MODELING WITHIN-WORD AND CROSS-WORD PRONUNCIATION VARIATION TO IMPROVE THE PERFORMANCE OF A DUTCH CSR

Judith M. Kessens, Mirjam Wester & Helmer Strik

A²RT, Dept. of Language & Speech, University of Nijmegen, The Netherlands
{J.Kessens, M.Wester, W.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. Within-word variants were automatically generated by applying five phonological rules to the words in the lexicon. For the within-word method, a significant improvement is found compared to the baseline. Cross-word pronunciation variation was modeled using two different methods: 1) adding cross-word variants directly to the lexicon, 2) only adding multi-words and their variants to the lexicon. Overall, cross-word method 2 leads to better results than cross-word method 1. The best results were obtained when cross-word method 2 was combined with the within-word method: a relative improvement of 8.8% WER was found compared to the baseline.

1. INTRODUCTION

The present research concerns the continuous speech recognition (CSR) component of a spoken dialogue system named OVIS 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 OVIS have been recorded. These data clearly show that the manner in which people speak to OVIS varies, ranging from using hypo-articulated speech to hyper-articulated speech. 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. We expect that by explicitly modeling pronunciation variation some of the errors introduced by the various ways in which people address the system will be corrected.

Our CSR consists of three components, which means there are three levels at which pronunciation variation can be modeled: the lexicon, the phone models (PMs) and the language model (LM). In our experiments, we test at all three levels. In this way, four testing conditions are obtained which are shown in Table 1. "S" denotes the use of single pronunciations, "M" denotes the use of multiple pronunciations.

	test condition	Lex	PMs	LM
baseline	SSS	S	S	S
step 1	MSS	M	S	S
step 2	MMS	M	M	S
step 3	MMM	M	M	M

Table 1. Test conditions

As most speech recognizers make use of a lexicon, a much used approach to modeling pronunciation variation has been to model it at the level of the lexicon. This can be done by using rules to generate variants which are then added to the lexicon [2]. In our

research we also adopted this approach. On the basis of five frequently occurring Dutch phonological processes, we formulated rules with which to model within-word pronunciation variation. The results of these experiments were promising. Since our ultimate aim is to find the optimal set of rules to model pronunciation variation, we also tested each rule in isolation to find out if the results obtained for rules in isolation can predict how rules will behave in combination. This issue is quite important, as it is at the core of how to proceed with a rule-based approach.

Besides modeling pronunciation variation at the lexical level it can also be incorporated in the PMs. In [3], we found that when including pronunciation variation in the lexicon, it is generally best to retrain the PMs too. This is done by automatically transcribing the training corpus using the CSR in foced recognition mode. Earlier experiments showed that the CSR performs comparable to expert listeners in selecting the appropriate pronunciation variant [4]. Therefore, we expect that the match between the new transcriptions and the acoustic signal will become better and consequently, that the PMs trained on these new transcriptions will be better. This process can be repeated in iteration to obtain even better PMs until no changes occur. In this paper, we investigated how many changes occur in the transcriptions of the training corpus as a result of each retraining and how often this process of retraining PMs should be carried out to obtain optimal results.

In [3], we found that modeling within-word pronunciation variation improves the CSR's performance. However in continuous speech, there is a lot of variation which occurs over word boundaries. In [3] we also showed that adding multi-words (i.e. sequences of words) and their variants to the lexicon is beneficial to recognition performance. Therefore, we decided to retain this approach in the current research. In addition, we tested a second method for modeling cross-word variation. For this method, multi-words were not used, but the separate parts of the multi-words and their variants were added to the lexicon.

For the cross-word methods, we measured the effect of interaction between cross-word variants and within-word variants by testing both methods in isolation and in combination with the within-word method. Here, once again the question is whether the sum of the effects of the methods tested in isolation is the same as the total effect of testing the combination of the methods.

2. METHOD AND MATERIAL

2.1. Method

In this section, we first describe the baseline system (section 2.1.1). In section 2.1.2, this is followed by an explanation of the general test procedure. Next, an explanation is given of how within-word (section 2.1.3) and cross-word variation (section 2.1.4 and section 2.1.5) were modeled. In section 2.2, more details about the CSR and

the speech material that we used are given.

2.1.1. Baseline. The baseline lexicon contains one transcription for each word. These transcriptions were automatically obtained using a Text-to-Speech system (TTS) for Dutch [5]. The baseline PMs were trained using the training corpus and the corresponding transcriptions in the baseline lexicon. The LM was calculated on the basis of the orthographic representation of words in the training corpus. The baseline performance was measured by performing a recognition test using the baseline lexicon, PMs and LM.

