USING DUTCH PHONOLOGICAL RULES TO MODEL PRONUNCIATION VARIATION IN ASR

Mirjam Wester, Judith M. Kessens & Helmer Strik

A^2RT, Dept. of Language and Speech, University of Nijmegen, the Netherlands
{M.Wester, J.Kessens, W.Strik}@let.kun.nl

Abstract

In this paper, we describe how the performance of a continuous speech recognizer for Dutch has been improved by modeling within-word and cross-word pronunciation variation. Within-word variants were automatically generated by applying five phonological rules to the words in the lexicon. Cross-word pronunciation variation was modeled by adding multi-words and their variants to the lexicon. The best results were obtained when the cross-word method was combined with the within-word method: a relative improvement of 8.8% in the WER was found compared to baseline system performance. We also describe an error analysis that was carried out to investigate whether rules in isolation can predict the performance of rules in combination.

1. Introduction

The present research concerns the continuous speech recognition component of a spoken dialogue system called OVIS (Strik et al., 1997). OVIS is employed to automate part of an existing Dutch public transport information service. A large number of telephone calls of the on-line version of OVIS have been recorded and are stored in a database called VIOS. The speech material consists of interactions between man and machine. The data show that the manner in which people speak to OVIS varies, ranging from using hypo-articulated speech to hyper-articulated speech. As pronunciation variation degrades the performance of a continuous speech recognizer (CSR) –if it is not properly accounted for– solutions must be found to deal with this problem. We expect that by modeling pronunciation variation some of the errors introduced by the various ways in which people address the system will be corrected. Ultimately our aim is to develop a method for modeling Dutch pronunciation variation that can be used to tackle the problem of pronunciation variation for Dutch CSRs.

Phonus 5, Institute of Phonetics, University of the Saarland, 2000, ??-??.

In this paper, we will first present the method we used to model within-word and cross-word pronunciation variation (Section 2). Section 3 deals with the main characteristics of our CSR and the speech material that we used. The results in terms of word error rates are presented in Section 4.1, followed by an analysis of the changes that take place due to modeling pronunciation variation in Section 4.2. Subsequently, in Sections 4.3 and 4.4, we address the problem of whether rules tested in isolation can predict the behavior of rules in combination. Finally, to conclude this paper, conclusions are drawn and the implications of the results are discussed.

2. Method

The type of pronunciation variation, which is modeled in the current experiments, consists of variation at the segmental level. More specifically, we investigated insertions and deletions of phones. Furthermore, a distinction was made between within-word and cross-word pronunciation variation. In this research, we used a general method, which consists of modeling pronunciation variation at three different levels: the lexicon, the phone models and the language model. In short, variants are added to the lexicon and language model and phone models are retrained using a corpus with variants included in the transcriptions. In our experiments, the effect of modeling pronunciation variation is measured at each level. In this way, four test conditions are obtained which are shown in Table 1. “S” denotes the use of single pronunciations; “M” denotes the use of multiple pronunciations.

Table 1. Test conditions

	test condition	**Lexicon**	**phone models**	**language model**
baseline	SSS	S	S	S
level 1	MSS	M	S	S
level 2	MMS	M	M	S
level 3	MMM	M	M	M

The first level at which we modeled pronunciation variation was in the lexicon. As most speech recognizers make use of a lexicon, this has been a frequently used approach to modeling pronunciation variation (Strik & Cucchiarini, 1999). There are various ways to obtain variants which can be added to the lexicon. For instance, rules can be used to generate variants, the variants can be obtained from a pronunciation dictionary, or the variants can be selected on the basis of their frequency of occurrence in the data. In our research, we used the rule-based approach. We formulated rules for five frequently occurring Dutch phonological processes: /n/-deletion, /r/-deletion, /t/-deletion, schwa-insertion, and schwa-deletion (Kessens et al., 1999b). The rules, which are context dependent, were applied to the words in the canonical lexicon and the resulting variants were added to the lexicon. As we would like to find the optimal set of rules with which to model pronunciation variation, we not only tested the rules in combination, but we also tested each of the rules in isolation in order to find out if the results obtained for rules in isolation can predict how rules will behave in combination. This issue is central to any rule-based approach to pronunciation modeling.

