condition parts but different productions, the condition parts have to be specialized. This is done by *expanding* the discrete symbolic context around the initial condition parts by just a sufficient amount to make the rules unique. The rules will then have condition parts of variable length. In this way, the generality of the rules will be optimum, and still, all the conflicts within the training cases are solved. The resulting rule set can be interpreted as a vast decision tree, where the decisions are made according to symbolic context.

Having constructed the rule set, it can be used to correct similar kind of errors that appeared in the training examples. The input stream is segmented like the training stream. Each segment is then searched for in the rule tree according to its context. The corrected result will be the concatenation of the found productions. In speech recognition experiments the DEC has been able to correct 50% - 75% of the errors remaining in phonemic transcriptions. For details of the algorithm, see [40] or [97].

Multi-layer feed-forward networks (MFNs) are widely used in this kind of mapping problems. The most famous example is probably the NETTALK system [93]. We compared the DEC to a MFN on the described task of correcting phonemic transcriptions [104]. Both methods used the same maximum amount of context, that is, four symbols around the center phoneme symbol to be corrected. The DEC corrected 50% of the errors while the MFN was able to correct only 37% of the errors. The explanation of this might be that the MFN has to take all of the context into account with equal importance, whereas the DEC is better able to concentrate on more relevant information. The order of relevance is explicitly defined by the expanding context. Alternatively, a MFN having more complex structure might be needed.

4.3 Taking Coarticulation into Account before Phonemic Decoding

Following the idea of DEC or NETTALK, a similar transformation can be applied to vector quantized (VQ) speech, too. In the example studied here, VQ produces a discrete phoneme symbol every centisecond. The duration of a phoneme extends over a period of many symbols, say four to forty symbols. The question that arises is what to map them to, or what could be the target sequence corresponding to a quasiphoneme sequence? The ideal quasiphoneme sequence was found to give the best results [98]. This kind of a sequence is defined as a sequence of phoneme labels which are constant within a phoneme area and change abruptly at phoneme borders. We could expect that if this mapping can be learnt from examples and it is able to transform the quasiphoneme sequences closer to ideal, better results could be expected from phonemic decoding. This was also verified by the experiments.

The difference to phoneme level mapping is that no insertion or deletion errors have to be corrected; we have only a symbol-to-symbol mapping. Additionally, in contrast to classifying long concatenated context vectors, here we take into account the *classification* of context, not the raw feature vectors as such.

A brief presentation of the ideas of the three different schemes evolved during this work follows. More complete descriptions can be found in the publications.

4.3.1 Applying DEC Directly to the Quasiphoneme Sequence

The DEC can, of course, be applied as such to a VQ'd speech signal. The problem with this data is that the period covered by one quasiphoneme symbol is too short to represent any relevant scale in speech, in contrast to one phoneme symbol in phoneme level DEC. Straightforward application of the DEC would thus produce a large number of too specific mapping rules and, consequently, a very large number of examples would be necessary to achieve any kind of coverage with the rules.

There are at least two solutions [100]. We can achieve generality by not requiring an exact match between the condition parts of a rule and the existing rulebase. In this case some kind of a distance score should be computed for all of the rules in the base and the closest one should be selected. This would, however, circumvent one of the merits of the DEC – its great speed based on tree search and exact rule matching. For this reason, another alternative was adopted.

Generality can also be achieved through the *representation* of condition parts of the rules. The context is still expanded by quasiphoneme symbols, but the durations of consecutive similar symbols are quantized quite strongly. It is equal to compressing or subsampling the condition parts made up from quasiphonemes. This leads to new condition parts, no longer using quasiphoneme symbols, but in which a single symbol covers a longer period of compressed quasiphoneme symbols. Exact match can then be required with these more general rules. The DEC-algorithm can be used as such [40] to construct the mapping rule set from examples.

4.3.2 Dynamically Focusing Context

The DEC has a fixed scale of representation (one symbol), but the width of context taken into account increases dynamically, while the condition parts of the rules are specialized. The algorithm implicitly defines an expanding hierarchy of importance for the context. The information contained in the context is also implicitly weighted following this importance. However, we are not confined to this expanding scheme only. Other hierarchies can be defined that better take into account the relevant scale of the data at hand. The scheme developed in this work uses context *focusing* to specialize the rules. Hence the name, Dynamically Focusing Context.

As an initial context, a larger section of data surrounding the area to be transformed is taken. A single discrete symbol is then computed to characterize this fine-grained data, and this symbol is taken as the condition part of the rule. In case of a conflict, the contexts have to be specialized. The initial section of data is then divided into smaller parts, from which their own characteristic symbols are computed and concatenated to construct the new condition part of the rule. This context focusing can be continued until no conflicting rules remain, or no more data division is possible. A maximum depth to stop the focusing can also be defined.

