IDIAP RESEARCH REPORT

New Nonsense Syllables Database – Analyses and Preliminary ASR Experiments

Petr Fousek [a b] Petr Svojanovský [a c]
František Grézl [a c] Hynek Hermansky [a c]

IDIAP–RR 04-29

May 2004

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP Research Institute, Martigny, Switzerland
[b] CTU Prague, Faculty of Electrical Engineering, Prague, Czech Republic
[c] Brno University of Technology, Faculty of Information Technology, Brno, Czech Republic

IDIAP Research Report 04-29

New Nonsense Syllables Database – Analyses and Preliminary ASR Experiments

Petr Fousek   Petr Svojanovský   František Grézl   Hynek Hermansky

May 2004

submitted for publication

Abstract. In the first half of the 20th century, series of experiments on human perception of nonsense syllables were carried out at Bell Laboratories. Since the original data were never recorded, Linguistic Data Consortium designed and recorded a corpus which loosely corresponds to the original setup. This may eventually allow for the replication of the perceptual experiments. The preliminary version of the corpus was recently made available to us and we used for the first series of machine recognition experiments, reported here. In order to understand the corpus and prepare it for phoneme recognition experiments, semi-automatic analyses were done on speech data as well as on prompt texts. Results have shown that the database still contained some artifacts to be fixed. A modified structure of the database was made. Some files containing errors were removed, all files were extended with emulated silence at boundaries, and a phonetic transcription with phoneme-alignment was made using forced alignment. Preliminary phoneme recognition experiments were done with TIMIT-trained models, syllables-adapted TIMIT models as well as with syllables-trained models. The performance was evaluated not only in terms of the overall error rates (that are still rather low - between 40-60%) but also using a set of evolving alternative criteria, that may allow for additional insights into the performance of the different recognition approaches.

1 Introduction

The *Syllables* corpus was designed to loosely correspond to Harvey Fletcher's perceptual experiments carried out since 1919 in order to allow for their replication, since the original data appear to be never recorded [1]. The corpus can also be used for phoneme-recognition experiments.

The Linguistic Data Consortium allowed us to use the pre-release version of the database. Due to a large amount of data, a semiautomatic analyses were performed. It has revealed that the database still contains some undesirable artifacts, which were isolated. Performed analyses and their conclusions are described in following sections as well as all the changes that were made to the corpus.

2 Preparation of database for recognition experiments

2.1 Database description

English syllables, some of which are real words, but most of which are nonsense syllables, were recorded in a quiet and anechoic environment. Speakers were asked to say a set of 2000 syllables *common* to all speakers, and also a set of 20 syllables *unique* to that speaker. As a result a following corpus was created.

- Over 33 hours of speech,
- 16 kHz PCM and 8 kHz μ-law (telephone line) data,
- 20 speakers (12 M + 8 F),
- about 2000 files per speaker,
- 400 syllables unique to a speaker (20 for each speaker),
- roughly 2000 syllables common to all speakers,
- each syllable recorded within phrase and isolated.

2.1.1 Syllable selection

First, all diphone syllables allowed in English were used. These were consonant-vowel (CV) and vowel-consonant (VC) and accounted for over 600. Second, to reflect the natural distribution of syllables in spoken English, the word-frequency table from Switchboard I (recorded pre-arranged telephone conversations corpus) was used and more than 1300 triphone syllables (CVC, CCV, VCC) were extracted from the most common words. No syllables containing schwa were used.

2.1.2 Prompts design

Each speaker pronounced each syllable exactly twice: once in a carrier phrase, and once in an isolation, see prompt structure in fig. 1.

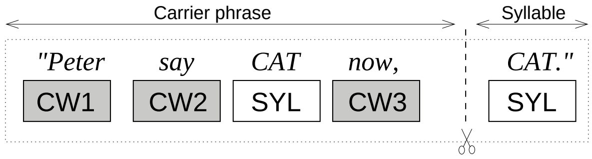

Figure 1: *Prompt structure with an example text.*

The carrier words *CW1, CW2, CW3* represented a subject, verb, and adverb. They were chosen randomly from appropriate words sets. If syllables were real English words, they were put directly to the prompt; when they were nonsense, a word containing the syllable was used instead (letters not to be pronounced were parenthesized).