2.1.2. General Procedure. As mentioned in the introduction, our general test procedure consists of three steps. Table 1 shows that in each step pronunciation variation is incorporated in another level. Thus, it is possible to test our methods at all three levels. In short the whole procedure is as follows:

* *In step 1,* pronunciation variation is incorporated in the lexicon by adding variants to the baseline lexicon. In this way, a multiple pronunciation lexicon is obtained. Since the baseline LM is calculated on the basis of the orthographic representation of the words, the a priori probabilities (given by the language model) of all variants of a word are equal.

* *In step 2,* pronunciation variation is included in the training of the PMs. A forced recognition is carried out on the training corpus using the multiple pronunciation lexicon from step 1. In this type of recognition the CSR is "forced" to choose between different pronunciation variants of a word instead of between all the words in the lexicon. In this way, an updated transcription of the training corpus is obtained, which is used to train new PMs.

* *In step 3,* pronunciation variation is included in the LM. The updated training corpus obtained in step 2 is used to generate a "multiple" LM. In this LM, different variants will generally have different a priori probabilities.

2.1.3. Within-word Method. To test the within-word method, we generated variants automatically by applying a set of Dutch phonological rules to the words in the baseline lexicon. The selected rules were: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion (SAMPA notation is used throughout this paper). A more detailed description of the phonological rules and the criteria for choosing them can be found in [6].

Each of the steps of the general test procedure were carried out, with the multiple pronunciation lexicon obtained using all five rules. Steps 2 and 3 were carried out four times. We also tested each of the rules in isolation by adding the variants for each separate rule to the lexicon and carrying out a recognition test (MSS).

2.1.4. Cross-word Method 1. Variants for cross-word method 1 were obtained as follows. First, the 50 most frequently occurring word sequences were selected from the training material. Next, those words sensitive to the cross-word processes; cliticization, contraction and reduction were chosen. This led to the following seven words being selected (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.

The main disadvantage of this is that there are no restrictions to the context in which the variants can occur. One would expect these variants to occur only in specific contexts whereas here the context is totally unrestricted. Therefore, especially for this method, the multiple LM could prove to be important. By using the multiple LM, the contexts in which a cross-word variant can occur are restricted. Thus, errors which are introduced by the increased confusability due to adding variants could be corrected by using the multiple LM. Besides unrestricted context forming a possible problem, a second disadvantage may be the length of the pronunciation variants which are added. Some of the variants are extremely short, for instance /k/, /t/, /d/ and /s/ consist of only one phone, and therefore they may easily be inserted.

2.1.5. Cross-word Method 2. The second method, which we adopted for modeling cross-word variation was to make use of multi-words. Multi-words are word-sequences which 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/). In order to be able to compare the results of this method to the results of the previous one, the same cross-word processes were modeled in both methods. On the basis of the seven words from cross-word method 1, multi-words were selected from the list of 50 word sequences. Only those word sequences in which at least one of the seven words was present was chosen. Thus, 22 multi-words were selected

Counting the parts of the multi-words as separate words, the total number of words which could have a pronunciation variant covers 6% of the total number of words in the training corpus. This percentage is lower than for cross-word method 1 due to the contextual constraints imposed by the multi-words.

2.2. CSR and Material
The main characteristics of the CSR are described in [1]. The speech material was selected from a database named VIOS, which contains a large number of telephone calls recorded with the on-line version of OVIS [1]. The training and test material consisted of 25,104 utterances (81,090 words) and 6,267 utterances (21,106 words), respectively.

The baseline training lexicon contains 1412 entries and the baseline recognition lexicon contains 1154 entries. In Table 2, the number of variants which were added to the lexica are given. The maximal number of variants that occurred for a single word was 16.

	within	cross1	Cross2
recognition	1119	8	28
training	1317	8	28

Table 2. Number of variants added to the different lexica.

Table 3 shows the number of variants which were added to the recognition lexica used to test each of the rules separately. The total number of added variants for which a single rule applied is 841. This means that 278 extra variants (in the within-word multiple pronunciation lexicon) were the result of applying at least two different rules to one word in the baseline lexicon.

	/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins
recognition	348	306	83	28	76

Table 3. Number of variants added to each lexicon to test each rule separately.

3. RESULTS

In section 3.1, results will be presented for the within-word method. Next, in section 3.2, the results of both methods of modeling cross-word variation are given, tested in isolation and in combination with the within-word method.

