ELSEVIER

Speech Communication 29 (1999) 193–207

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation

Judith M. Kessens *, Mirjam Wester, Helmer Strik

A² RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Received 22 December 1998; received in revised form 2 August 1999; accepted 4 August 1999

Abstract

This article describes how the performance of a Dutch continuous speech recognizer was improved by modeling pronunciation variation. We propose a general procedure for modeling pronunciation variation. In short, it consists of adding pronunciation variants to the lexicon, retraining phone models and using language models to which the pronunciation variants have been added. First, within-word pronunciation variants were generated by applying a set of five optional phonological rules to the words in the baseline lexicon. Next, a limited number of cross-word processes were modeled, using two different methods. In the first approach, cross-word processes were modeled by directly adding the cross-word variants to the lexicon, and in the second approach this was done by using multi-words. Finally, the combination of the within-word method with the two cross-word methods was tested. The word error rate (WER) measured for the baseline system was 12.75%. Compared to the baseline, a small but statistically significant improvement of 0.68% in WER was measured for the within-word method, whereas both cross-word methods in isolation led to small, non-significant improvements. The combination of the within-word method and cross-word method 2 led to the best result: an absolute improvement of 1.12% in WER was found compared to the baseline, which is a relative improvement of 8.8% in WER.

Zusammenfassung

Dieser Artikel beschreibt, wie die Leistung eines automatischen Spracherkenners, der niederländische gesprochene Sprache erkennt, mit Hilfe der Modellierung von Aussprachevarianten verbessert wurde. Für diese Modellformung wird eine allgemeine Prozedur vorgeschlagen, die – kurz gesagt – darin besteht, dem Lexikon Aussprachevarianten hinzuzufügen, die Phonmodelle erneut einer Lernphase zu unterziehen und Sprachmodelle dabei zu verwenden, in denen die Aussprachevarianten mithineinbezogen wurden. Durch Anwendung einer Gruppe von fünf optionalen phonologischen Regeln wurden im Basislexikon zunächst Aussprachevarianten innerhalb von Wörtern generiert. Dann wurde mit Hilfe zweier Methoden eine begrenzte Anzahl von Sandhiprozessen (Prozesse auf Wordgrenzen) modelliert. Die erste bestand darin, die Sandhivarianten direkt dem Lexikon hinzuzufügen und bei der zweiten wurden Multiwörter gebraucht. Letztendlich wurden die wortinternen Ausprachevarianten mit den zwei Sandhivarianten kombiniert getestet. Die Basisleistung des Spracherkenners, d.h. ohne Anwendung des Modells der Aussprachevariation, betrug 12.75% "word error rate" (WER). Bei Anwendung der wortinternen Aussprachevarianten wurde eine geringe, aber statistisch signifikante Verbesserung von 0.68% WER gemessen. Die Anwendung der zwei Sandhimodelle hingegen ergab einen

* Corresponding author. Tel.: +31(0)24-3612055; fax: +31(0)24-3612907.
E-mail address: j.kessens@let.kun.nl (J.M. Kessens)

0167-6393/99/$ - see front matter
PII: S0167-6393(99)00048-5

sehr kleinen, nicht signifikanten Verbesserung. Die Kombination des wortinternen Modells mit dem zweiten Sandhimodell hingegen ergab schließlich das beste Ergebnis: eine absolute Verbesserung von 1.12% WER, was einer relativen Verbesserung von 8.8% WER entspricht.

Résumé

Cet article décrit comment les performances d'un reconnaisseur de parole continue (CSR) pour le néerlandais ont été améliorées en modelant la variation de prononciation. Nous proposons une procédure générale pour modeler cette variation. En bref, elle consiste à ajouter des variantes de prononciation au lexique et dans le ré-apprentissage des modèles de phones en utilisant des modèles de langage auxquels les variantes de prononciation ont été ajoutées. D'abord, des variantes de prononciation à l'intérieur de mot ont été produites en appliquant un ensemble de cinq règles phonologiques optionnelles aux mots dans le lexique de base. Ensuite, un nombre limité de processus entre-mots ont été modelés, en utilisant deux méthodes différentes. Dans la première approche, des processus entre-mots ont été modelés en ajoutant directement les variantes "entre-mots" au lexique, et dans la deuxième approche ceci a été fait en utilisant des "mots-multiples". En conclusion, la combinaison de la méthode qui se limite aux processus à l'intérieur de mot avec les deux méthodes "entre-mots" a été testée. La performance de base était un taux d'erreur de 12.75% mots (WER); comparée à cette performance de base, une amélioration petite mais significative de 0.68% dans WER a été obtenue avec la méthode 'à l'intérieur de mot', tandis que les deux méthodes d'entre-mots en isolation ont mené à des petites améliorations non significatives. La combinaison de la méthode "à l'intérieur de mot" avec la méthode 2 "entre-mots" a mené au meilleur résultat: une amélioration absolue de 1.12% dans le WER a été trouvée comparée à la ligne de base, qui est une amélioration relative de 8.8% dans le WER.