The way in which the condition parts are gradually specialized is as follows [98]: The initial (most general) condition part is based on a 9 centisecond window of codes (9 symbols) centred at the code symbol to be transformed. The symbol to be the condition part will be the simple majority of codes in this window. At the next two specializing stages the context expands 16 centiseconds to both sides of the initial window. Thereafter only focusing is used to specialize the rules. These three parts are gradually decomposed into smaller ones, from which the majorities are again computed. The majority voting also has the property of smoothing out the irregular or random errors in the quasiphoneme sequences, that occur more frequently than in phoneme sequences. The presented scheme is thus a combination of expansion and focusing, and it was experimentally found to give the best results. After having defined this hierachy of importance, the rule set construction part of the original DEC-algorithm can again be used as such.

4.3.3 Mapping based on Feed-Forward Nets

Following the idea of NETTALK, MFNs were also experimented with in this work in the same task. The basic principle is exactly the same as with the DEC and the DFC: Use the classification of context as parallel inputs to a MFN, and train it to produce the corresponding ideal label. The difference compared to the DFC is that there is not any kind of weighting for different parts of the context. All of the context is fed to a MFN with equal importance.

Three variants of this scheme were developed. After first trying to feed a wide context of quasiphonemes as such to the network, majority windows were incorporated to extend the coverage of input context without excessively increasing the number of input units to the network [51]. However, this approach was replaced by another one inspired by the DFC. Multiple resolution focusing windows were adopted as parallel inputs to the net [101]. Simple majorities of codes within windows were also replaced by proportions of different quasiphoneme labels in each of the windows. Finally, an experiment was conducted to study whether the phonemic interpretations given to the vector quantizer could be replaced by a representation of the speech signal [102]. Sequential dependencies in speech can be represented by studying the sequences of SOM grid locations. This representation is based on the topological ordering of the model vectors in the two-dimensional grid of a SOM [44]. This scheme is almost equivalent to the previous one except that SOM grid locations are used as inputs to the net instead of their labels. The performance, however, fell short of the previous approach. This would indicate that emphasizing phoneme differences by LVQ is more advantageous than trying to construct a SOM grid based representation of the speech signal and to classify such representations, at least in this speaker-dependent case.

4.3.4 Relaxation Labeling

Relaxation labeling is a method widely used in image analysis to improve the consistency of contradictory or inconsistent classification [34], [29]. We applied it also to quasiphoneme data, after having derived probabilistic classification information from the LVQ [53]. However, in experiments its performance fell slightly short of that of the DFC and MFNs (The comparison results are yet to be published in [103]). The reason for this might be that relaxation is basically a heuristic approach, and it does not seem to be applicable to data that has details of varying size, for example, short plosive bursts and long double vowels. Relaxation tends to favor long steady periods at the expense of short details. Under the circumstances, it seems better to use methods that can be trained by explicit examples to take these different kind of phenomena into account.

4.4 Conclusion

In conclusion, the subject of this thesis was to demonstrate that it is possible to improve short-time feature vector based phonemic speech recognition significantly by taking the classification of local context into account. The following general statement by Haralick [29] was thus shown to be applicable also to phonemic speech recognition: "More is to be gained by discovering suboptimal ways of handling context than by discovering optimal ways of handling the local structure".

Two main approaches that learn from examples were developed to utilize this context information. The DFC derived from the DEC of Kohonen is able to take into account the order of importance of context. The other approach based on MFNs lacks this property. On the other hand, MFNs can be more easily fed with uncertain or fuzzy information. Both of the developed approaches highlight one of the main points in pattern recognition: It is important first to construct a suitable or even an intelligent representation of the data (in this case the focusing context windows), and only then to try to classify it.

In the future, this research could be expanded to other languages, such as English. Another direction might be speech spoken in a more continuous manner. A further future line of research in connection to MFNs could be to incorporate probabilistic classification information into the network. This information could be integrated over each multiple resolution window. Relaxation labeling could also be developed so as to better take into account the charcteristics of different phonemes, such as lengths.

5 Summary of the Publications

This thesis consists of eight publications dealing with the problem of improving phonemic classification of speech by taking context into account. The publications can be divided into the following three categories:

1. Alignment of speech with corresponding phonemic transcriptions to obtain labeled training data (Publication 1).
2. Correction of phonemic transcriptions by using phoneme-level context (Publications 3 and 8).
3. Taking context into account at quasiphoneme level (Publications 2, 4, 5, 6, 7, and 8).

The publications are presented in chronological order that reflects the progress of ideas during this work.