Besides modeling pronunciation variation at the lexical level it can also be incorporated in the phone models (PMs). This can be done by automatically transcribing the training corpus using the CSR in forced recognition mode. Forced recognition, in our case, means that the recognizer does not choose between all the words in the lexicon, but only between the different pronunciation variants of the same word. In this way, the variant that most closely resembles the spoken word can be chosen. Experiments described in Kessens et al. (1998) and Wester et al. (1999) have shown that the performance of forced recognition is comparable to expert listeners in selecting the appropriate pronunciation variant. Pronunciation variation is incorporated in the phone models by retraining them using the transcriptions obtained with the forced recognition. It is to be expected that retraining will lead to better PMs. This process of retranscribing and retraining can be repeated in iteration to obtain increasingly better phone models until no changes occur. In general, most of the changes take place after the first iteration. Therefore, one iteration is usually sufficient (Kessens et al., 1999a).

The third level at which we modeled pronunciation variation is in the language model. To calculate the baseline language model, the orthographic representation of the words in the training corpus is used. Because there is only one variant per word this suffices. However, when there is more than one variant per word and this is not accounted for in the language model, all variants of the same word will have equal a priori probabilities, which may not suffice. Therefore, we also based the calculation of the language model on the phone transcriptions of the corpus obtained through forced recognition instead of on the original transcriptions.

In continuous speech, a substantial part of the variation occurs across word boundaries in addition to the variation that occurs within words. Therefore, we also paid attention to modeling some of the pronunciation variation that takes place across word boundaries. In Kessens et al. (1999b), we compared two different methods for modeling cross-word pronunciation variation. It was shown that using multi-words to model cross-word processes leads to better results than adding the variants of cross-word processes as separate items to the lexicon. Therefore, in this paper we only report on the cross-word method, which makes use of the multi-word approach.

Multi-words are word-sequences that are added to the lexicon as separate entities. Examples of multi-words (with the transcriptions of their variants between brackets) are: “het_is” (/hEtIs/, /@tIs/, /tIs/) and “is_het” (/IshEt/, /Is@t/, /Ist/)[1]. The multi-words used in these experiments were selected by first selecting the 50 most frequently occurring word sequences from the training material. Next, those words to which the cross-word processes of cliticization, contraction and reduction could apply were chosen from the list. This led to the selection of 22 multi-words (Kessens et al. 1999b). The cross-word variation was tested in isolation and the combination of within-word variation and cross-word variation was also tested. Here again the question was whether the sum of the effects of the methods tested in isolation could predict the total effect of testing the combination of the methods.

3. CSR and Material

The main characteristics of the CSR are as follows. The input signals consist of 8 kHz, 8 bit A-law coded samples. Feature extraction is done every 10 ms for 16 ms frames. The first step in feature analysis is an FFT analysis to calculate the spectrum. In the following step, the energy in 14 mel-scaled filter bands between 350 and 3400 Hz is calculated. Next, a discrete cosine transformation is applied to the log filterband coefficients. The final processing stage is a running cepstral mean subtraction. Besides 14 cepstral coefficients (c_0-c_{13}), 14 delta coefficients are also used. This makes a total of 28 feature coefficients.

The CSR uses acoustic phone models, word-based language models (unigram and bigram) and a lexicon. The acoustic models are continuous density hidden Markov

[1] Sampa notation is used throughout this paper. http://www.phon.ucl.ac.uk/home/sampa/dutch.htm

models (HMMs) with 32 Gaussians per state. The topology of the HMMs is as follows: each HMM consists of six states, three parts of two identical states, one of which can be skipped. In total, 39 HMMs were trained. For each of the phonemes /l/ and /r/, two models were trained, because a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/). For each of the other 33 phonemes context-independent models were trained. In addition, one model was trained for non-speech sounds and a model consisting of only one state was employed to model silence.

