

# PRONUNCIATION MODELLING IN THE RWTH LARGE VOCABULARY SPEECH RECOGNIZER

K. Beulen, S. Ortmanns, A. Eiden, S. Martin, L. Welling, J. Overmann, H. Ney  
Lehrstuhl für Informatik VI, RWTH Aachen, University of Technology, D-52056 Aachen

## SUMMARY

In this paper we describe the application of pronunciation variants for our large vocabulary continuous speech recognizer. We will explain how the pronunciation variants were used in training and recognition and give some recognition results on three different corpora. The recognition tests were performed on the *Wall Street Journal* (WSJ) November 92 development and evaluation corpora (5 000 words), the *North American Business* (NAB) H1 development corpus (20 000 words) and on the *Verbmobil* 1996 evaluation corpus (5 000 words). For the WSJ and NAB corpora, a slight improvement in recognition accuracy can be observed, while for the *Verbmobil* corpus the error rate remains unchanged.

In addition, we will discuss the incorporation of phrases in combination with pronunciation variants in the pronunciation lexicon as well as the language model. The recognition results on the WSJ November 92 development and evaluation corpora show that the main improvement due to phrases is caused by the language model.

## 1. INTRODUCTION

Pronunciation variants have been successfully used in several speech recognition systems. In [1] an improvement in the error rate of about 2-3% for pronunciation variants only in recognition and 6% for pronunciation variants in training and recognition is reported. In [6] the improvement for pronunciation variants only in recognition is about 2% while for pronunciation variants in training and recognition the reduction is significantly higher, about 7%. In our recognition system, the use of pronunciation variants for the English corpora showed an improvement of 2-6% of the error rate, while for the German corpus the error rate did not change. Additionally exploiting the pronunciation variants for the training has only a slight effect in the specific test corpus we have used.

In addition, we will present our work on phrases. Phrases are short sequences of words which occur very frequently in human speech. These phrases can be used to enhance acoustic modelling by capturing the across-word effects between the component words of the phrase, and for language modelling to

enlarge the span of the language model. Using phrases, the error rate on the *Wall Street Journal* (WSJ) November 92 test corpus can be reduced by about 8%, in combination with pronunciation variants even by about 13% relative.

## 2. PRONUNCIATION VARIANTS

One source of the variability of the speech signal are pronunciation variants of words. In order to model them in speech recognition, they can simply be added to the standard lexicon. The lexica we used for our recognition tests were produced in two different manners. For the WSJ and NAB tasks, we used the Philips lexicon as described in [1]. Here, the pronunciation variants were generated by a set of rules for suffixes like “-ally” and context dependent modifications like the “t” in “twen(ty)”. For function words as “the” additional pronunciation variants were added by hand. For the German *Verbmobil* task [3], the transcriptions of the training corpus already contain the pronunciation variants transcribed manually. So we simply counted these pronunciation variants and then added the most frequent ones to the lexicon. In both cases, the vocabulary size was increased by about 10%.

### 2.1. TRAINING

To use the pronunciation variants in the acoustic training, it is necessary to recognize the correct pronunciation variant for each word in the training corpus. For that, we implemented a search algorithm using a linear lexicon [8] where the word transitions are constrained by the sequence of words in the sentence, and trained acoustic monophone models with low acoustic resolution (about 8 densities per mixture). Using this algorithm, a corrected transcription for the training corpus is determined. With this new transcription a second training is performed, then the training corpus is retranscribed again and so on. This procedure is iterated until no (major) changes in the transcription are noticed. For the WSJ S184 training corpus two iterations of this procedure are sufficient. Then this final transcription is used to train the acoustic models for recognition.

To enhance the robustness of pronunciation variant recognition, we incorporated prior probabilities for the pronunciation variants. To do so, we simply counted the frequencies of the

pronunciation variants of a word in the training corpus. These frequencies were then used as a unigram language model during the recognition of the pronunciation variants. For unseen words, we assumed an equal distribution of the pronunciation variants in the training corpus.

## 2.2. RECOGNITION

In recognition, the pronunciation variants are treated as separate words. Each pronunciation variant is represented by one leaf in the lexical tree. At word transitions where the language model has to be incorporated, a mapping function of the pronunciation variants onto the lexical word in the language model is used. Every pronunciation variant of a lexical word results in the same language model probability, although one can think of incorporating the prior probability of a pronunciation variant given the lexical word.

## 3. PHRASES

Although it is possible to form a sentence by using any combination of words from a given lexicon, there are certain combinations of words which occur very frequently in human communication. These word combinations are called phrases. Examples are "it\_is" or "there\_are" for two word phrases or "are\_in\_the" for a three word phrase. Because of their high frequency of occurrence, it is advantageous to treat them in a special way for speech recognition, which will be discussed in the following sections.