In **Publication 1** a system to align speech with the corresponding phonetic transcriptions is described. The aligment algorithm first splits the speech signal into stationary events, after which an interpretation of them is searched for on the basis of the given correct phonemic transcription of the utterance. The search for an interpretation can be based strongly on the quasiphoneme sequence, because an LVQ-tuned SOM is already able to produce a rather accurate raw phonemic classification. Thus, the need to build an expert system to aid the segmentation, or extensive manual segmentation, is avoided. The system is in frequent use in building the speech data base of the Laboratory of Information and Computer Science. The resulting alignment is also used to derive the training data for the quasiphoneme mapping methods studied in this thesis.

Publication 2 is the first paper in this thesis where quasiphoneme-level context is taken into account. The DEC is applied quite straightforwardly to quasiphoneme sequences in order to derive a mapping to ideal sequences. A problem in applying the DEC directly to fine-grained data is that the rules tend to become too specific. This problem was handled by compressing the condition parts of the rules to increase generality. About 22% of the errors could be corrected with this scheme.

Publication 3 concerns the use of context at phoneme level, i.e., after phonemic decoding. A scheme using feed-forward networks was developed to correct errors in phoneme strings. This scheme is compared to the DEC at the same task. The experiments indicated that while the DEC was able to correct 50% of the errors in the transcriptions, the same figure was 37% for the feed-forward net. On the basis of this publication it could be concluded that the DEC was able to concentrate better on more relevant context when symbolic phoneme data were used.

Publication 4 and the rest of the publications concentrate on using context at the quasiphoneme level. A scheme using feed-forward nets to replace unwanted coarticulatory effects by explicit transition codes in quasiphoneme sequences is introduced. The durational segmentation rules remained unnecessary due to removal of coarticulatory effects. Better results were achieved than by using short-time feature vector classifications with durational rules.

Publication 5 can be considered as the main part of this thesis. A modification of the DEC called Dynamically Focusing Context (DFC) is presented. Since the relevant scale in speech is more than a single quasiphoneme symbol, an alternative to expanding context by single symbols was developed. A large area of speech is taken as the initial context, and its resolution is focused to solve the conflicts between rules. In this publication, the aim was to create a mapping to the ideal quasiphoneme sequence as in Publication 2. The experiments indicated that the method is able to correct 38% of the errors measured at the phoneme level. The method was also studied in combination with the DEC. The correction capability of the DEC alone was almost the same as that of the DFC. Using first the DFC at quasiphoneme level and the DEC at phoneme level about 50% of the errors could be corrected.

The quasiphoneme mapping scheme based on feed-forward nets introduced in Publication 4 was further developed in **Publication 6**. Inspired by the focusing resolution context windows adopted for the DFC, a multiple resolution input structure was introduced here too. In the experiments this method attained almost the same performance level as the DFC.

A further modification to the quasiphoneme mapping scheme is presented in **Publication 7**. Instead of having a LVQ-based quasiphoneme sequence as the input to the feed-forward network, the topological ordering of the speech data in a SOM is utilized. Trajectories formed on the map are fed into a classifier. Viewing the trajectories at multiple resolution levels, feature vectors are formed that take contextual information into account. Experiments with such feature vectors indicated that better accuracies are obtainable than by using a LVQ-tuned SOM classifier based on instantaneous acoustic features. However, the performance fell short of the DFC and the quasiphoneme-based feed-forward net mapping.

Finally, in **Publication 8** the best two of the developed quasiphoneme mapping methods are presented in the framework of a whole phonemic speech recognizer. Other improvements to the basic system at the feature computation level and the phonemic decoding level are also studied. On the basis of the experiments, it can be concluded that the two developed methods increase significantly the phonemic recognition accuracy, and when used with durational rules, approach the performance of hidden Markov model based phonemic decoding. In this publication the performance figures of the two developed methods compared to, for example, phoneme level DEC are not as good as reported in previous publications. Previously pre-segmented utterance limits were used, and the unvoiced plosives were treated as a single class. Errors due to these were thus not present in the figures. Here the figures include errors due to detection of the limits of the utterance and due to unvoiced plosives. This kind of errors are independent of how close the quasiphoneme sequences are to the ideal (mapped by the developed methods). However, these errors can be taken into account by the phoneme-level DEC. Furthermore, in the HMM-based decoding statistical models are built to describe the silence and/or noise outside the actual utterance as well as for the unvoiced plosives. This gives a clear advantage over heuristic RMS-based rules to detect utterance limits, and over the single-shot plosive recognition, that are used with durational rule-based decoding, and thus included in the figures of the two developed methods.