Our training and test material, selected from the VIOS database (Strik et al., 1997), consisted of 25,104 utterances (81,090 words) and 6,267 utterances (21,106 words), respectively. We used a third dataset for the purpose of error analysis. This corpus (further referred to as the error analysis corpus) contains the same type of speech material as the training and test corpora and consists of 6,245 utterances (18,371) words. There is no overlap between the three corpora.

4. Results

4.1. *Word error rates*

Table 2 shows the results for the various test conditions. For our baseline CSR, we used a canonical lexicon with one phone transcription for each word. The word error rate (WER=ins+del+subs/N) for the baseline system was 12.75%. Row 2 in Table 2 (within) shows the results of modeling within-word pronunciation variation. Each step in the method leads to an improvement. In total, a significant improvement of 0.68% was found (from SSS to MMM) for modeling within-word pronunciation variation.

The cross-word method in isolation does not lead to a significant change in performance. Besides, the improvement that is found is mainly due to adding the multi-words to the lexicon and language model. However, testing the cross-word method in combination with the within-word method shows a significant improvement of 0.44% in WER compared to the within-word condition in isolation (MMM). Summarizing, the best results are obtained when pronunciation variants are used during training and recognition, and when they are added to the language model (MMM). The total improvement from SSS to MMM is 1.12% WER absolute (8.8% relative).

Table 2. Results in WER for the various test conditions for different variants: within-word variants (within), cross-word variants (cross), and the combination of within-word with cross-word variants. Significant improvements, compared to SSS, are shown in bold.

	SSS	**MSS**	**MMS**	**MMM**
within	12.75%	12.44%	12.22%	**12.07**%
cross	12.41%*	12.74%	12.99%	12.45%
within + cross	12.41%*	12.37%	12.30%	**11.63**%

*=multi-words added to the lexicon and language model (no variants)

4.2. *Changes due to modeling pronunciation variation*

Up until now our results have been presented in terms of WER (as is done in most studies). Although WER gives a global idea of the merits of a method, it certainly does not reveal all details of the effect a method has. Therefore, we carried out an error analysis in which we compared the utterances recognized in the baseline test to those recognized in the test condition MMM for within + cross. The results in Table 3 show that 75.7% of the utterances are recognized correctly in both conditions, and 17.3% of the utterances are recognized incorrectly in both conditions. Improvements are found for 4.3% of the utterances, and deteriorations are found for 2.9% of the utterances.

The comparison of the utterances recognized differently in the two conditions can also be used to study how many changes truly occur. These results are presented in Table 4. The group of 1,083 utterances (17.3%) which are recognized incorrectly in both tests consist of 609 utterances (9.7%) for which both tests produce the same incorrect recognition results and 474 utterances (17.3 - 9.7 = 7.6%) with different mistakes. In addition, improvements were found for 267 utterances (4.3%) and deteriorations for 183 utterances (2.9%), as was already mentioned above. Consequently, the net result is an improvement for only 84 utterances (267 - 183), whereas in total the recognition result changes for 924 utterances (474 + 267 + 183). These changes are a consequence of our methods of modeling pronunciation variation, but they cannot be seen in the WER. The WER only reflects the net result obtained, and our error analysis has shown that this is only a fraction of what actually happens due to applying our methods.

Table 3. Comparison between baseline and MMM condition for within and cross-word variation: number of correct utterances, incorrect utterances, improvements and deteriorations

		SSS			
		correct		**incorrect**	
MMM within + cross	**correct**	4,743	75.7%	267	4.3%
	incorrect	183	2.9%	1,083	17.3%

Table 4. Types of changes in utterances going from the baseline condition to the MMM within + cross test condition

Type of change	**Number of utterances**
Same utterance, different mistake	474 (7.6%)
Improvements	267 (4.3%)
Deteriorations	183 (2.9%)
Net result	+84 (1.3%)

4.3. *Isolation versus combination?*

In order to get a first indication whether results obtained for rules in isolation can predict how rules will behave in combination, we employed the test corpus to test each of the five phonological rules in isolation and in combination. Figure 1 shows the differences in WER between the results of adding variants of each of the five phonological rules to the lexicon separately, the summation of these results (sum) and the result of the combination of all five rules (combi). The differences in Figure 1 are all on the basis of the MSS condition, i.e. variants are only added to the lexicon. These results seem to indicate that there is no way of predicting the result of a combination of rules on the basis of the rules in isolation. The principle of superposition clearly does not apply here.