3.1. Results for the Within-word Method

3.1.1. Application of Each Rule. For the within-word method, we repeated steps 2 and 3 of the general procedure four times in iteration. To find out how many times a rule was applied in the training material, a forced recognition was performed on the training corpus. For each rule, we calculated the number of times the rule was applied in the training corpus. A rule is applied whenever a variant is recognized in which an /n/, /t/, /r/, or /@/ is deleted in case of the deletion-rules, or a /@/ is inserted in case of the /@/-insertion rule. This number is divided by the total number of times the conditions for rule application were met in the training corpus. In Figure 1, the percentage application is shown for the different transcriptions of the training corpus. "Iteration 0" means no forced recognition was performed and the training corpus was analyzed using the baseline transcriptions. "Iteration x" means that a forced recognition was performed to update the transcriptions, and the PMs which were used were retrained "x" times.
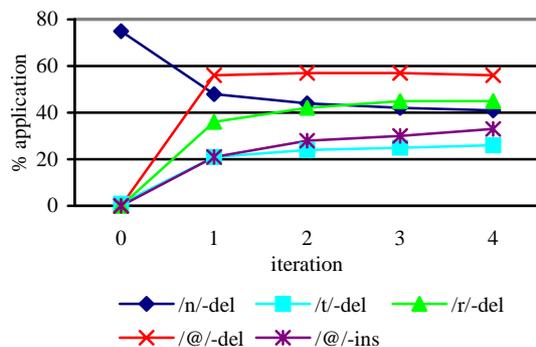

Figure 1. Percentage application for each phonological rule

In Figure 1, it can be seen that only the /n/-deletion rule was applied in the baseline system. This is because in the TTS system we used to generate the transcriptions, the /n/-deletion rule is applied in about 75% of the cases. The number of times the /n/-deletion rule is applied in the speech material decreases, when going from iteration 0 to iteration 4. For all rules, the changes in percentage application are largest when going from iteration 0 to iteration 1. For the other cases only very small changes in percentage application occur.

Even if no changes in percentage application are observed, it is still possible that different variants are chosen in each forced recognition. To investigate this, we counted the percentage of times a rule applied whereas it was not in the previous iteration, and visa versa. We found that for the fourth iteration both percentages were smaller than 5% for all rules.

Since for all rules the changes in percentage application are largest when going from iteration 0 to iteration 1, steps 2 and 3 of the general procedure are only performed once in the rest of this paper.

3.1.2. WERs for Different Test Conditions. In Table 4, the Word Error Rates (WER=100*(S+D+I)/N) are given for the different test conditions for the within-word method. It can be seen that stepwise incorporating pronunciation variation in the lexicon, PMs and LM improves the recognition performance. In total, a significant improvement of 0.68% WER was found for test condition MMM compared to the baseline (SSS).

SSS	MSS	MMS	MMM
12.75	12.44	12.22	12.07

Table 4. WERs for the within-word method.

3.1.3. Testing Rules in Isolation. Figure 2 shows the differences in WER for test condition MSS compared to the baseline condition (SSS), for each rule tested in isolation (gray bars). "Sum" denotes the sum of the results of the five rules tested in isolation (black bar) and "combi" denotes the results for all five rules tested in combination (white bar).
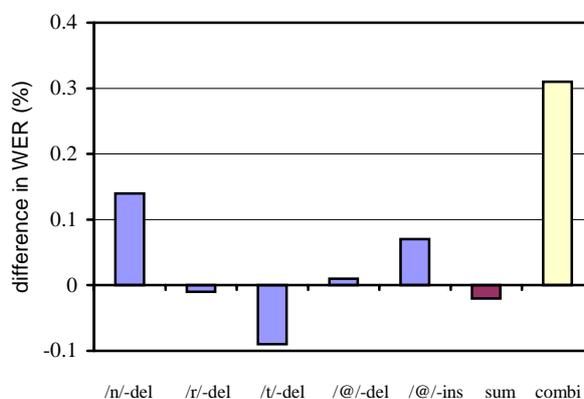

Figure 2. Difference in WER for test condition MSS compared to SSS for each rule tested in isolation, the sum of the results, and the improvement gained when testing all rules together.

It can be seen in Figure 2, that the sum of the differences in WER for the rules tested in isolation is not equal to the improvement obtained when testing the combination of rules. When testing variants of different rules together there is interaction between variants of different rules, and this interaction can cause either improvements or deteriorations.

In section 2.2, we mentioned that 278 variants are present in the within-word multiple pronunciation lexicon, whereas they are not in any of the lexica which were used to test each rule separately. These are exactly the variants to which two (or more) different rules applied. Such variants can cause improvements, for example if the realized pronunciation resembles the variant to which two different rules applied more closely, than that it resembles the variants to which only one rule applied. On the other hand, these variants can be confused with other words/combination of words which were already present in the lexicon, and this results in deteriorations.