2.1.3 Recording and segmentation

All recordings were made in the same environment. The prompt displaying was controlled by speaker, which had to read the text and press the "next" button after each prompt. If the text was badly pronounced then the whole sentence was re-recorded. Speakers were asked to say the phrase fluently and pause at the comma after the carrier phrase so that the second occurrence of syllable was isolated.

Recordings were made in wide-band (16 kHz, 16-bit PCM) and narrow-band (8 kHz μ-law transmitted via telephone channel) versions. The wide-band data were listened to and each recording was manually split into *phrase* file and isolated *syllable* file. The narrow-band data were aligned with the wide-band data using cross-correlation. Some of the files in narrow-band were not recorded (about 10%).

Some parts of the text in this section were adopted from the database documentation [2], which contains more details.

2.2 Corpus analyses

The aim of the analyses was to understand the corpus and to detect possible outlier files. All analyses were done on the wide-band version of data.

2.2.1 SNR check

A ratio between the beginning/end part of speech file and a part with expected speech activity (segment symmetrically surrounding the highest absolute value in the file) was evaluated. The length of all compared segments was 25 ms, which corresponds to typical speech segmentation for short-term analysis.

The analysis has shown very similar SNRs for boundaries of both phrase and syllable files. The average SNR was 45.6 dB. Files with values lower than 20 dB were audio-visually checked. The low SNR was caused either by breath-in of speaker or by starting of speech, no errors were observed.

2.2.2 Check for sequences of zeros

Since longer sequences of zeros within speech file can cause numerical problems, in each file the longest zero sequence was found. Results have shown that the zero sequences are not the problem in this corpus.

2.2.3 Check for high energy at speech file boundaries

All files were checked for the energy within 25 ms frame at boundaries. The goal was to find files which might had been cut in a wrong position while being created from long files. A high energy in these regions could mean that there was a part of speech. Since the files were clean, energy could be used as a criterion.

First 100 files out of each subset (phrases and syllables) were audio-visually checked. The improper cut was found in 7 files and these files were removed.

2.2.4 Prompt texts check

A check of prompt texts consistency was performed. Since the prompt structure was always the same, <*CW1 space CW2 space SYL space CW3 comma space space SYL*>, prompts were searched for this pattern. There were 29 prompts found where the syllable was missing. These files were discarded, because their possible correction could have been problematic.

2.3 Padding with silence

Some feature extraction techniques require long time context and in *Syllables* there are almost no silence parts at files boundaries. All files were therefore extended in the beginnings and ends with 600 ms of emulated silence.

In each file a 25 ms segment with the lowest energy was found. This segment was repeated 24 times at the beginning and end of the file.

All further sections refer to silence-padded data.

2.4 Phoneme alignment of the database

For the purposes of additional analyses a phoneme alignment was done using forced alignment of the phonetic transcription of the data. Following set of TIMIT-trained models was used.

- 42 context-independent phoneme models,
- 2 silence models <sp>, <sil>,
- 5 emitting states per model,
- 32 Gaussian mixtures per state[1].

A phonetic transcription of *Syllables* was made using Switchboard/ICSI Meetings pronunciation dictionary. Initial models were retrained on *Syllables* by running Baum-Welch algorithm 6 times prior to the final alignment.

2.4.1 Analysis of aligned data

The point of our concern were the phonemes with long duration, since the extreme duration of a phoneme can mean that the word transcription in prompt file does not correspond to the text actually spoken.

Ten longest occurrences of each phoneme were found. Files containing these occurrences were audio-visually checked. 21 files containing either a repetition or a corrected mispronunciation of a word were found. All of these have been discarded.

Besides truly long phonemes (mainly in syllables), some phonemes were extended to a neighboring silence or breath noise. This mainly affected plosive closures. In these 56 occurrences the alignment was manually corrected.