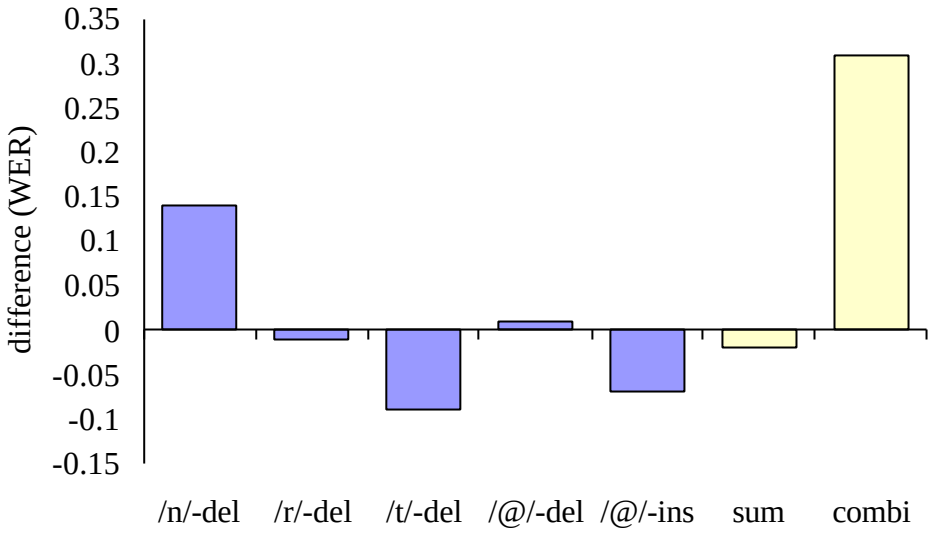


Figure 1. Difference in WER between SSS and MSS for each of the rules in isolation, sum of those results, and combination result of all rules (test corpus)

4.4. *Error analysis*

The set of tests described in the previous section was repeated using the error analysis corpus (described in Section 3). The use of another corpus was necessary because if we were to carry out any further detailed analysis of the test corpus its validity as an independent test set would be scrutinized. Therefore, in order to further investigate the isolation-combination problem, tests were carried out on the error analysis corpus. For the sake of comparison, the results of the tests on the error analysis corpus are presented in the same manner as the results for the test corpus (see Figure 2). It is clear that the overall picture is quite different. In Figure 1 it seems obvious that the superposition principle does not hold whereas in Figure 2 this is not nearly as evident. The differences between “sum“ and “combi“ are large in Figure 1 and extremely small by comparison in Figure 2.

