IMPROVING THE PERFORMANCE OF A DUTCH CSR BY MODELING PRONUNCIATION VARIATION

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 in order 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 accounted for by adding multi-words and their variants to the lexicon. Thirdly, probabilities of pronunciation variants were incorporated in the language model (LM), and thresholds were used to choose which pronunciation variants to add to the LMs. For each of the methods, recognition experiments were carried out. A significant improvement in error rates was measured.

1. INTRODUCTION

The work reported on here concerns the Continuous Speech Recognition (CSR) component of a Spoken Dialogue System (SDS) that is employed to automate part of an existing public transport information service [1]. A large number of telephone calls of the on-line version of the SDS have been recorded. These data clearly show that the manner in which people speak to the SDS varies, ranging from using very sloppy articulation to hyper articulation. 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.

Pronunciation variation can be divided into two main kinds of variation. First, variation in the order and number of phones a word consists of, and second, variation in the acoustic realization of phones. In the present research, we are mainly interested in the first kind of pronunciation variation, because we expect this variation to be more detrimental to speech recognition than the second kind. After all, most of the variation in producing phones should be modeled implicitly when using mixture models.

Our objectives are to improve the performance of the CSR, but also to gain more understanding of the processes which play a role in spontaneous speech. The work reported on in this paper is exploratory research into how pronunciation variation can best be dealt with in CSR.

In section 2, the general method for modeling pronunciation variation is described. It is followed by a detailed description of three different approaches which we used to model pronunciation variation. 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

The approach we use resembles those used previously with success in [2, 3]. Earlier experiments using this method are reported on in [4]. First, our baseline lexicon is described followed by an explanation of the general method for modeling pronunciation variation. Next, an explanation of the manner in which the general method is used for modeling within-word variation (method 1) and cross-word variation (method 2) is given. The last method (method 3), which is an expansion of the general method, describes how probabilities of pronunciation variants were incorporated in the language model (LM).

2.1.1 Baseline

As a baseline we used a CSR with an automatically generated lexicon. This lexicon is a canonical lexicon which means it contains one transcription per word. It is crucial to have a well-described lexicon to start out with. This is especially so in light of pronunciation variation, because the variants chosen for each word in the canonical lexicon have great consequences for the results of the recognition. Since improvements or deteriorations in recognition due to modeling pronunciation variation are measured compared to the result of the baseline system, the choice of this baseline is quite crucial. Furthermore, the pronunciation variants which we generate are based on the canonical transcriptions, therefore the canonical lexicon must be well-defined.

Our lexicon was automatically generated using the Text-to-Speech (TTS) system [5] developed at the University of Nijmegen. Phone transcriptions for the

words in the lexicon were obtained by looking up the transcriptions in two lexica; ONOMASTICA [6], a lexicon with proper names, and CELEX, a lexicon with words from mainly fictional texts. The grapheme-to-phoneme converter is employed whenever a word cannot be found in either of the lexica. There is also the possibility of manually adding words to a user lexicon, if the words do not occur in either of the lexica and are not correctly generated by the grapheme-to-phoneme converter. In this way, transcriptions of new words are easily obtained automatically and consistency in transcriptions is achieved.

2.1.2 Rule-based lexicon expansion

As explained above, our baseline is a canonical lexicon, with one entry per word. Pronunciation variants are added to this lexicon, thus 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. Choose a set of phonological rules.
3. Generate a multiple-pronunciation lexicon using the rules from step 2.
4. Use forced recognition to improve the transcription of the training corpus.
5. Train new phone models using the improved transcriptions.

In step 4, 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.

Steps 4 and 5 can be repeated in iteration in order to gradually improve the transcriptions and the phone models. Steps 2 through to 5 can be repeated for different sets of phonological rules.

2.1.3 Method 1: Within-word variation

Pronunciation variants were automatically generated by applying a set of phonological rules of Dutch to the pronunciations in the canonical 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 canonical lexicon thus creating a multiple-pronunciation lexicon.

In the first set of experiments, we modeled within-word variation using four phonological rules: /n/-deletion, /t/-deletion, /ə/-deletion and /ə/-insertion. In the next set of experiments, we added a fifth rule; the rule for post-vocalic /r/-deletion. 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, 7, 8].