Keywords: Continuous speech recognition; Modeling pronunciation variation; Within-word variation; Cross-word variation

1. Introduction

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

Since the early seventies, attempts have been made to model pronunciation variation for automatic speech recognition (for an overview see (Strik and Cucchiarini, 1998)). 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 (e.g. Cohen and Mercer, 1974; Cohen, 1989; Lamel and Adda, 1996). In our research, we also adopted this approach. First, we used four phonological rules selected from Booij (1995), which describe frequently occurring within-word pronunciation variation processes (Kessens and Wester, 1997). The results of these preliminary experiments were promising and suggested that this rule-based approach is suitable for modeling pronunciation variation. Therefore, we decided to pursue this approach and for the current research another frequent rule was added: the /r/-deletion rule (Cucchiarini and van

den Heuvel, 1995). Our long-term goal is to find the set of rules which is optimal for modeling pronunciation variation.

Our experiments showed that modeling within-word pronunciation variation in the lexicon improves the CSR's performance. However, in continuous speech there is also a lot of variation which occurs over word boundaries. For modeling cross-word variation, various methods have been tested in the past (see e.g. Cremelie and Martens, 1998; Perennou and Brieussel-Pousse, 1998; Wiseman and Downey, 1998). In our previous research (Kessens and Wester, 1997), we showed that adding multi-words (i.e. sequences of words) and their variants to the lexicon can be beneficial. Therefore, we decided to retain this approach in the current research. However, we also tested a second method for modeling cross-word variation. For this method, we selected from the multi-words the set of words which are sensitive to the cross-word processes that we focus on; cliticization, reduction and contraction (Booij, 1995). Next, the variants of these words are added to the lexicon. In other words, in this approach no multi-words (or their variants) are added to the lexicon.

In this paper, we propose a general procedure for modeling pronunciation variation. This procedure affects all three levels of the CSR at which modeling can take place: i.e. the lexicon, the phone models and the language models (Strik and Cucchiarini, 1998). Table 1 shows at which levels pronunciation variation can be incorporated in the recognition process, and the different test conditions which are used to measure the effect of adding pronunciation variation. In the abbreviations used in Table 1, the first letter indicates which type of recognition lexicon was used; either a lexicon with single (S) or multiple (M) pronunciations per word. The second letter indicates whether single (S) or multiple (M) pronunciations per word were present in the corpus used for training the phone models. The third letter indicates whether the language model was based on words (S) or on the pronunciation variants of the words (M).

The general procedure is employed to test the method for modeling within-word variation, as well as the two methods for modeling cross-word variation. First of all, the three methods were tested in isolation. We were however also interested in the results obtained when combining the different methods. Therefore, we tested a combination of modeling within-word variation together with each of the methods we used to model cross-word variation.

The question which arises here is whether the trends in recognition results measured when testing different methods for modeling pronunciation variation in isolation are the same when testing them in combination. More precisely, the question is whether the sum of the effects of the methods in isolation is (almost) the same as the total effect of the combination of the methods. The answer to this question has implications for our own research and the research on modeling pronunciation variation in general. If there are no differences in results between testing methods in isolation or in combination, it would suffice to test each method in isolation. However, if this is not the case, then all combinations will have to be tested (which poses a large practical problem, because potentially numerous combinations are possible).

This issue is important when combining methods for modeling within-and cross-word variation, but the problem can also exist within one method. Above we already mentioned that our ultimate goal is to find the optimal set of rules which describe Dutch pronunciation variation appropriately. Indeed, finding an optimal set of rules is the

Table 1
The test conditions used to measure the effect modeling pronunciation variation

	Test condition	Lexicon	Phone models	Language models
Baseline	SSS	S	S	S
1	MSS	M	S	S
2	MMS	M	M	S
3	MMM	M	M	M

goal of many rule-based approaches. If each rule can be tested in isolation the way forward is quite obvious. If, however, the outcome of modeling pronunciation variation is enormously influenced by interaction between rules, the way forward is much less straightforward. That is why we decided to pay attention to this issue.

The outline of our article is as follows. In Section 2, the CSR's baseline performance and the general procedure which we used for modeling pronunciation variation are described. A detailed description of the approaches which we used to model pronunciation variation is provided. Subsequently, in Section 3, more details about the CSR and the speech material which we used for our experiments are given. The results obtained with these methods are presented in Section 4. Finally, in Section 5, we discuss the results and their implications.

2. Method

In our research, we tested a method for modeling within-word variation (Section 2.3) and two methods for modeling cross-word variation (Section 2.4). We also tested the combination of the within-word method with each of the cross-word methods (Section 2.5). For all methods, in isolation and in combination, we employed the same general procedure. This general procedure is described in Section 2.2. The starting point, our CSR's baseline performance, is described in Section 2.1.

