MODELING PRONUNCIATION VARIATION FOR A DUTCH CSR: TESTING THREE METHODS

Mirjam Wester, Judith M. Kessens & Helmer Strik

A²RT, Dept. of Language & Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
{wester, 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 pronunciation variation. We used three methods to model pronunciation variation. First, within-word variation was dealt with. Phonological rules were applied to the words in the lexicon, thus automatically generating pronunciation variants. Secondly, cross-word pronunciation variation was modeled using two different approaches. The first approach was to model cross-word processes by adding the variants as separate words to the lexicon and in the second approach this was done by using multi-words. For each of the methods, recognition experiments were carried out. A significant improvement was found for modeling within-word variation. Furthermore, modeling cross-word processes using multi-words leads to significantly better results than modeling them using separate words in the lexicon.

1. INTRODUCTION

The work reported on here concerns the Continuous Speech Recognition (CSR) component of a Spoken Dialogue System called OVIS [1]. 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 data base called VIOS. The speech material consists of interactions between man and machine. The data clearly show that the manner in which people speak to OVIS varies, ranging from using very sloppy articulation to hyperarticulation. As pronunciation variation - if it is not properly accounted for - degrades the performance of the CSR, solutions must be found to deal with this problem.

In this paper, we make a distinction between two types of pronunciation variation: within-word and cross-word variation. We model within-word variation with a rule-based method using phonological rules. In previous experiments [2], we modeled cross-word variation by treating frequently occurring sequences of words as separate entities, i.e. multi-words. The type of cross-word processes which we modeled were cliticization, reduction and contraction, like for example: “ik_wil” which has the variants /IkwIL/ and /kwIL/ and “het_is” with the variants /hEtIs/, /@tIs/ and /tIs/ (transcriptions throughout this paper are in SAMPA).

A substantial improvement was obtained with the rule-based method [2]. However, the results were less clear for the multi-word method:

- Adding multi-words to the lexicon leads to a significant improvement.
- Adding pronunciation variants of these multi-words leads to a deterioration in word error rates (WERs).
- Adding probabilities of the variants to the language model (LM) alleviates the deterioration and even leads to a significant improvement compared to the baseline system.

By combining both methods a total relative reduction of 8.5% in the WER was found [2]. However, because the multi-word method was only studied in combination with the rule-based method it was difficult to interpret the results for the multi-word method. Therefore, we decided to conduct a new set of tests in which the multi-word method was studied in isolation. In addition we also studied a second method for modeling cross-word variation, in which cross-word variants are included as separate entities in the lexicon (/hEt/, /@t/ and /t/ for the word “het”, /Ik/ and /k/ for the word “ik”, etc.). Both cross-word methods and the within-word method are studied in isolation, by applying them to the baseline system.

In [2] the main criteria for selecting multi-words was frequency. As a consequence, some of the multi-words did not have any cross-word variants but only one canonical variant. In our new tests we only included the multi-words in which cross-word variation can occur. Furthermore, since we want to compare the two cross-word methods, the same cross-word processes are modeled in both methods.

When pronunciation variants are included in the lexicon there are two options for the LM: (1) do not include the variants in the LM, and (2) calculate probabilities for the variants and include them in the LM. For each method we will compare the results for both types of LMs.

The aim of this paper is to determine what the effect of modeling cross-word variation is and how it best can be modeled. To this end the two cross-word methods will be compared with each other and with the within-word method. Furthermore, the effect of adding pronunciation variants to the LM will be analyzed for the three methods being studied.

In section 2, the methods we used for modeling pronunciation variation are described. Subsequently, in section 3, the results obtained with these methods are presented. Finally, in the last section, we discuss the results and their implications.

2. METHOD AND MATERIAL

2.1. Method

In this section, we first describe our baseline lexicon followed by an explanation of the general method for modeling pronunciation variation. Next, an explanation is given of the manner in which the general method is used for modeling within-word variation and cross-word variation.

2.1.1. Baseline

For our baseline system, we used a CSR with an automatically generated lexicon. This lexicon contains one transcription for each word. The transcriptions were obtained using the Text-to-Speech system developed at the University of Nijmegen [3]. In this way, transcriptions of new words are easily obtained automatically and consistency in transcriptions is achieved.