6 Contributions of the Author

In all of the cases the author has been mainly responsible for producing the actual publications. In Publication 2 the idea of applying the DEC to quasiphoneme sequences was developed by both of the authors, whereas the present author did all the programming and the experiments. The whole idea and development of the DEC is, of course, due to Professor Kohonen. In Publication 3 the author was responsible for the part concerning the DEC. The author also developed the feed-forward net based scheme to be suitable for correcting phonemic transcriptions, especially deletion errors. The co-author ran the experiments concerning feed-forward nets. In Publication 4 the idea of using majority windows to extend context coverage was due to the author. The co-author did the programming work and ran the experiments. In Publication 6 the DFC part of the work is due to the author. For the feed-forward net part, the author developed the multiple resolution context windows, whereas the co-author was responsible for the programming work and running the experiments. In Publication 7 the idea of using SOM-coordinates instead of phoneme labels as input for the feed-forward net can be credited to both of the authors. The present author was responsible of formulating the approach as taking advantage of the topology-preserving properties of SOMs and modifications connected to that. In Publication 8 the author performed the DFC- and the DEC-experiments. The author also directed the parts of the work that were concerned with feature selection, relaxation labeling, and HMM-based phonemic decoding.

References

[1] William A. Ainsworth. *Speech Recognition by Machine.* IEE Computing series. Peter Peregrinus Ltd., 1988.

[2] Anders Baekgaard and Paul Daalsgaard. Recognition of continuous speech using neural nets and expert system. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech89)*, volume 2, pages 545–548, Paris, France, September 26-28 1989.

[3] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. Speech recognition with continuous parameter hidden Markov models. In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 40–43, New York City, USA, April 1988.

[4] Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, March 1983.

[5] Frederic Bimbot, Gerard Chollet, Paul Deleglise, and Claude Montacie. Temporal decomposition and acoustic-phonetic decoding of speech. In *Proceeding of the IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing, pages 445–448, New York City, USA, April 1988.

[6] Herve Bourlard and Christian J. Wellekens. Multilayer perceptrons and automatic speech recognition. In *Proceeding of the First International Conference on Neural Networks*, volume IV, pages 407–416, San Diego, June 1987.

[7] Herve Bourlard and Christian J. Wellekens. Speech dynamics and recurrent neural networks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages S 33–36, Glasgow, Scotland, May 1989.

[8] Sing-Tse Bow. *Pattern Recognition.* Marcel Dekker Inc., 1984.

[9] Martin Brenner, Harald Hoge, Erwin Marschall, and Jorge Romano. Word recognition in continuous speech using a phonological based two-network matching parser and a synthesis-based prediction. In *Proceedings of the IEEE 1989 International Conference on Acoustics, Speech, and Signal Processing (ICASSP89)*, pages 457–460, Glasgow, Scotland, May 23-26 1989.

[10] Khalid Choukri. Speech recognition using connectionist approaches. In *Proceedings of the Annual ESPRIT Conference*, pages 835–850, Brussels, Belgium, November 12-15 1990. Kluwer.

[11] Ronald A. Cole and Lily Hou. Segmentation and broad classification of continuous speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 453–456, New York City, USA, April 1988.

[12] Paul Daalsgaard. Semi-automatic phonemic labelling of speech data using a self-organizing neural network. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech89)*, volume 2, pages 541–544, Paris, France, September 26-28 1989.

[13] Paul Daalsgaard and Anders Baekgaard. Recognition of continuous speech using neural nets and expert system processing. *Speech Communication*, 9(5/6):509–520, December 1990.

[14] Stephen B. Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, August 1980.

[15] Anne-Marie Deroualt. Context-dependent phonetic Markov models for large vocabulary speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP87)*, volume 1, pages 360–363, Dallas, Tx., April 6-9 1987.

[16] L.D. Erman and V.R. Lesser. The HEARSAY-II speech understanding system: A tutorial. In Lea [55], chapter 16.

[17] F. Fallside, H.Lucke, T.P. Marsland, P.J.O'Shea, M. St J. Owen, R.W. Prager, A.J. Robinson, and N.H. Russel. Continuous speech recognition for the TIMIT database using neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 445–448, Albuquerque, NM, USA, April 3-6 1990.

[18] Gunnar Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, 1970.

[19] J. Flanagan. *Speech Analysis, Synthesis, and Perception.* Springer, Berlin, Heidelberg, 1972.

[20] D. Fohr, N. Carbonell, and J.P. Haton. Phonetic decoding of continuous speech with the APHODEX expert system. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech89)*, volume 2, pages 609–612, Paris, France, September 26-28 1989.

[21] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, New York, London, 1972.

[22] Sadaoki Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59, February 1986.