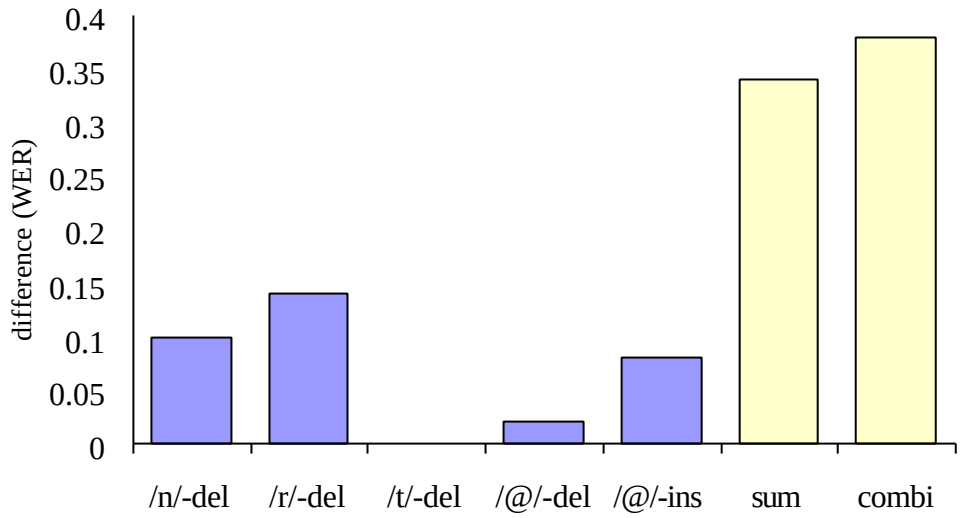
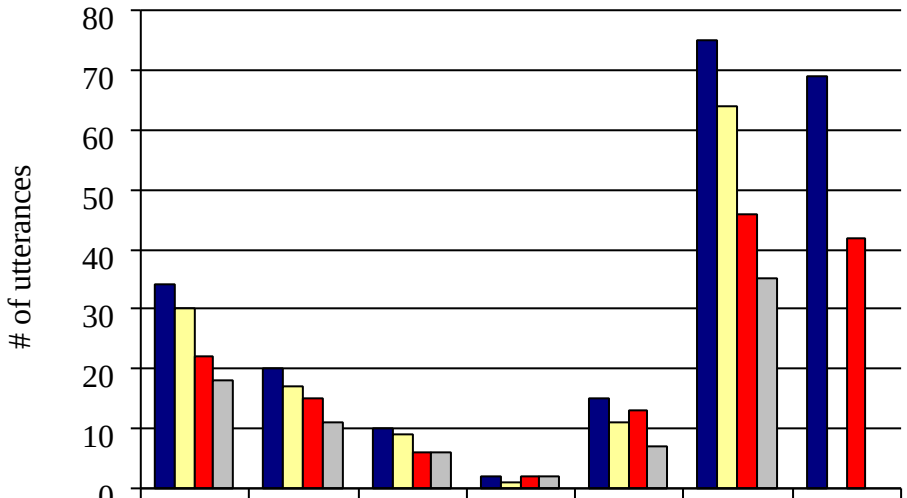


Figure 2. Difference in WER between SSS and MSS for each of the rules in isolation, sum of those results and combination result of all rules (error analysis corpus)

Above and also in Kessens et al. (1999b) we concluded that the principle of superposition does not apply for the five rules of the within-word method. Some possible explanations for this finding are: 1. More than one rule can apply to the same word, creating variants in combination that are not present in isolation. 2. Confusion can occur between pronunciation variants of different methods. 3. During decoding, the words in the utterances are not recognized independently of each other. To investigate if these explanations are correct we carried out a more detailed analysis of the results of the error analysis corpus.

Figure 3 shows the number of differently recognized utterances compared to the baseline due to adding variants of one of the rules to the lexicon, in addition to the sum of the individual rules and the combination result of all rules. The first bar (+) in each set indicates the number of improvements due to the addition of variants to the lexicon which are obtained by the application of a rule, i.e. the utterance was recognized incorrectly in the baseline test condition and correctly after adding the pronunciation variants. The second bar (++) indicates how many of those improvements are also present in the combination test. The third bar (-) indicates the number of deteriorations, i.e. the number of utterances that was recognized correctly in the baseline test condition but incorrectly after the addition of pronunciation variants. Finally, the fourth bar (--) indicates how many of those deteriorations are also found in the combination test. It

should be noted that sum is not simply the addition of improvements/deteriorations for each of the rules in isolation, because we only allowed each utterance to be counted once. In some cases more than one rule in isolation influences the recognition of the same utterance.



	/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins	sum	combi
+	34	20	10	2	15	75	69
++	30	17	9	1	11	64	x
-	22	15	6	2	13	46	42
--	18	11	6	2	7	35	x

Figure 3. Number of changed *utterances* due to adding variants of the separate rules to the lexicon in isolation, the sum of those results, and the result of the combination of the rules. For more information see text.