2.1.2. Lexicon expansion

In all three methods, pronunciation variants are added to the baseline lexicon, resulting in a lexicon with multiple pronunciation variants. This lexicon can be used either during recognition or training, or during both. In short, the whole procedure for training is as follows:

1. Train the first version of phone models using a canonical lexicon.
2. Generate a multiple-pronunciation lexicon.
3. Use forced recognition to improve the transcription of the training corpus.
4. Train new phone models using the improved transcriptions.

In step 3, forced recognition is used to determine which pronunciation variants are realized in the training corpus. Forced recognition involves "forcing" the recognizer to choose between variants of a word, instead of between different words. In this way, an improved transcription of the training corpus is obtained, which is used to train new phone models. The improved transcription of the training corpus is also used to calculate probabilities of variants, i.e. to add the variants to the LM.

2.1.3. Within-word variation

The pronunciation variants were automatically generated by applying a set of phonological rules of Dutch to the words in the baseline lexicon. The rules were applied to all words in the lexicon where possible, using a script in which rules and conditions were specified. All variants generated by the script were added to the baseline lexicon thus creating a multiple-pronunciation lexicon.

We modeled within-word variation using five phonological rules: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion. These rules were chosen according to four criteria. The rules had to be rules of word-phonology, they had to concern insertions and deletions, they had to be frequently applied, and they had to regard phones that are relatively frequent in Dutch. A more detailed description of the phonological rules and the criteria for choosing them can be found in [4].

2.1.4. Cross-word variation 1

The multi-words from the previous experiments [2] were used as a starting point to choose which variants to add to the lexicon. From those multi-words, only those words which were affected by cross-word processes were selected. This led to the following seven words being chosen (with their various transcriptions between brackets): "ik" (/Ik/, /k/), "het" (/hEt/, /@t/, /t/) "is" (/Is/, /s/), "dit" (/dIt/, /dI/) "dat" (/dAt/, /dA/), "niet" (/ni:t/, /ni:/), and "de" /d@/, /d/). These words make up 9% of all the words in the training corpus.

If these variants of the seven words (together with their very short transcriptions) are added to the lexicon it is likely that the confusability will increase. Since these cross-word variants are entered as separate entities in the lexicon, there is also no restriction as to where the variant can occur. Consequently, it is possible that simple adding these variants deteriorates the performance of the CSR. Especially in this case calculating probabilities for the variants (i.e. adding them to the LM) could prove to be important, as it should restrict the contexts in which a cross-word variant can occur and thus reduce part of the introduced confusability. Furthermore, there may be reasons to assume that certain pronunciation variants will occur in succession in the course of one utterance. For instance, if the speaking rate is high, it can be expected that it will be high during the whole utterance. Including the variants in the LM is a way of modeling this effect.

2.1.5. Cross-word variation 2

In this approach multi-words are added to the lexicon. In order to be able to compare this method with the previous one the same cross-word processes as listed in section 2.1.4. were modeled. On the basis of the 7 words selected in the former approach, 22 multi-words were chosen. 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/). All selected multi-words had at least two variants. These 22 multi-words make up 7% of the training material.

2.2. CSR and Material

The main characteristics of the CSR are described in [1, 2, 4]. Our training and test material, selected from VIOS, consisted of 25,104 utterances (81,090 words) and 6267 utterances (21,106 words), respectively. Recordings with high levels of background noise were excluded from the material used for training and testing.

The single-variant training lexicon contains 1412 entries and the single-variant recognition lexicon contains 1158 entries. Adding pronunciation variants generated by the five phonological rules increases the size of the training lexicon to 2729 entries and the recognition lexicon to 2273 entries (an average of about 2 entries per word). The maximum number of variants that occurs for a single word is 16.

The testing corpus does not contain any out-of-vocabulary (OOV) words. This is a somewhat artificial situation, but we did not want the recognition performance to be influenced by words which could never be recognized correctly, simply because they

were not present in the lexicon.

3. RESULTS

Recognition can be carried out with phone models trained on a corpus with a single pronunciation per word (S), or with phone models trained on a corpus with multiple pronunciations (M). In addition, either a single (S) or a multiple (M) pronunciation lexicon can be used during recognition. In the following tables the different conditions are indicated in the row entitled "CSR". The first letter indicates which type of training corpus was used and the second letter denotes what type of lexicon was used during recognition.