3.2. Results for Cross-word Method 1 and 2

In Table 5, the WERs are given for the three different test-conditions (see section 2.1.2) for cross-word method 1 and 2, both tested in isolation and in combination with the within-word method.

	MSS	MMS	MMM
cross1	13.00	12.89	12.59
cross2	12.74	12.99	12.45
within + cross1	12.70	12.58	12.14
within + cross2	12.37	12.30	11.63

Table 5. WERs for the cross-word methods tested in isolation, and in combination with the within-word method.

Table 5 shows that overall cross-word method 2 gives better results than cross-word method 1 both in isolation and in combination with the within-word method. In combination with the within-word method, cross-word method 1 performs even worse than the within-word method in isolation (compare "within+cross1" in Table 5 to Table 4).

Cross-word method 1 in combination with the within-word method gives an improvement of 0.34% compared to the within-word method alone (compare "within+cross2" Table 5 to Table 4, MMM). However, most of the improvement is due to adding the 22 multi-words to the lexicon and LM. Only adding multi-words to the lexicon and LM gives an improvement ranging from 0.29% to 0.41% for the different test conditions.

For cross-word method 2, WER increases when retrained phone models are used. This is the case for cross-word method 2 tested in isolation (compare MMS to MSS, "cross2" in Table 5) and tested in combination (compare MMS, "within+cross2" in Table 5 to MMS in Table 4). A possible cause for this deterioration in performance could be that the PMs were not retrained properly. During forced recognition, the option for recognizing a pause between the separate parts of the multi-words was not given. As a consequence, if a pause occurred in the acoustic signal of a multi-word, the pause will have been used to train the surrounding phone models, which results in contaminated phone models.

To investigate this hypothesis, we repeated the forced recognition for the combination of cross-word method 2 and the within-word method and let the CSR decide whether a pause was present within a multi-word or not. We then trained new phone models on the basis of the output of this forced recognition and repeated test MMS. The WER for this new test was 12.19%. Since, this is an improvement compared to test condition MMS for the within-word method (Table 4), our hypothesis proves to be correct.

4. DISCUSSION

It is clear that testing a combination of methods leads to different results than when methods are tested in isolation. This is the case for testing the combinations of the cross-word methods with the within-word method, and testing them in isolation. The results for the within-word method show the difference which exists between testing methods in isolation or in combination even more clearly. The sum of the results for separate rules led to a degradation in WER (compared to the baseline) whereas the combination led to an improvement. This is mainly due to the fact that variants are not recognized independently of each other the during the recognition process, i.e. interaction takes place between pronunciation variants. This interaction takes place at different levels: within words (e.g. two different rules apply to a word), within a whole utterance (e.g. variants of two different methods are contained in a possible hypotheses for an utterance), within the lexicon (e.g. confusability between different variants), etc. These findings implicate that it will not suffice to study methods in isolation. Instead, they will have to be studied in combination. However, this poses a practical problem as there are many possible combinations.

5.CONCLUSIONS

The percentage application of each rule as a function of the number of iterations behaves as expected. Since earlier experiments showed that the CSR performs comparable to expert listeners in selecting the appropriate pronunciation variant [6], we can conclude that iteration of step 2 and 3 of the general method works well. Furthermore, since for all rules the changes in percentage application are largest when going from iteration 0 to iteration 1, we can conclude that it is usually only necessary to iterate once.

Modeling pronunciation variation in the lexicon, the PMs and the LM, gives a total significant improvement of 0.68% for the within-word method. Overall, cross-word method 2 leads to better results than cross-word method 1, both when tested in isolation and in combination with the within-word method. The best results were obtained when cross-word method 2 was combined with the within-word method: a relative improvement of 8.8% WER was found compared to the baseline.

Finally, it is clear that the principle of superposition does not apply to testing different methods for modeling pronunciation variation. This poses a problem as how to test the different methods, as it is practically impossible to test all combinations. Therefore, we are looking for other solutions to this problem.

REFERENCES

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

[2] Lamel, L. F. and Adda, G. 1996. On Designing Pronunciation Lexica for Large Vocabulary Continuous Speech Recognition, *Proc. of ICSLP-96*, Philadelphia, pp. 6-9.

[3] Wester, M., Kessens, J. M., Strik, H. 1998. 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.

[4] 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 International Conference on Spoken Language Processing, Sydney*, Australia.

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

[6] Kessens, J. M. and Wester, M. 1998. Improving Recognition Performance by Modeling Pronunciation Variation," *Proc. of the CLS opening Academic Year '97 '98*: 1-19.