This analysis shows that many of the changes found in isolation are also found in combination. 93% (64/69) of the improvements found in the combination test condition are also found in isolation and 83% (35/42) of deteriorations. Figure 3 shows that five improvements and seven deteriorations are found in the combination test condition and not in isolation. Inspection of these utterances shows that one of the improvements and five of the deteriorations are the result of a combination of rules. This means that more than one rule could apply to the same word within an utterance. An example is shown in Table 5. It is unclear what caused the remaining four improvements and two deteriorations. The example in Table 5 shows the original transcription of an utterance (this is the transcription present in the baseline lexicon) followed by the result of

recognition with the baseline lexicon, and with the combination lexicon. When the baseline lexicon is used during recognition “Delft” is not recognized correctly. Whereas, when the combination lexicon is used the whole utterance is recognized correctly. This is probably due to the combination of the rules for schwa-insertion and /t/-deletion applied to the word “Delft”. In addition, /r/-deletion is applied to “Amsterdam” and this may also influence the correct recognition of the utterance.

Table 5. Example of how the combination of rules leads to correct recognition. R= recognized, O= original

		orthography<phone transcription>
O	transcription	delft<dELft> naar<na:R> amsterdam<Amst@RdAm>
R	baseline lex	terug<t@rYx> naar<na:R> amsterdam<Amst@RdAm>
R	combi lex	delft<dEl@f> naar<na:R> amsterdam<Amst@dAm>

6. Conclusions

Our conclusions are that modeling pronunciation variation by using phonological rules indeed improves the recognizer's performance. The best results are obtained when within-word variation is modeled in combination with cross-word variation, and when the variation is incorporated at all three levels: the lexicon, the phone models and the language models. In total, WER decreased by 1.12%; a relative improvement of 8.8%.

Figures 1 and 2 showed that differences in WER for methods in isolation and in combination on two different corpora (test and error analysis corpus) lead to two quite different pictures. It seems the results are corpus dependent. Furthermore, in both cases, the superposition principle does not apply. In section 4.2, we saw that the changes due to modeling pronunciation variation are only partially visible in terms of WER. Analysis on the level of utterances showed that 14.7% of the recognized utterances changed, whereas a net improvement of only 1.3% was found. A lot of changes occur that are not visible in the error rates. This implies that error rates alone are neither suitable for measuring the effect of modeling pronunciation variation, nor for analyzing the performance of rules in isolation vs. in combination.

A more detailed error analysis showed that a substantial part of the improvements/deteriorations, which were found in the results for a rule in isolation, can also be found in the results of the combination of various rules. Further error analyses and tests will show whether the behavior of rules in isolation –if correctly interpreted– can partially predict how they will perform in combination with other rules.

Acknowledgments

The research by Judith M. Kessens was carried out within the framework of the Priority Programme Language and Speech Technology, sponsored by NWO (Dutch Organization for Scientific Research). The research by Helmer Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

References

Kessens, J.M., Wester, M., Cucchiarini, C. and Strik, H. (1998). The Selection of Pronunciation Variants: Comparing the Performance of Man and Machine. *Proc. of the Int. Conf. on Spoken Language Processing*, Sydney, Vol. 6, 2715-2718.

Kessens, J.M., Wester, M. & Strik, H. (1999a). Modeling Within-word and Cross-word Pronunciation Variation to Improve the Performance of a Dutch CSR. *Proc. of the14th Int. Congress of Phonetic Sciences*, San Francisco, 1665 - 1668.

Kessens, J.M., Wester, M. & Strik, H. (1999b). Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation. *Speech Communication* **29**, 193-207.

Strik, H. & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication* **29**, 225-246.

Strik, H., Russel, A.J.M., van den Heuvel, H. Cucchiarini, C. & Boves, L. (1997). A spoken dialog system for the Dutch public transport information service. *International Journal of Speech Technology*, Vol. 2, No. 2, 119-129.

Wester, M., & Kessens, J.M. (1999). Comparison between Expert Listeners and Continuous Speech Recognizers in Selecting Pronunciation Variants. *Proc. of the14th Int. Congress of Phonetic Sciences*, San Francisco, 723 - 726.