2.1.4 Method 2: Cross-word variation

Cross-word variation was modeled by joining words together with underscores, thus forming new words which we refer to, in this paper, as *multi-words*. This changes the lexica, corpora, and LMs. The multi-words are added to a lexicon in which the separate parts that make up the multi-words are still present. Multi-words are substituted in the corpora wherever the word sequences occur. The LMs are calculated on the basis of these adapted corpora.

We used the following criteria to decide if a word classifies as a multi-word or not. First, the sequence of words had to occur frequently in the training material. We considered a minimum of 20 occurrences of the word sequence in the training material to be adequate. The second criterion which we adopted was that word sequences had to form an articulatory or linguistic unit. Thirdly, when a two part multi-word, for example "ik_wil" is selected, it is no longer possible to create a multi-word consisting of three parts which includes "ik_wil". Thus, the three-part multi-word "ik_wil_graag" is then no longer a possible multi-word.

Experiments were carried out to measure the effect of adding multi-words to the lexicon, and the effect of adding pronunciation variants of multi-words. The pronunciation variants of the multi-words were automatically generated using the five within-word phonological rules mentioned earlier and a number of cross-word phenomena, namely: cliticization, contraction and reduction. The underscores were disregarded during the scoring procedure, so whether the word sequence was recognized as a multi-word or in separate parts had no effect on the word error rates.

2.1.5 Method 3: Probabilities

In previous experiments [4], we found that it is crucial to determine which pronunciation variants should be added to the lexicon. Adding variants to the lexicon can lead to a higher degree of confusability during recognition. Consequently, pronunciation variants not only correct some of the mistakes made, but also introduce new mistakes. Therefore, we started looking for automatic ways to reduce this confusability. First, we incorporated probabilities in the LMs, and second, we applied a threshold to determine which pronunciation variants should be included in both the LMs and the lexicon.

A forced recognition was carried out on a large corpus (see section 2.2) with a lexicon containing 50 multi-words and pronunciation variants. Word counts and counts of

pronunciation variants were made on the basis of the resulting corpus. These counts were used to create new LMs (unigram and bigram). Pronunciation variants were added to the LMs, thus creating new entries. This is in contrast to the earlier described methods 1 and 2, where the pronunciation variants were not incorporated in the LMs, but only in the lexicon.

We assumed that not all words occurred frequently enough in the training material to correctly estimate the probabilities of all variants. Therefore, a number of thresholds were chosen, to find out how often a word must occur in order to correctly estimate the probabilities of the pronunciation variants.

The thresholds (N) are applied to both the LM and the test lexicon. The word count is used to determine if pronunciation variants are included in the LM. If a word occurs N times or more, all pronunciation variants of that word and their counts are included in the LM and the lexicon. If a word occurs less times than the threshold, only the most frequent pronunciation variant is included in the LM and the lexicon.

2.2 CSR and Material

The CSR used in this experiment is part of an SDS [1], as was mentioned earlier. The speech material was collected with an online version of the SDS, which was connected to an ISDN line. The input signals consisted of 8 kHz 8 bit A-law coded samples. The speech can be described as spontaneous or conversational. Recordings with high levels of background noise were excluded from the material used for training and testing.

The most important characteristics of the CSR are as follows. Feature extraction is done every 10 ms for frames with a width of 16 ms. The first step in feature analysis is an FFT analysis to calculate the spectrum. Next, the energy in 14 Mel-scaled filter bands between 350 and 3400 Hz is calculated. The final processing stage is the application of a discrete cosine transformation on the log filterband coefficients. 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 models (HMMs), language models (unigram and bigram), and a lexicon. The continuous density HMMs consist of three segments of two identical states, one of which can be skipped. In total 38 HMMs were used, 35 of these models represent phonemes of Dutch, two represent allophones of the phonemes /l/ and /r/, and one model is used for the non-speech sounds.

For the experiments conducted using methods 1 and 2, our training and test material consisted of 25,104 utterances (81,090 words) and 6267 utterances (21,106 words), respectively. The training material was used to train the HMMs and the LMs. In a later stage, the training corpus was expanded with 49,822 utterances leading to a total of 74,926 utterances (225,775 words). The enlarged training corpus is only used for method 3 to estimate the probabilities of pronunciation variants. In the future, this enlarged corpus will also be used in methods 1 and 2.