[23] Sadaoki Furui. *Digital Speech Processing, Synthesis, and Recognition.* Marcel Dekker Inc., New York, Basel, 1989.

[24] John S. Garofolo. *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic-Phonetic Continuous Speech Database.* National Institute of Standards and Technology (NIST), Gaithersburg, MD, 1988.

[25] James R. Glass and Victor W. Zue. Multi-level acoustic segmentation of continuous speech. In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 429–432, New York City, USA, April 1988.

[26] Morris Halle. Phonology. In Osherson and Lasnik [80], chapter 3.

[27] Seppo Haltsonen. Improvement and comparison of three phonemic segmentation methods of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1160–1163, Atlanta, Georgia, USA, March 30,31, April 1 1981.

[28] Toshiyuki Hanazawa, Kenji Kita, Satoshi Nakamura, Takeshi Kawabata, and Kiyohiro Shikano. ATR HMM-LR continuous speech recognition system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 53–56, Albuquerque, NM, USA, April 3-6 1990.

[29] Robert M. Haralick. Decision making in context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(4):417–428, July 1983.

[30] Nobuo Hataoka and Alexander Waibel. Speaker-independent phoneme recognition on TIMIT database using integrated time-delay neural networks (TDNNs). Technical Report CMU-CMT-89-115, Carnegie Mellon University, November 27 1989.

[31] Hynek Hermansky. An efficient speaker-independent automatic speech recognition by simulation of some properties of human auditory perception. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP87)*, pages 1159–1162, Dallas, Tx., April 6-9 1987.

[32] John Hertz, Anders Krogh, and Richard G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley Publishing Company, 1991.

[33] William Y. Huang and Richard P. Lippmann. Comparison between neural net and conventional classifiers. In *Proceedings of the First International Conference on Neural Networks*, volume IV, pages 485–493, San Diego, June 1987.

[34] Robert A. Hummel and Steven W. Zucker. On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(3):267–287, May 1983.

[35] Melvyn J. Hunt and Claude Lefebvre. A comparison of severeal acoustic representations for speech recognition with degraded and undegraded speech. In *Proceedings of the IEEE 1989 International Conference on Acoustics, Speech, and Signal Processing (ICASSP89)*, pages 262–265, Glasgow, Scotland, May 23-26 1989.

[36] H. Iwamida, S. Katagiri, E. McDermott, and Y. Tohkura. A hybrid speech recognition system using HMMs with an LVQ-trained codebook. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 489–492, Albuquerque, NM, USA, April 3-6 1990.

[37] Jari Kangas and Teuvo Kohonen. Transient map method in stop consonant discrimination. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech89)*, Paris, France, September 26-28 1989.

[38] R.L. Kashyap and Mahesh C. Mittal. Recognition of spoken words and phrases in multitalker environment using syntactic methods. *IEEE Transactions on Computers*, 27(5):442–452, May 1978.

[39] Donald G. Kimber, Marcia A. Bush, and Gary N. Tajchman. Speaker-independent vowel classification using hidden Markov models and LVQ2. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 497–500, Albuquerque, NM, USA, April 3-6 1990.

[40] Teuvo Kohonen. Dynamically Expanding Context, with application to the correction of symbol strings in the recognition of continuous speech. In *Proceedings 8th International Conference on Pattern Recognition (8th ICPR)*, pages 1148–1151, Paris, France, Oct. 27-31 1986.

[41] Teuvo Kohonen. Learning vector quantization for pattern recognition. Technical Report TKK-F-A601, Helsinki University of Technology, 1986.

[42] Teuvo Kohonen. An introduction to neural computing. *Neural Networks*, 1(1):3–16, 1988.

[43] Teuvo Kohonen. The 'Neural' Phonetic Typewriter. *IEEE Computer*, 21(3):11–22, March 1988.

[44] Teuvo Kohonen. *Self-Organizing and Assosiative Memory.* Springer, Berlin-Heidelberg-New York-Tokio, 3 edition, 1989.

[45] Teuvo Kohonen, Gyorgy Barna, and Ronald Chrisley. Statistical pattern recognition with neural networks: Benchmarking studies. In *Proceedings of the IEEE International Conference on Neural Networks*, pages I 61–68, San Diego, July 1988.

[46] Teuvo Kohonen, Kai Mäkisara, and Tapio Saramäki. Phonotopic maps – insightful representation of phonological features for speech recognition. In *Proceedings of the 7th International Conference on Pattern Recognition (7th ICPR)*, pages 182–185, Montreal, Canada, July 1984.

[47] Teuvo Kohonen, Kimmo Raivio, Olli Simula, Olli Venta, and Jukka Henriksson. An adaptive discrete-signal detector based on self-organizing maps. In *Proceedings of the International Joint Conference on Neural Networks*, pages II 249–252, Washington D.C., January 1990.