### 3.1. PHRASES IN SPEECH RECOGNITION

For speech recognition, it is possible to add such phrases to the lexicon as single words. The two benefits are:

- The language model treats these phrases as single words so the span of the modelled word dependencies is increased.
- The acoustic models are able to consider the acoustic context at the word boundaries of the phrase components (across-word context).

It is obvious that according to the two effects mentioned above, there are two steps for incorporating the phrases in the speech recognition system:

- Retrain the language model using the defined phrases.
- Add the phrases to the lexicon with respect to the acoustic context at the word boundaries.

Our experimental results show that the improvement of the error rate is mainly due to the language model while the modifications of the acoustic modelling have almost no effect.

For our recognition system, we used the phrases which were included in the Philips lexicon [5]. These phrases were selected by the counting method described in [5] where the threshold

for including the phrase was about 7000. For these phrases a bigram language model was calculated as follows. First of all, a pass over the training corpus was made to replace the proper word sequences by the phrases. Using this modified corpus, a reasonable language model for phrases can be trained.

Next, the phrases were added to the pronunciation lexicon. Therefore, the pronunciations of the words were combined to form a phrase. This was done in two ways:

- The transcriptions were combined with a silence phoneme between the component words. Hence the triphones at the word boundaries were independent of the across-word context so the acoustic modelling is unchanged.
- The transcriptions were combined without a silence phoneme. This has the benefit that the across-word context can be taken into account for the word boundaries. Since this results in a lot of unseen triphones in the recognition corpus, it may be necessary to retrain the acoustic models. However, in our experiments the acoustic models were not retrained.

These two possibilities of combining the pronunciations can be used separately or, which will be shown to perform best, can be combined to allow both alternatives (coarticulation and no coarticulation).

### 3.2. PHRASES AND PRONUNCIATION VARIANTS

An obvious next step is to combine phrases and pronunciation variants. This is very simple because the only effort which has to be made is to examine each phrase and, if it contains a word with a pronunciation variant, add a copy of the phrase with the proper word replaced by this variant. In the worst case this can result in an exponential growth of the number of phrases in the lexicon. For the lexicon which was used in the recognition tests, the number of phrases was only enlarged by a factor of 2-3.

## 4. RESULTS

The speech recognition system which was used for the tests is described in [9]. It has the following properties:

- 16 cepstral coefficients together with 16 first and one second order derivatives resulting in a 33 component Cepstrum vector [10],
- dimension reduction by LDA [2],
- continuous HMM with Laplacian mixture densities,
- gender dependent models,
- one single vector of absolute deviations for all distributions,
- HMM generalized triphone models without across-word modelling [4],
- Viterbi approximation for training,

- word conditioned search algorithm using a lexical prefix tree in combination with a bigram language model for recognition [7].

The tests were performed on three different corpora:

- WSJ-5k

The training was done on the WSJ SI84 training data, the tests on the WSJ November 92 development & evaluation data for 5 000 word lexicon (WSJ-5k).

- NAB-20k

The training was done on the WSJ SI284 training data, the tests on the *North American Business* (November '94) H1 development data for 20 000 word lexicon (NAB-20k).

- VM-5k

The training was done on the Verbmobil 1996 training data, the tests on the Verbmobil 1996 evaluation data for 5 000 word lexicon (VM-5k).

Table 1 contains the results on these corpora for the baseline lexica (*BASE*), for pronunciation variants only for recognition (*REC*), and for pronunciation variants both in training and recognition (*TRN & REC*)

**Table 1. Results for pronunciation variants on WSJ-5k and NAB-20k and on VM-5k.**

Corpus	WER [%]		
	BASE	REC	TRN & REC
WSJ-5k	6.9	6.5	6.6
NAB-20k	16.3	16.0	–
VM-5k	25.5	25.6	–

For WSJ-5k and NAB-20k, there is a slight improvement of about 2-4% relative for pronunciation variants only in recognition. The additional use of pronunciation variants in training which was only tested for WSJ-5k does not change the result. For VM-5k, no gain in recognition accuracy is observed.

Table 2 focuses on the recognition of the pronunciation variants for transcription correction. Here, we have tested the effect of using single densities versus mixture densities (8 mixture components), and the benefit of additional prior probabilities for the pronunciation variant recognition. These tests were performed only on the male corpus of WSJ-5k:

**Table 2. Results for pronunciation variants in training and different recognition methods, WSJ-5k male test set, 10 speakers, 4067 spoken words.**

method	WER [%]
REC	6.8
TRN & REC, 1 density	7.6
TRN & REC, 8 densities	7.1
TRN & REC, 8 densities & priors	7.2

For eight densities, the error rates for the recognition with and without priors are slightly worse compared to the result with pronunciation variants only in recognition while for single densities the result is deteriorated by more than 10% relative.