The single variant training lexicon contains 1412 entries, which are all the words in the training material. Adding pronunciation variants generated by the five phonological rules increases the size of the lexicon to 2729 entries (an average of about 2 entries per word). Adding 50 multi-words plus their variants leads to a lexicon with 2845 entries. The maximum number of variants that occurs for a single word is 16.

The single variant test lexicon contains 1158 entries, which are all the words in the test corpus, plus a number of words which must be in the lexicon because they are part of the domain of the application. 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. Adding pronunciation variants generated by the five phonological rules leads to a lexicon with 2273 entries (also about 2 entries on average per word). Adding 50 multi-words and their variants results in a lexicon with 2389 entries.

The results presented in the next section are best-sentence word error rates. The word error rate (WER) is determined by :

$$WER = \frac{S+D+I}{N} \qquad \textbf{(1)}$$

where S is the number of substitutions, D the number of deletions, I the number of insertions and N the total number of words. During the scoring procedure only the orthographic representation is used. Whether or not the correct pronunciation variant was recognized is not taken into account.

3. RESULTS

Recognition can be carried out with phone models trained on a corpus with single-pronunciation variants (S), or with phone models trained on a corpus with multiple-pronunciation variants (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 what kind of training corpus was used and the second letter denotes what type of lexicon was used during testing.

3.1 Method 1: Within-word variation

Table 1 shows the results obtained for two rule sets: four and five rules (see 2.1.3). Adding a pronunciation rule, in this case the /r/-deletion rule, gives the same result for the SM condition, but leads to an improvement, 0.32% and 0.31% in WER, for the MS and MM conditions, respectively. Therefore, the rest of the results discussed here concern the CSR with five rules.

Table 1: WERs for different lexica with 4 and 5 rules during training and testing .

CSR	SS	SM	MS	MM
4 rules WER(%)	12.75	12.49	13.14	12.47
5 rules WER(%)	12.75	12.46	12.82	12.16

The effect of adding pronunciation variants during recognition can be seen when comparing the SS and SM conditions. In column 2, the results are shown for the baseline condition (SS). Adding pronunciation variants to the lexicon (resulting in a multiple-pronunciation lexicon, SM) leads to an improvement of 0.29% in WERs.

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 0.30% compared to the condition SM.

Testing with the single-pronunciation lexicon while using updated phone models leads to a slight decrease in WERs compared to the SS condition. It seems the best results are found when the phone models are trained on a corpus which is based on the same lexicon as the lexicon which is used during recognition. (SS is better than MS and MM is better than SM.)

3.2 Method 2: Cross-word variation

On the basis of the criteria explained in section 2.1.4, we selected multi-words which were added to the lexicon. Table 2 shows the effect of adding 25, 50 and 75 multi-words compared to the WER for the case where 0 multi-words have been added to the lexicon (the SS column in Table 1). The first 50 multi-words were as general as possible, no real application specific word sequences were included. The next 25 multi-words which were added to get a total of 75 multi-words were application specific. They consisted of frequently occurring station names. This was necessary because no more than 50 word sequences, which were not application specific, adhered to all the criteria listed in 2.1.4. The station names which we added were of the type “Driebergen-Zeist”, which is simply a station name consisting of two parts.

Table 2: WERs for different numbers of multi-words

# multi	0	25	50	75
WER(%)	12.75	12.43	12.26	12.41

Adding 50 multi-words leads to an improvement of 0.49% in WERs. It seems as if there is a maximum to the number of variants which should be added. On the basis of the results shown in Table 2, we decided to continue using the lexicon containing 50 multi-words, because this gave the largest improvement in WERs.

In the following stage, we added different pronunciation variants to the lexicon containing 50 multi-words. The results are shown in Table 3. The second column shows the result for the condition without pronunciation variants, but with 50 multi-words (see also column 4, Table 2). Next, we added pronunciation variants generated by the five phonological rules (see 2.1.3). First, the rules were only applied to the separate words in the lexicon, not to the multi-words (column 3). The result in column 4 is due to adding only pronunciation variants of the 50 multi-words (see 2.1.4) to the lexicon. In the last column, the result is shown for the situation where all of the pronunciation variants (5 rules and multi) were added to the lexicon.