[48] Teuvo Kohonen, Heikki Riittinen, Erkki Reuhkala, and S. Haltsonen. On-line recognition of spoken words from a large vocabulary. *Information Sciences*, 33:3–30, 1984.

[49] Teuvo Kohonen, Kari Torkkola, Makoto Shozakai, Jari Kangas, and Olli Ventä. Implementation of a large vocabulary speech recognizer and phonetic typewriter for Finnish and Japanese. In *Proceedings of the European Conference on Speech Technology*, pages 377–380, Edinburgh, U.K., September 1987.

[50] Teuvo Kohonen, Kari Torkkola, Makoto Shozakai, Jari Kangas, and Olli Ventä. Phonetic Typewriter for Finnish and Japanese. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP88)*, pages 607–610, New York City, USA, April 1988.

[51] Mikko Kokkonen and Kari Torkkola. Using self-organizing maps and multi-layered feed-forward nets to obtain phonemic transcriptions of spoken utterances. *Speech Communication*, 9(5-6):541–549, December 1990.

[52] Yasuhiro Komori, Kaichiro Hatazaki, Takaharu Tanaka, and Takeshi Kawabata. Combining phoneme identification neural networks into an expert system using spectrogram reading knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 505–508, Albuquerque, NM, USA, April 3-6 1990.

[53] Mikko Kurimo. Relaksaation käyttö kvasifoneemien tunnistuksessa, December 1990. Special project (erikoistyö), Helsinki Univ. of Technology, Lab. of Information and Computer Science (in Finnish).

[54] Lori F. Lamel, R.H. Kassel, and Stephanie Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings of DARPA Speech recognition Workshop*, pages 100–109, February 1986.

[55] Wayne A. Lea, editor. *Trends in Speech Recognition.* Prentice-Hall, Englewood Cliffs, N.J., USA, 1980.

[56] Chin-Hui Lee, Biing-Hwang Juang, Frank K. Soong, and Lawrence R. Rabiner. Word recognition using whole word and subword models. In *Proceedings of the IEEE 1989 International Conference on Acoustics, Speech, and Signal Processing (ICASSP89)*, pages 683–686, Glasgow, Scotland, May 23-26 1989.

[57] Kai-Fu Lee. *Automatic Speech Recognition: The Development of the SPHINX System.* Kluwer Academic Publishers, Boston, 1989.

[58] Kai-Fu Lee. Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(4):599–609, April 1990.

[59] Kai-Fu Lee and Hsiao-Wuen Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, November 1989.

[60] Hong C. Leung and Victor W.Zue. Some phonetic recognition experiments using artificial neural nets. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP88)*, pages 422–425, New York City, USA, April 11-14 1988.

[61] S.E. Levinson. Structural methods in automatic speech recognition. *Proceedings of the IEEE*, 73:1625–1650, 1985.

[62] S.E. Levinson, M.Y. Libermann, A. Ljolje, and G.L. Miller. Speaker independent phonetic transcription of fluent speech for large vocabulary speech recognition. In *Proceedings of the IEEE 1989 International Conference on Acoustics, Speech, and Signal Processing (ICASSP89)*, pages 441–444, Glasgow, Scotland, May 23-26 1989.

[63] S.E. Levinson, A. Ljolje, and G.L. Miller. Continuous speech recognition from a phonetic transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 93–96, Albuquerque, NM, USA, April 3-6 1990.

[64] J-S. Lienard, M. Mlouka, J.J. Mariani, and J. Sapaly. Real-time segmentation of speech. In *Preprints of the Speech Communications Seminar*, volume 3, pages 183–187, Stockholm, Sweden, August 1-3, 1974 1974.

[65] Jean-Sylvain Lienard. Perception, data variability and inductive inference. In *Proceedings of COGNITIVA-90*, pages 154–166, Madrid, Spain, November 20-23 1990.

[66] Richard P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, April 1987.

[67] Richard P. Lippmann. Review of neural networks for speech recognition. *Neural Computation*, 1(1):1–38, 1989.

[68] V. Le Maire, R. Andre-Obrecht, and D. Jouvet. An acoustic-phonetic decoder based on an automatic segmentation algorithm. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech89)*, volume 2, pages 392–395, Paris, France, September 26-28 1989.

[69] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63:561–580, 1975.

[70] J. Mariani. Recent advances in speech processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP89)*, pages 429–440, Glasgow, U.K., May 23-26 1989.

[71] Erik McDermott and Shigeru Katagiri. Shift Invariant, Multi-Category Phoneme Recognition Using Kohonen's LVQ2. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 81–84, Glasgow, Scotland, May 1989.