Average phoneme lengths were computed for syllable-files and phrase-files. Comparison to TIMIT lengths is shown in fig. 2.

2.5 Final notes

If either a text prompt, isolated syllable or syllable in a phrase were found and error, all files containing the given syllable were removed. Table 1 compares the original and the resulting corpus[2]

3 Preliminary recognition experiments

To gain the first experience with *Syllables* database, we set up following recognition experiments. All experiments were run on wide-band data. The database was divided into three subsets.

Training set – phrase files from half of speakers.

Test set 1 – the rest of phrase files.

Test set 2 – isolated syllable files from all speakers.

[1] This setup was found to be optimal for TIMIT phoneme recognition task.

[2] Numbers for original corpus except for total number of files were adopted from documentation [2].

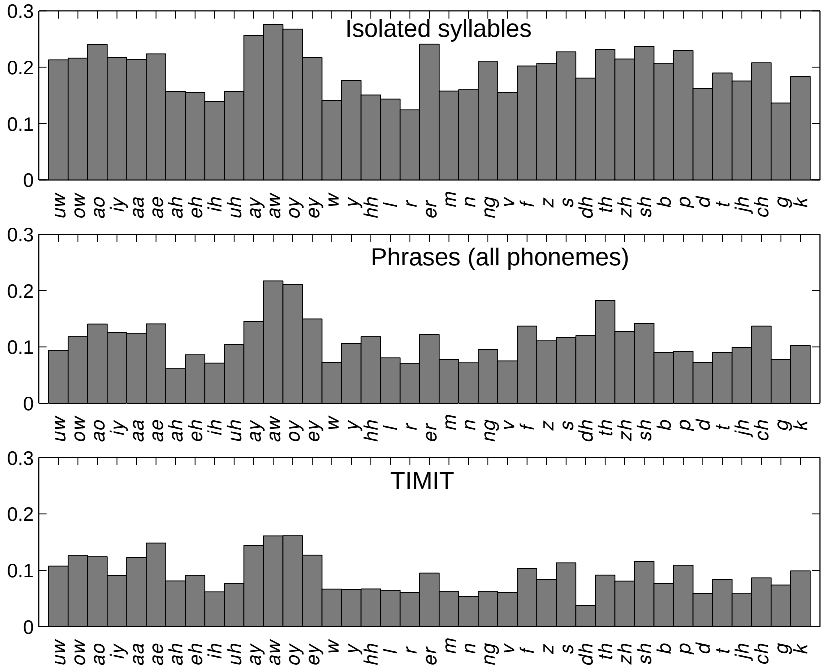

Figure 2: *Comparison of average phoneme lengths in TIMIT and Syllables, durations in seconds.*

Corpus	Original	Resulting
Common Total	2005	2005
Common by All	1845	1811
Unique	400	393
Total files	40278	40217

Table 1: Syllable counts. *Common Total* – all "common" syllables; *Common by All* – syllables pronounced by each speaker; *Unique* – syllables unique to a speaker; *Total files* – number of pairs phrase-syllable.

MFCC features were extracted from the whole database (segmentation 25/10ms, 12 cepstral coeffs. + energy + Δ + $\Delta\Delta$ coeffs.). The same set of 42 + 2 TIMIT models as described in section 2.4 was used in back-end.

3.1 Recognition with known number of phonemes

Both test sets were recognized using TIMIT-trained phoneme models (the same set as was used for the forced alignment). In order to be close to Fletcher's experiments [3], recognizer grammar was restricted. Original human listeners knew the number of phonemes to expect and they also knew the structure of the carrier phrase.

Isolated syllables – In each file the recognizer was forced to find a sequence of phonemes <*pause – Phn1 – Phn2 (– Phn3) – pause*>. Number of phonemes was fixed to match prompt text.