Table 3: WERs for CSRs with 50 multi-words, and different pronunciation variants

CSR	SS	SM	SM	SM
variants	none	5 rules	multi	all
WER(%)	12.26	11.92	12.77	12.35

Adding variants generated by the five phonological rules (5 rules) gives roughly the same improvement (0.34% compared to 0.29%) as was found in Table 1 when going from SS to SM. When only variants of the multi-words are added (multi), a deterioration of 0.51% in WERs is found. Adding both multi-word variants and the variants generated by the five rules (all) leads to a deterioration in WERs when compared to the SS condition.

3.3 Method 3: Probabilities

Probabilities for separate pronunciation variants were estimated using the enlarged corpus. A forced recognition was carried out on this corpus in order to obtain the pronunciation variants for each word. The lexicon which

was used for the forced recognition contained the 50 multi-words and all of the pronunciation variants (same lexicon as for SM_{all}, last column in Table 3). The probabilities of the pronunciation variants were incorporated in the LMs. Column 2 in Table 4 shows the result of adding probabilities of all pronunciation variants to the LMs. When this is compared to the same test situation, without probabilities (last column, Table 3), an improvement of 0.61% in WERs is achieved.

Table 4: WERs for different thresholds

threshold	0	20	50	100	∞
WER(%)	11.74	11.72	11.70	11.67	11.94

Next, we decided to apply thresholds for adding pronunciation variants to the lexica and LMs as was described in section 2.1.5. We expected that this would also influence recognition, but the improvements proved to be small, as can be seen in columns 3 through 5 in Table 4.

3.4 Overall Results for the 3 Methods

In all of the above results, the effects of adding pronunciation variants can not be seen clearly, because WERs only give an indication of the total improvement or deterioration. Table 5 shows the changes in the utterances, which occur due to the combination of all three methods which were tested. A comparison is made between the baseline condition and the final test (the best condition in Table 4, threshold 100). In the first column (Table 5) the type of change is given, in the second column the number of utterances which are affected.

Table 5: Type of change in utterances going from baseline to final test

type of change	number of utterances
same utterance different mistake	480
improvements	248
deteriorations	147
net result	+101

In total 875 of the 6276 utterances changed. The net result is improvements in 101 utterances, as Table 5 shows, but that is only part of what actually happens due to applying the three methods. For instance, in 480 cases the mistakes made in the utterances change. Although they remain incorrect, the mistakes which are made are different, so pronunciation modeling has an effect here which can not be seen in the WERs.

A significant improvement of 1.58% in sentence error rates (SERs) is found (McNemar test for significance [9]) when going from the baseline condition to the final test. The McNemar test for significance cannot be performed on WERs because the errors (insertions, deletions and substitutions) are not independent of each other. All three methods separately, also show significant improvement for SERs. Table 6 shows the SERs for each of the three methods.

Table 6: SERs for each of the 3 methods

	baseline	method 1	method 2	method 3	
condition	SS	MM	SS	SM_{all}	SM_{all}
multi-word	-	-	50	50	50
prob. LM	-	-	-	-	100
SER(%)	21.51	20.84	20.78	20.57	19.93

Adding variants of five rules, and using updated phone models (method 1), leads to a significant improvement of 0,67% in SERs, when it is compared to the baseline. Adding 50 multi words to the baseline condition (method 2) leads to a significant improvement of 0.73% in SERs. For method 3, a comparison is made between the SM_{all} condition (see column 5 in Table 3) and the condition with a threshold of 100 for the LM. The improvement is 0.64% in SERs, which is also a significant improvement.

4. DISCUSSION AND CONCLUSIONS

The results of method 1, modeling within-word variation, show that adding pronunciation variants generated by applying four phonological rules, reduces the WER. Adding another pronunciation rule, the rule for /r/-deletion also improves recognition performance. A further improvement is found when using updated phone models. This improvement is larger for five rules than for four rules. In total, for method 1, the WERs improve by 0.59% which is a significant improvement of 0.67% in SERs. Therefore, we can conclude that this method works for improving the performance of our CSR. It is important to realize, however, that with each rule that is applied, the variants which are generated will introduce new mistakes in addition to correcting others. In the future, we will look for ways to minimise confusability and to maximise the efficiency of the variants which are added by finding the optimal set of phonological rules.