[72] G. Mercier, D. Bigorgne, L. Miclet, L. Le Guennec, and M. Querre. Recognition of speaker-dependent continuous speech with KEAL. *IEE Proceedings*, 136(2), April 1989.

[73] Masanori Miyatake, Hidefumi Sawai, Yasuhiro Minami, and Kiyohiro Shikano. Integrated training for spotting japanese phonemes using large phonemic time-delay neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 449–452, Albuquerque, NM, USA, April 3-6 1990.

[74] Yeshwant K. Muthusamy, Ronald A. Cole, and Malcolm Slaney. Speaker-independent vowel recognition: Spectrograms versus cochleagrams. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 533–536, Albuquerque, NM, USA, April 3-6 1990.

[75] Cory S. Myers and Lawrence R. Rabiner. Connected digit recognition using a level-building DTW algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):351–363, 1981.

[76] Seiichi Nagakawa and Yoshimitsu Hirata. Comparison among time-delay neural networks LVQ2, discrete parameter HMM and continuous parameter HMM. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 509–512, Albuquerque, NM, USA, April 3-6 1990.

[77] Hermann Ney. Dynamic programming parsing for context-free grammars in continuous speech recognition. *IEEE Transactions on Signal Processing*, 39(2):336–340, February 1991.

[78] Douglas O'Shaughnessy. *Speech Communication*. Addison-Wesley, 1987.

[79] Douglas O'Shaughnessy. *Speech Perception*, chapter 5. In [78], 1987.

[80] Daniel N. Osherson and Howard Lasnik, editors. *An Invitation to Cognitive Science: Language*, volume 1. MIT Press, 1990.

[81] Yoh-Han Pao. *Adaptive Pattern Recognition and Neural Networks*. Addison-Wesley Publishing Company, Inc., 1990.

[82] R.W. Prager and F. Fallside. The modified Kanerva model for automatic speech recognition. *Computer Speech and Language*, 3:61–81, 1989.

[83] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[84] Lawrence R. Rabiner and Ronald W. Schaefer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[85] D. Raj Reddy. Speech recognition by machine: A review. *Proceedings of the IEEE*, 64(4):502–531, April 1976.

[86] Stephen Renals and Richard Rohwer. Learning phoneme recognition using neural networks. In *Proceedings of the IEEE 1989 International Conference on Acoustics, Speech, and Signal Processing (ICASSP89)*, pages 413–416, Glasgow, Scotland, May 23-26 1989.

[87] Tony Robinson and Frank Fallside. Phoneme recognition from the TIMIT database using recurrent error propagation networks. Technical Report CUED/F-INFENG/TR.42, Cambridge University Engineering Department, Trumpington Street, Cambridge, England, March 1990.

[88] Tony Robinson, John Holdsworth, Roy Patterson, and Frank Fallside. A comparison of preprocessors for the Cambridge recurrent error propagation network speech recognition system. In *Proceedings of 1990 International Conference on Spoken Language Processing (ICSLP90)*, volume 1, pages 1033–1036, Kobe, Japan, November 18-22 1990.

[89] S. Roucos and M.O. Dunham. A stochastic segment model for phoneme based continuous speech rcognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP87)*, volume 1, pages 73–76, Dallas, Tx., April 6-9 1987.

[90] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, volume 1. Foundations, chapter 8: Learning Internal Representations by Error Propagation, pages 318–362. MIT Press/Bradford, Cambridge, 1986.

[91] Hiroaki Sakoe. Two-level DP-matching - a dynamic programming-based pattern matching algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(6):588–595, 1979.

[92] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[93] Terrence Sejnowski and Charles R. Rosenberg. Parallel networks that learn to pronounce english text. *Complex Systems*, 1(1):145–168, 1987.

[94] Stephanie Seneff. A computational model for the peripheral auditory system: Application to speech recognition research. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 1983–1986, Tokyo, Japan, April 7-11 1986. IEEE.

[95] A.R. Smith and M. Sambur. Hypothesizing and verifying words for speech recognition. In Lea [55], pages 139–165.

[96] Charles W. Therrien. *Decision, Estimation, and Classification.* John Wiley and Sons, 1989.

[97] Kari Torkkola. Yhteysherkät korjailusäännöt puheentunnistuksessa. Licentiate's thesis, Helsinki University of Technology, 1988. (in Finnish).

[98] Kari Torkkola. A combination of neural network and low level AI-techniques to transcribe speech into phonemes. In *Proceedings of COGNITIVA-90*, pages 637–644, Madrid, Spain, November 20-23 1990.