Syllables in phrases – In phrases the grammar was <*pause – CW1 – CW2 – Phn1 – Phn2 (– Phn3) – short pause – CW3 – pause*>, while recognizer was given sets of possible carrier words *CW1, CW2, CW3.*

While evaluation, to focus only on phonemes in syllables, everything except for the sequence <*Phn1 – Phn2 (– Phn3)*> was discarded. Results were compared to a reference using NIST alignment procedure, which resulted in the same number of insertions and deletions. To incorporate both insertions and deletions into a confusion matrix, a virtual model called *No-phoneme* (Null) was introduced.

3.1.1 Results

Results of the experiment with isolated syllables were plot in the confusion matrix (see fig. 5), which was rearranged according to phonetic categories introduced in [3] and normalized with respect to phoneme frequencies shown below.

Rows represent confusion of expected phoneme with other phonemes and sum of each row equals one. The diagonal is then a measure of model accuracy (lower values are caused either by low quality of model or by model mismatch), see fig. 3.

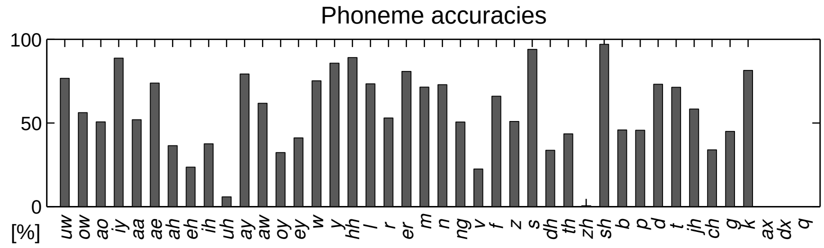

Figure 3: *Accuracy of models.*

Columns of the normalized matrix represent relative model usage (how often the model was picked for a phoneme). A sum over each column then determine a relative frequency of model usage – model "priority", which is shown in fig. 4.

It can be observed that most confusions happen within phonetic categories (*b,p,d,t,jh,ch – v,f,z,s,dh,th – m,n,ng*). Confusions in vowels are widespread due to high number of classes.

There are significant asymmetries in the matrix. For example phoneme *uh* is quite often substituted (see also fig. 3) and almost never used (see fig. 4). On the contrary, phonemes *d,t* often substitute others within the same phonetic category (see also fig. 4).

Last column and row reflect misrecognitions. Phoneme *q* is often inserted (substitutes *No-phoneme*) even if it is not expected at all. Stop consonants *b,p,d,t* are often deleted (substituted by *No-phoneme*).

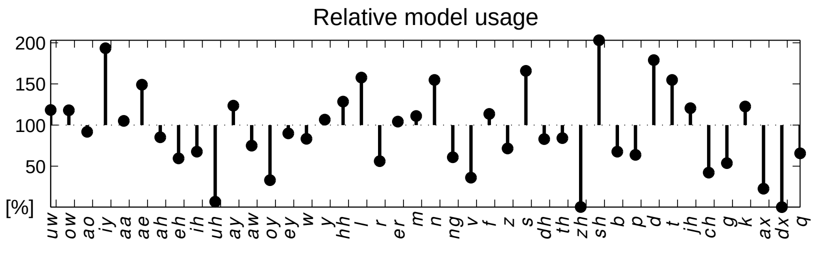

Figure 4: *Relative frequency of model usage.*

3.2 Blind phoneme recognition

For a comparison purposes, both test sets were recognized using TIMIT models with no prior information (44-phonemes recognition). In case of syllables any sequence of phonemes was allowed, in