Method 2 shows that adding multi-words leads to an

improvement of 0.49% in WERs and a significant improvement of 0.73% in SERs. This improvement may be due to the fact that by adding multi-words a type of trigram is created in the LM, only for the most frequent word sequences in the training corpus.

It is unclear why modeling pronunciation variants of multi-words does not lead to an improvement in WERs. The multi-words are all frequent word sequences and we expected that modeling pronunciation variation at that level would have an effect. Furthermore, the pronunciation phenomena which were modeled, i.e. cliticization, reduction processes and contractions are all phenomena which are thought to occur frequently in Dutch [8]. An analysis of the changes which occur due to adding pronunciation variants for multi-words show that the variants correct some errors but also introduce new ones. Other methods might model cross-word variation more effectively. Therefore, we will examine other ways of modeling cross-word variation and we will also attempt to minimize the confusability between variants in the future.

The results of method 3 show an improvement of 0.68% in WERs and a significant improvement of 0.64% in SERs. The steps undertaken in method 3 consisted of adding counts of the pronunciation variants to the LMs and defining a number of thresholds. In the set of experiments, in which probabilities for pronunciation variants were included in the LM, they were included in both the unigram and the bigram. An alternative to this method is to keep the bigram intact and to add the information about frequency of pronunciation variants to the unigram only.

The question is whether or not information about pronunciation variants should be modeled in the bigram. In some cases, there may be reasons to assume that certain pronunciation variants will follow up each other 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. The exact relationships between different pronunciation variants are currently, however, not well understood, and in addition to that, methods to decide when those relationships occur are also not available. So, it may not be optimal to model pronunciation variation at word level in the bigram. In the future, we will experiment with modeling the unigrams independently of the bigrams to find out if they should be modeled separately or together.

In our experiments we found a relative improvement of 8.5% WER (1.08% WER absolute) when going from our baseline condition to the condition in which a lexicon containing multi-words and pronunciation variants was used, and an LM with probabilities of pronunciation variants was used. Our results show that all three methods lead to significant improvements. We found an overall, significant improvement of 1.58% in SERs. These results are very promising and we will continue to seek ways to elaborate on this research in order to understand the processes which play a role to a fuller extent and to gain further degrees of improvement in the performance of the CSR.

5. ACKNOWLEDGMENTS

This work was funded by the Netherlands Organisation for Scientific Research (NWO) as part of the NWO Priority Programme Language and Speech Technology. The research of Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

6. REFERENCES

[1] H. Strik, A. Russel, H. Van den Heuvel, C. Cucchiarini & L. Boves (1997) A spoken dialogue system for the Dutch public transport information service *Int. Journal of Speech Technology*, Vol. 2, No. 2, pp. 119-129.

[2] M. H. Cohen (1989) *Phonological Structures for Speech Recognition*. Ph.D. dissertation, University of California, Berkeley.

[3] L. F. Lamel & G. Adda (1996) On designing pronunciation lexica for large vocabulary, continuous speech recognition. *Proc. of ICSLP '96, Philadelphia*, pp 6-9.

[4] J. M. Kessens, M. Wester (1997) Improving Recognition Performance by Modeling Pronunciation Variation. *Proc. of the CLS opening Academic Year '97 '98*, pp. 1-19. http://lands.let.kun.nl/literature/kessens.1997.1.html

[5] J. Kerkhoff & T. Rietveld (1994) Prosody in Niros with Fonpars and Alfeios, *Proc. Dept. of Language & Speech, University of Nijmegen*, Vol.18 pp. 107-119.

[6] Onomastica http://www2.echo.lu/langeng/en/lre1/onomas.html

[7] C. Cucchiarini & H. van den Heuvel (1995) /r/ deletion in Standard Dutch, *Proc. of the Dept. of Language & Speech, University of Nijmegen*, Vol. 19, pp. 59-65.

[8] G. Booij (1995) *The Phonology of Dutch* Oxford: Clarendon press.

[9] S. Siegel & N.J. Castellan (1956) *Nonparametric Statistics for the Behavioral Sciences*, McGraw Hill, pp.63-67.