3.1. Within-word Variation

Table 1 shows the results of modeling within-word pronunciation variation. In column 2 the WER for the baseline condition (SS) is given. The effect of adding pronunciation variants during recognition can be seen when comparing the SS and SM conditions. Adding pronunciation variants to the lexicon (SM) leads to an improvement of 0.31% in WER. When the multiple-pronunciation lexicon is used to perform a forced recognition and new phone models are trained on the resulting updated training corpus (MM), it leads to a further improvement of 0.28% compared to the SM condition.

CSR	SS	SM	MM	prob-LM
WER(%)	12.75	12.44	12.16	no
WER(%)	-	12.41	12.04	yes

Table 1: WERs for CSRs with within-word variation modeled.

The results in row 2 were obtained without changing the LM. Next, probabilities of the pronunciation variants were added to the LM. By comparing the results of row 3 with those in row 2 the effect of the LM can be seen. For the SM condition the difference is only 0.03%, while for the MM condition an improvement of 0.12% is observed.

3.2. Cross-word Variation 1

Table 2 shows the results of modeling cross-word pronunciation variation by adding the separate parts to the lexicon. A deterioration of 0.25% is found when cross-word variants are added to the lexicon, as was expected. However, if retrained phone models are used, part of the deterioration is eliminated.

CSR	SS	SM	MM	prob-LM
WER(%)	12.75	13.00	12.89	no
WER(%)	-	12.81	12.59	yes

Table 2: WERs for CSRs with cross-word variation 1 modeled.

In the third row, the WERs are given after probabilities of pronunciation variants were added to the LM. This leads to improvements in both the SM and MM conditions, with a larger improvement for the MM condition.

3.3. Cross-word Variation 2

Table 3 shows the results of modeling cross-word pronunciation variation by adding multi-words to the lexicon. When we add multi-words to the lexicon, we also add them to the LM. So even in the case that no variants of the multi-words are used, the LM is changed. This change alone leads to an improvement of 0.34% (compare SS of Table 3 to SS of table 1). Also for this method we find an increases in the WER when variants of cross-word processes are included in the lexicon (SM). The next step, using retrained phone models, barely influences the WER (MM).

CSR	SS	SM	MM	prob-LM
WER(%)	12.41	12.74	12.72	no
WER(%)	-	12.48	12.39	yes

Table 3: WERs for CSRs with cross-word variation 2 modeled.

Once again we added probabilities of the pronunciation variants to the LM. This step seems to have a larger effect here than in the previous two methods. For the SM condition an improvement of 0.26% is found and for the MM condition an improvement of 0.33% WER is found.

3.4. Overall Results for the Three Methods

Fig. 1 shows the effect of adding the probabilities of the different pronunciation variants to the LM. Each bar represents the percentage of improvement in WERs found when comparing the WERs in row 2 and 3, for the SM and MM conditions, in tables 1, 2 and 3. Fig. 1 shows that adding probabilities of pronunciation variants to the LM has a larger effect when retrained phone models are used. Moreover, the effect is larger for cross-word variation than for within-word variation.

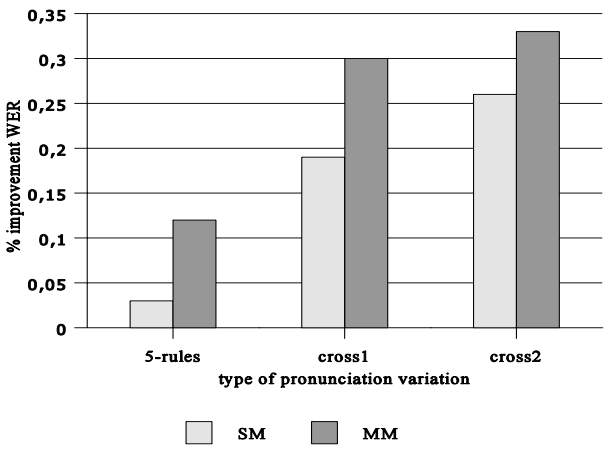

Figure 1: Percentage of improvement in WER caused by adding probabilities of pronunciation variants to the LM for different types of pronunciation variants.

Throughout this paper we have used WER as criterion to measure the performance of the CSR. However, tests for

significance cannot be performed on WERs because the errors (insertions, deletions and substitutions) are not independent of each other. Therefore, to test significance we use sentence error rates (SERs) and a McNemar test [5].