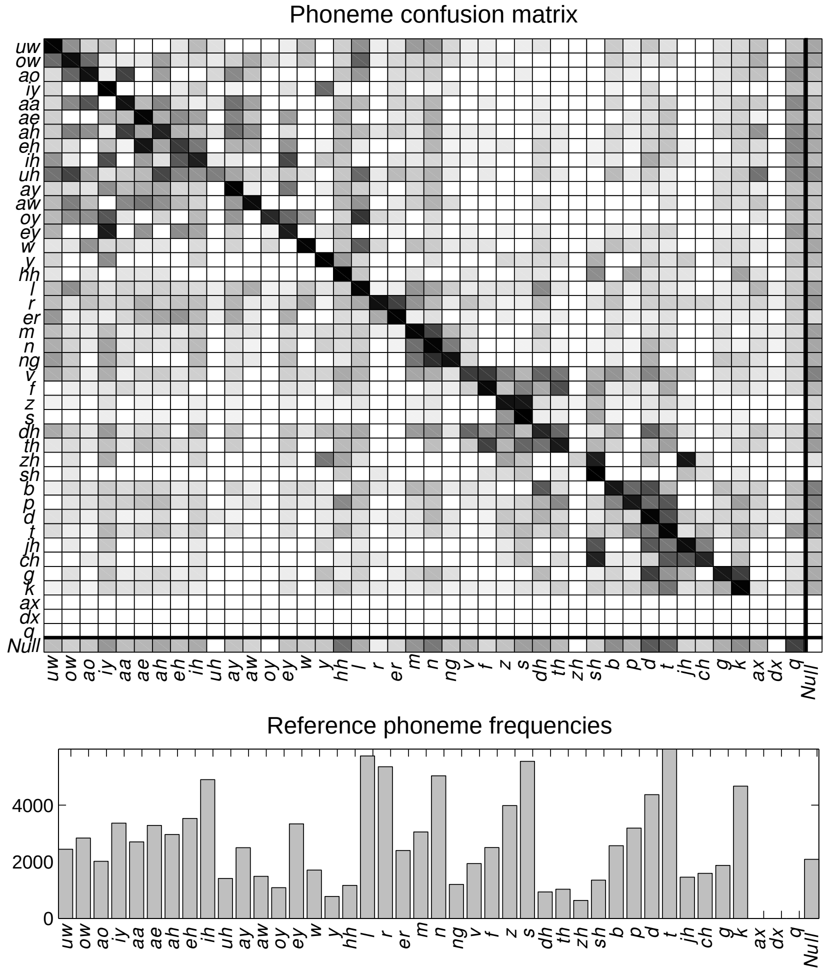

Figure 5: *Confusion matrix for isolated syllables with known number of phonemes. Rows – stimuli, columns – observation.*

phrases the grammar was <*pause – CW1 – CW2 – any sequence of phonemes – short pause – CW3 – pause*>, given carrier words sets.

Since there were only several phonemes expected in each tested file (only one syllable), the silence models <sil> and <sp> were not considered in evaluation (except for confusions with other models). The experiment was tuned to reach a balanced ratio of insertions and deletions.

Overall recognition accuracies are compared in tab. 2.

3.3 Recognition with re-estimated models

TIMIT phoneme models were re-estimated on *Syllables* using Baum-Welch procedure on training set and both test sets were recognized. Half of speakers from isolated syllables test set were also present in train set. The relative improvement was 35.5% in phrases and 23.4% in isolated syllables.

4 Short Discussion and Conclusions

The pre-release version of the Syllables database was checked and used in the first series of ASR experiments. The attempt was made to design the experiments so that they would relate to perceptual

Acc[%]	known # of phns	"blind" recog.
SYLs in phrases	51.5	44.5
SYLs isolated	59.0	44.8

Table 2: *Overall recognition accuracy*

experiments done at Bell Laboratories in the first half of the last century [1]. Comparing to the human performance reported then, the observed machine performance is so far rather low.

5 Acknowledgments

We thank the Linguistic Data Consortium, Nelson Morgan and Jont B. Allen for their initiative in initiating, designing and recording the Syllables database. The work was supported by DARPA under the EARS Novel Approaches grant no. MDA972-02-1-0024. The other sources of support were the IM2 Swiss National Center for Competence in Research, managed by Swiss National Science Foundation on behalf of Swiss authorities, and the European Community AMI and M4 grants.

References

[1] Allen, J., "How Do Humans Process and Recognize Speech?", IEEE Interactions on Speech and Audio Processing, Vol. 2, No. 4, Oct 1994. speech corpus", NISTIR 4930 , feb 1993

[2] Wright, J., "Articulation Index Corpus", corpus documentation on DVD, 2003.

[3] Fletcher, H., "Speech and Hearing in Communication", Ac. Soc of America, 1995.