PRONUNCIATION VARIATION IN ASR: WHICH VARIATION TO MODEL?

Mirjam Wester, Judith M. Kessens & Helmer Strik

A²RT, Dept. of Language and Speech, University of Nijmegen, the Netherlands
{M.Wester, J.Kessens, Strik}@let.kun.nl; http://lands.let.kun.nl/

ABSTRACT

This paper describes how the performance of a continuous speech recognizer for Dutch has been improved by modeling within-word and cross-word pronunciation variation. A relative improvement of 8.8% in WER was found compared to baseline system performance. However, as WERs do not reveal the full effect of modeling pronunciation variation, we performed a detailed analysis of the differences in recognition results that occur due to modeling pronunciation variation and found that indeed a lot of the differences in recognition results are not reflected in the error rates. Furthermore, error analysis revealed that testing sets of variants in isolation does not predict their behavior in combination. However, these results appeared to be corpus dependent.

1. INTRODUCTION

The present research deals with modeling Dutch pronunciation variation in ASR. Pronunciation variation is one of the factors that can degrade the performance of an ASR system, if it is not properly accounted for. Therefore, in recent years, a lot of attention has been paid to dealing with pronunciation variation in ASR and various methods have been proposed and tested [1]. In our work, we have been using a knowledge-based approach in which variants are generated using phonological rules [2]. This approach has led to a significant improvement in WER. However, one of the problems that remains is finding the optimal set of rules or variants with which to model the remaining pronunciation variation present in the speech material. Whether or not a data-driven or knowledge-based approach is used to obtain variants, it is necessary to have some way to decide which of these variants should be included in the lexicon.

In this paper, we address this issue in two different ways. First, by performing a detailed analysis of the differences in recognition results which occur due to modeling pronunciation variation. Second, by comparing results of rules tested in isolation and in combination in order to find out if the results obtained for rules in isolation can predict how rules will behave in combination.

2. METHOD

The general method that we use to tackle the problem of pronunciation variation operates at three different levels: the lexicon, the phone models (PMs) and the language model. In this section, modeling pronunciation variation at each of these levels is discussed. This is followed by a description of two types of pronunciation variation that we have modeled using this method: within-word and cross-word variation.

The first level at which pronunciation variation is modeled is in the lexicon. Pronunciation variants are added to the canonical lexicon (which contains a single transcription per word), thus, creating a multiple pronunciation lexicon.

To incorporate pronunciation variation in the PMs first forced recognition of the training data is carried out using a lexicon with multiple pronunciations per word. The recognizer aligns the signal with the closest matching pronunciation variant, thus including pronunciation variation in the transcription of the training corpus. Pronunciation variation is then integrated in the PMs by retraining them using these new transcriptions. Experiments described in [4] have shown that the performance of forced recognition is comparable to the performance of expert listeners in selecting the appropriate pronunciation variant.

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. However, when there is more than one variant per word the a priori probabilities for the different variants of that word are probably not equal and for that reason should not be based on the orthographic transcription. We therefore also calculated a language model based on the frequency counts of the variants in the training corpus, which was obtained through forced recognition.

In our experiments, the effect of modeling pronunciation variation is measured at each of the levels. In this way, you get the four test conditions shown in Table 1. “S” denotes the use of single pronunciations; “M” denotes the use of multiple pronunciations.

	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

Table SEQARABIC1: Test Conditions

Within-word variation was dealt with by using a rule-based approach. We selected five phonological processes, which are described in the literature, to formulate rules with which pronunciation variants were generated; i.e. /n/-deletion, /r/-deletion, /t/-deletion, schwa-insertion, and schwa-deletion [2]. (Sampa phoneme notation is used throughout this paper.) The rules, which are context dependent, were applied to the

words in the canonical lexicon and the resulting variants were added to the lexicon. We did not only test the rules in combination, but 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.

In continuous speech, a substantial part of the variation occurs across word boundaries in addition to the variation that occurs within words. In [2], we compared two different methods for modeling cross-word pronunciation variation. In the first method, we used multi-words, which are word-sequences that are added to the lexicon as separate entities. An example of a multi-word and its transcriptions is "het_is" (/hEtIs/, /@tIs/, /tIs/). The second method consisted of adding the separate parts of the multi-words to the lexicon. It was shown that using multi-words to model cross-word processes leads to better results than adding the variants as separate items to the lexicon. Therefore, in this paper, we only report on the multi-word approach to cross-word pronunciation variation modeling.

The multi-words were obtained 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 [2]. The cross-word variation was tested in isolation and the combination of within-word variation and cross-word variation was also tested.

3. CSR AND SPEECH MATERIAL

The main characteristics of the CSR are described in [2, 4]. Our training and test material, selected from the VIOS database [5], consisted of 25,104 utterances (81,090 words) and 6,267 utterances (21,106 words), respectively. A third dataset, consisting of 6,245 utterances (18,371 words) from the VIOS database, was used for error analysis. The use of this corpus (further referred to as the error analysis corpus) was necessary because if we were to carry out detailed error analysis on the test corpus, its validity as an independent test set would be scrutinized. 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+sub/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).

	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**

Table 2: WER (%) for the various test conditions for different variants: within-word variants, cross-word variants, and the combination of within-word with cross-word variants. Significant improvements, compared to SSS, are shown in bold. *Multi-words added to lexicon and language model.

Summarizing, the best results are obtained when a combination of cross-word and within-word 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). For a more detailed discussion of these results see [2].

4.2. Differences in recognition results due to modeling pronunciation variation

As error rates do not give a complete picture of the effect a method has, 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. For the moment we have restricted this analysis to the utterance level, mainly for practical reasons. 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 and deteriorations are found for 4.3% and 2.9% of the utterances, respectively.