[99] Kari Torkkola, Jari Kangas, Pekka Utela, Sami Kaski, Mikko Kokkonen, Mikko Kurimo, and Teuvo Kohonen. Status report of the Finnish phonetic typewriter project. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN-91)*, Espoo, Finland, June 24-28 1989. (to be published).

[100] Kari Torkkola and Teuvo Kohonen. Correction of quasiphoneme strings by the dynamically expanding context. In *Proceedings of the 9th International Conference on Pattern Recognition (9th ICPR)*, pages 487–489, Rome, Italy, November 14-17 1988.

[101] Kari Torkkola and Mikko Kokkonen. A comparison of two methods to transcribe speech into phonemes: A rule-based method vs. back-propagation. In *Proceedings of 1990 International Conference on Spoken Language Processing (ICSLP90)*, volume 1, pages 673–676, Kobe, Japan, November 18-22 1990.

[102] Kari Torkkola and Mikko Kokkonen. Using the topology-preserving properties of SOFMs in speech recognition. In *Proceedings of the IEEE 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP91)*, Toronto, Canada, May 14-17 1991. (to be published).

[103] Kari Torkkola, Mikko Kokkonen, Mikko Kurimo, and Pekka Utela. Improving short-time speech frame recognition results by using context. In *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech91)*, Genova, Italy, September 24-26 1991. (to be published).

[104] Kari Torkkola and Kimmo Raivio. Comparison of symbolic and connectionist approaches to eliminate coarticulation effects in phonemic speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech89)*, volume 1, pages 9–12, Paris, France, September 26-28 1989.

[105] Julius T. Tou and Rafael C. Gonzalez. *Pattern Recognition Principles.* Addison-Wesley Publishing Company, 1974.

[106] E.J.M. van Mierlo, E. Blaauw, and G. Bloothooft. Phoneme segmentation of speech, based on temporal decomposition using band filter spectra and phonetic rules. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech89)*, volume 2, pages 71–74, Paris, France, September 26-28 1989.

[107] Olli Ventä. *Associative and syntactic text correction methods for continuous speech recognition.* PhD thesis, Helsinki University of Technology, 1990.

[108] Enrique Vidal and Andres Marzal. A review and new approaches for automatic segmentation of speech signals. In *Proceedings of EUSIPCO-90, Fifth European Signal Processing Conference*, volume I, pages 43–53, Barcelona, Spain, September 18-21 1990.

[109] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin Lang. Phoneme Recognition Using Time-Delay Neural Networks. *IEEE, Transactions on Acoustics, Speech and Signal Processing*, 37(3):328–339, March 1989.

[110] R. L. Watrous, L. Shastri, and Alexander H. Waibel. Learned phonetic dicrimination using connectionist networks. In *Proceedings of the European Conference on Speech Techology*, volume 1, pages 377–380, Edinburgh, Scotland, September 1987.

[111] Clifford Weinstein, Stephanie McCandless, Lee Mondshein, and Victor Zue. A system for acoustic-phonetic analysis of continuous speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):54–67, February 1975.

[112] Terry Winograd. Computer software for working with language. *Scientific American*, 251(3):90–101, 1984.

[113] J. Wiren and H.L. Stubbs. Electronic binary selection system for phoneme classification. *Journal of Acoustical Society of America*, 28(6):1082–1091, 1956.

[114] George Yu, William Russel, Richard Schwartz, and John Makhoul. Discriminant analysis and supervised vector quantization for continuous speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 685–688, Albuquerque, NM, USA, April 3-6 1990.

[115] B.P. Yuhas, M.H. Goldstein Jr., Terrence J. Sejnowski, and R.E. Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, 78(10):1658–1668, October 1990.

[116] Victor Zue, James Glass, David Goodine, Michael Phillips, and Stephanie Seneff. The SUMMIT speech recognition system: Phonological modelling and lexical access. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP90)*, volume 1, pages 49–52, Albuquerque, NM, USA, April 3-6 1990.

[117] Victor Zue, James Glass, Michael Phillips, and Stephanie Seneff. Acoustic segmentation and phonetic classification in the SUMMIT system. In *Proceedings of the IEEE 1989 International Conference on Acoustics, Speech, and Signal Processing (ICASSP89)*, pages 389–392, Glasgow, Scotland, May 23-26 1989.

[118] Victor W. Zue. Experiments on spectrogram reading. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 116–119. IEEE, 1979.

[119] Victor W. Zue. The use of speech knowledge in automatic speech recognition. *Proc. of the IEEE*, 73(11):1602–1615, November 1985.

[120] Victor W. Zue and Richard M. Schwartz. Acoustic processing and phonetic analysis. In Lea [55], pages 101–124.