For all three methods the best results are obtained when variants are used during training, recognition and in the LM, i.e. the MM condition with probabilities for the variants. These final results of the three methods and the results for the baseline system are presented in Table 4. For all three methods improvements in WERs and SERs are found compared to the baseline. Tests of significance with the McNemar test and the SERs reveal that a significant improvement is obtained with the rule-based method, the results for the two cross-word methods do not significantly differ from those of the baseline, but cross-word method 2 is significantly better than cross-word method 1.

	baseline	5-rules	cross-word 1	cross-word 2
condition	SS	MM	MM	MM
WER(%)	12.75	12.04	12.59	12.39
SER(%)	21.51	20.60	21.34	20.94

Table 4: SERs for each of the 3 methods

Looking at the absolute percentages it may seem a bit strange, that an absolute improvement of 0.40% in SER when going from cross-word 1 to cross-word 2 is significant, while an improvement of 0.57% in going from baseline to cross-word 2 is not significant. The reason for this is that in the latter case overall much more changes occur, and thus a larger improvement is needed in order for it to be significant. This is simply a property of the McNemar test of significance.

4. DISCUSSION AND CONCLUSIONS

In the within-word method variants were generated by rule. The results show that using these variants during recognition alone reduces the WER. A further improvement is found when using updated phone models and a LM in which probabilities of the variants are incorporated. In total, the WERs improve by 0.71% (a significant improvement of 0.91% in the SER). Therefore, we can conclude that this method works well for improving the performance of our CSR.

In cross-word method 1 cross-word variants were added to the lexicon as separate entities. Only using the variants during recognition increases the WER. Using updated phone models and probabilities of the variants lowers the WER. The net result is a reduction in the WER of 0.16%. The final results for cross-word method 2 are significantly better than those of cross-word method 1. It should be noted however, that most of the improvement for cross-word method 2 is caused by the addition of the multi-words alone (i.e. without using variants of the multi-words). Remember that in this case we also added the multi-words to the LM. In this way the scope of the LM is enlarged and this probably is the main explanation for the improvement in the performance. If in addition to the multi-words also the variants of the multi-words, retrained phone models, and the LM variants are used, only a slight reduction in the WER is observed.

Modeling within-word variation seems to have more effect than modeling cross-word variation. However, in comparing these results one should keep in mind that only very few cross-word processes were modeled, while the phonological rules apply to many words. It is possible that larger improvements for the cross-word methods can be obtained by increasing the number of handled cross-word processes. The best results will probably be obtained when both kinds of methods are combined, one to model within-word variation and the other to model cross-word variation. In [2] we found that the improvement for the combination of methods is larger than that of each separate method.

By studying all three methods in isolation we had the possibility to analyze the effect of adding variants to the LM. We found that the improvement is larger for the cross-word methods than for the within-word method. The deterioration caused by adding variants in the cross-word methods is alleviated by using probabilities of those variants in the LM.

To summarize, the most optimal results are found when lexicon, phone models and LM are balanced in the sense that the same variants are present in all cases. Larger improvements are found for the within-word method than for the cross-word methods, and of the cross-word methods the multi-word approach gave the best results.

5. ACKNOWLEDGMENTS

The research by J.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. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. We thank Lou Boves for thoroughly reviewing an earlier version of this paper.

6. REFERENCES

1. Strik, H., Russel, A., van den Heuvel, H., Cucchiarini, C. and Boves, L. A Spoken Dialogue System for the Dutch Public Transport Information Service *Int. Journal of Speech Technology*, Vol. 2, No. 2: 119-129, 1997.

2. Wester, M., Kessens, J. M., Strik, H. "Improving the Performance of a Dutch CSR by Modeling Pronunciation Variation," *Proc. of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, 145-150, 1998.

3. Kerkhoff, J. and Rietveld, T. "Prosody in Niros with Fonpars and Alfeios," *Proc. Dept. of Language & Speech, University of Nijmegen*, Vol.18: 107-119, 1994.

4. Kessens, J. M. and Wester, M. "Improving Recognition Performance by Modeling Pronunciation Variation," *Proc. of the CLS opening Academic Year '97 '98*: 1-19., 1998.
http://lands.let.kun.nl/literature/kessens.1997.1.html

5. Siegel, S. and Castellan N.J. *Nonparametric Statistics for the Behavioral Sciences*, McGraw Hill, 63-67, 1956.