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

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

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 consists 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 is different for 924 utterances (474 + 267 + 183). These differences are a consequence of modeling pronunciation variation, but they cannot be seen in the WER. The WER only reflects the net result obtained, and this error

analysis shows that this is only a fraction of what actually happens due to applying our methods.

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%)

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

4.3. Isolation versus combination?

Figure 1 shows the differences in WER between the results of adding variants for 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. As the principle of superposition clearly does not apply here, 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.

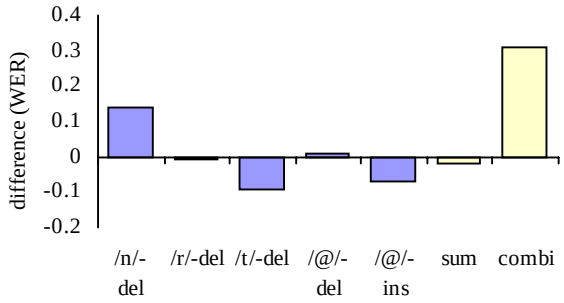

Figure 1: Difference in WER between SSS and MSS for each of the rules in isolation, sum of those results, and combination result of all rules (test corpus).

4.4. Error analysis

The set of tests described in the previous section was repeated using the error analysis corpus. 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 small in Figure 2.

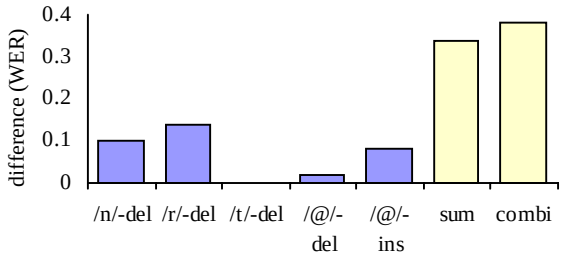

Figure 2: Difference in WER between SSS and MSS for each of the rules in isolation, sum of those results and combination result of all rules (error analysis corpus).

Above, and also in [2], 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 rules/variants.
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.

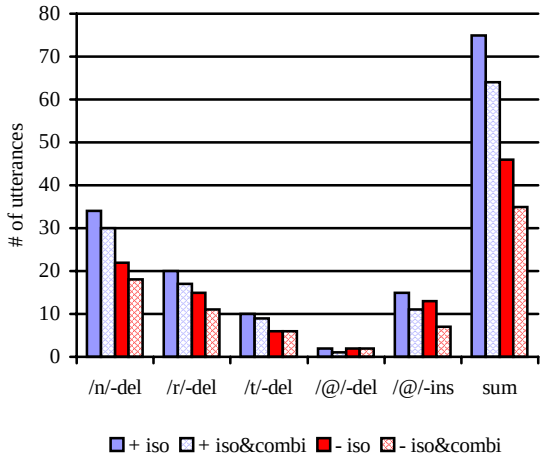

Figure 3: Number of improved (+) and deteriorated (-) *utterances* due to adding variants of the individual rules to the lexicon in isolation (iso). Iso&combi indicates how many of those improvements/deteriorations are also present in the combination test, and the last set of bars indicates the sum of all these results.

Figure 3 shows the number of utterances that are recognized differently compared to the *baseline* due to adding variants of one of the rules to the lexicon. It also shows how many of those differences are also found in the combination test and the sum of all these results. Table 5 shows the overall differences between the isolation and combination tests. Column 3 shows the overlap in the results of the two test conditions, column 4 shows the number of improvements/deteriorations only found in

isolation, column 5 shows these results for the combination test and in the last column the totals are shown.

					total
improvements	isolation	64	11	-	75
	combination		-	5	69
deteriorations	isolation	35	11	-	46
	combination		-	7	42

Table 5: Number of improvements and deteriorations that are found in both the isolation and combination tests, and solely in the isolation or the combination test.

This analysis shows that 93% (64/69) of improvements are found in both the combination and the isolation test conditions and 83% (35/42) of deteriorations. The utterances in columns 4 and 5 of Table 5 were inspected to see if the three points mentioned above could explain the differences between isolation and combination. Point 1, more than one rule applying to a word, explains one of the improvements and five of the deteriorations in the combination case. Of these utterances, two could also be explained by point 2. As to why the deteriorations and improvements in isolation do not occur in combination this can also be explained by point 2 in eleven of the cases, i.e. other rules are present in the combination test condition, which prevent the deteriorations/improvements from occurring. Also, two of the cases can be accounted for by point 1. It is unclear what caused the remaining improvements and deteriorations. They should probably be attributed to point 3, although this can not be verified by examining the output of the decoder.

The example in Table 6 shows the original transcription of an utterance (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, "Delft" is recognized incorrectly as terug't@rYx', whereas, when the combination lexicon is used the whole utterance is recognized correctly. This is 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" which may also influence the result.

	Orthography'phone transcription'
Original	delft'dELft' naar'na:R' amsterdam'Amst@RdAm'
Baseline	*terug't@rYx'* naar'na:R' amsterdam'Amst@RdAm'
Combi	delft'dEl@f' naar'na:R' amsterdam'Amst@dAm'

Table 6: Example of how the combination of rules leads to correct recognition.

5. 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 and cross-word variation are modeled in combination, 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%, which corresponds to 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 that 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 were recognized differently, whereas a net improvement of only 1.3% was found. A lot of the differences in recognition results are not visible in the error rates. Therefore, it can be concluded 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, are also found in the results of the combination of various rules. This indicates that rules in isolation can to some extent indicate what will happen in combination, however how this can facilitate deciding which variation to model is still an unanswered question.

6. 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 Dr. Helmer Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

7. REFERENCES

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

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

3. 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.

4. 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.

5. 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.