Table 3 contains the results with phrases on the WSJ November 92 5k corpus. For phrases without any coarticulation modelling (“ONLY SIL”), the error rate is improved by about 8% relative. For phrases with coarticulation modelling between the phrase components (“NO SIL”) the error rate is almost the same as for the baseline result. That means, the effect of the improved language model (“ONLY SIL”) is completely compensated by the modified acoustic modelling. By a combination of both methods (“BOTH”), meaning two representations in the lexical tree for each phrase, one with coarticulation modelling between *all* word transitions and one without, the gain in error rate is slightly higher, about 10% relative.

It can be argued that for phrases containing more than two words all combinations of coarticulation/no coarticulation at the word boundaries have to be added to the lexicon. Taking into account that for this corpus less than 10% of the recognized words are phrases and about 90% of the phrases are two word phrases, the possible improvements are expected to be negligible. However, additional tests have to be carried out to clarify this.

**Table 3. Results for phrases, WSJ-5k, 18 speakers, 12137 spoken words.**

phrase type	number of phrases	WER [%]
NONE	0	6.9
ONLY SIL	229	6.5
NO SIL	229	7.0
BOTH	458	6.3

Table 4 shows the results for the combination of phrases and pronunciation variants. Again the plain application of the modified language model reduces the error rate by about 7% relative, while adding the across-word modelling deteriorates it. A combination of both models yields the same error rate as the method without any across-word modelling.

**Table 4. Results for phrases and pronunciation variants in recognition, WSJ-5k, 18 speakers, 12137 spoken words.**

phrase type	number of phrases	WER [%]
NONE	0	6.5
ONLY SIL	229	6.1
NO SIL	229	6.7
BOTH	458	6.1

## 5. CONCLUSIONS

In this paper we have presented the application of pronunciation variants in combination with phrases for our large vocabulary continuous speech recognition system. The results with pronunciation variants show an improvement of about 2-6% for WSJ-5k and NAB-20k while for VM-5k the error rate remains unchanged. The additional use of pronunciation variants in training for the WSJ-5k corpus does not change the results significantly. However, since other groups report improvements by using pronunciation variants in training, additional tests on other corpora have to be carried out. For phrases the reduction of the error rate is about 8% relative on WSJ-5k. This improvement is mainly due to the increased span of the language model while the changes in the acoustic modelling we made show only a slight influence on the error rate. The reason for this surprising result can be that there was no acoustic retraining carried out due to the unseen across-word triphones. Another potential problem is that we have not added all combinations of coarticulation/no coarticulation word transitions for the phrases to the lexicon. Additional tests have to be carried out to clarify these points. In combination with phrases the error rate on WSJ-5k can be reduced by 13% compared to the baseline system.

## 6. ACKNOWLEDGEMENTS

We would like to thank *Philips Research Labs Aachen* for providing the lexica for the English corpora, and especially Dietrich Klakow for additional information about the phrases in the lexicon. This research was partly funded by grant 01 IV 701 T4 from the German Ministry of Science and Technology (BMBF) as a part of the Verbmobil project.

## REFERENCES

- [1] X. Aubert, C. Dugast, "Improved Acoustic-Phonetic Modelling in Philips' Dictation System by Handling Liaisons and Multiple Pronunciations," *Proc. Europ. Conf. on Speech Communication and Technology*, Madrid, Spain, Vol. 1, pp. 767-770, September 1995.
- [2] K. Beulen, L. Welling, H. Ney, "Experiments with Linear Feature Extraction in Speech Recognition," *Proc. Europ. Conf. on Speech Communication and Technology*, Madrid, Spain, pp. 1415-1418, September 1995.
- [3] T. Bub, W. Wahlster, A. Waibel, "Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 71-74, Munich, Germany, April 1997.
- [4] K. Beulen, E. Bransch, H. Ney, "State Tying for Context Dependent Phoneme Models," *Proc. Europ. Conf. on Speech Communication and Technology*, Rhodes, Greece, pp. 1179-1182, September 1997.
- [5] D. Klakow, "Language Model Optimization by Mapping of Corpora," to appear in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998.
- [6] L. Lamel, G. Adda, "On Designing Pronunciation Lexicons for Large Vocabulary Continuous Speech Recognition," *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, PA, Vol. 1, pp. 6-9, September 1996.
- [7] H. Ney, R. Haeb-Umbach, B.-H. Tran, M. Oerder, "Improvements in Beam Search for 10000-Word Continuous Speech Recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 13-16, San Francisco, CA, March 1992.
- [8] H. Ney, D. Mergel, A. Noll, A. Paeseler, "Data Driven Search Organization for Continuous Speech Recognition," *IEEE Transactions on Signal Processing*, Vol. 40, Nr. 2, February 1992.
- [9] H. Ney, L. Welling, S. Ortmanms, K. Beulen, F. Wesel, "The RWTH Large Vocabulary Continuous Speech Recognition System," to appear in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998.
- [10] L. Welling, N. Haberland, H. Ney, "Acoustic Front-End Optimization for Large Vocabulary Speech Recognition," *Proc. Fifth European Conference on Speech Communication and Technology*, pp. 2099-2102, Rhodes, Greece, September 1997.