2.1. Baseline

The starting point of our research was to measure the CSR's baseline performance. It is crucial to have a well-defined lexicon to start out with, since any improvements or deteriorations in recognition performance due to modeling pronunciation variation are measured compared to the results obtained using this lexicon. Our baseline lexicon contains one pronunciation for each word. It was automatically generated using the transcription module of the Text-to-Speech (TTS) system developed at the University of Nijmegen (Kerkhoff and Rietveld, 1994). In this transcription module, phone transcriptions of words were obtained by looking up the transcriptions in two lexica: ONOMASTICA [1] and CELEX (Baayen, 1991). A grapheme-to-phoneme converter was employed whenever a word could not be found in either of the lexica. All transcriptions were manually checked and corrected if necessary. By using this transcription module, transcriptions of the words were obtained automatically, and consistency was achieved. A further advantage of this procedure is that it can also easily be used to add transcriptions of new words to the lexicon.

The phone models were trained on the basis of a training corpus in which the baseline transcriptions were used (see Sections 3.1 and 3.2). The language models were trained on the orthographic representation of the words in the training material. The baseline performance of the CSR was measured by carrying out a recognition test using the lexicon, phone models, and language model described above (test condition: SSS).

2.2. General procedure

Our general procedure for testing methods of modeling pronunciation variation consists of three steps:

1. In the first step, the baseline lexicon is expanded by adding pronunciation variants to it, thus creating a multiple pronunciation lexicon. Using the baseline phone models, baseline language model and this multiple pronunciation lexicon a recognition test is carried out (test condition: MSS).
2. In the second step, the multiple pronunciation lexicon is used to perform a forced recognition. In this type of recognition the CSR is "forced" to choose between different pronunciation variants of a word instead of between different words. Forced recognition is imposed through the language model. For each utterance, the language model is derived on the basis of 100 000 repetitions of the same utterance. This

[1] http://www2.echo.lu/langeng/projects/onomastica/

means that it is virtually impossible for the CSR to choose other words than the ones present in the utterance. Still, a small percentage of sentences (0.4–0.5%) are incorrectly recognized. In those cases, the baseline transcriptions are retained in the corpus. In all other cases, the baseline transcriptions are replaced by the transcription of the recognized pronunciation variants. A new set of phone models is trained on the basis of the resulting corpus containing pronunciation variants. We expect that by carrying out a forced recognition, the transcriptions of the words in the training corpus will match more accurately with the spoken utterance. Consequently, the phone models trained on the basis of this corpus will be more precise. A recognition test is performed using the multiple pronunciation lexicon, the retrained phone models and the baseline language model (test condition: MMS).

3. In the third step, the language model is altered. To calculate the baseline language model the orthographic representation of the words in the training corpus is used. Because there is only one variant per word this suffices. However, when a multiple pronunciation lexicon is used during recognition and the language model is trained on the orthographic representation of the words, all variants of the same word will have equal a priori probabilities (this probability is determined by the language model). A drawback of this is that a sporadically occurring variant may have a high a priori probability because it is a variant of a frequently occurring word, whereas the variant should have a lower a priori probability on the basis of its occurrence. Consequently, the variant may be easily confused with other words in the lexicon. A way of reducing this confusability is to base the calculation of the language model on the phone transcription of the words instead of on the orthographic transcription, i.e. on the basis of the phone transcriptions of the corpus obtained through forced recognition. A recognition test is performed using this language model, the multiple pronunciation lexicon and the updated phone models (test condition: MMM).

2.3. Method for modeling within-word pronunciation variation

The general procedure, described above, was employed to model within-word pronunciation variation. Pronunciation variants were automatically generated by applying a set of optional phonological rules for Dutch to the transcriptions in the baseline lexicon. The rules were applied to all words in the lexicon wherever it was possible and in no specific order, using a script in which the rules and conditions were specified. All of the variants generated by the script were added to the baseline lexicon, thus creating a multiple pronunciation lexicon. We modeled within-word variation using five optional phonological rules concerning: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion (SAMPA [2]-notation is used throughout this article). These rules were chosen according to the following four criteria.

First, we decided to start with rules concerning those phenomena that are known to be most detrimental to CSR. Of the three possible processes, i.e. insertions, deletions and substitutions, we expect the first two to have the largest consequences for speech recognition, because they affect the number of segments present in different realizations of the same word. Therefore, using rules concerning insertions and deletions was the first criterion we adopted. The second criterion was to choose rules that are frequently applied. Frequently applied is amenable to two interpretations. On the one hand, a rule can be frequent because it is applied whenever the context for its application is met, which means that the most frequent form would probably suffice as sole transcription. On the other hand, a rule can be frequent because the context in which the rule can be applied is very frequent (even though the rule is applied e.g. only in 50% of the cases). It is this type of frequent occurrence which is interesting because in this case it is difficult to predict which variant should be taken as the baseline form. Therefore, all possible variants should probably be included in the lexicon. The third criterion (related to the previous

[2] http://www.phon.ucl.ac.uk/home/sampa/dutch.htm

one) was that the rules should be relevant to phones that are relatively frequent in Dutch, since rules that concern infrequent phones probably have fewer consequences for the recognizer's performance. Finally, we decided to start with rules that have been extensively described in the literature, so as to avoid possible effects of overgeneration and undergeneration due to incorrect specification of the rules.

The description of the four rules: /n/-deletion, /t/-deletion, /@/-deletion and /@/-insertion is according to Booij (1995), and the description of the /r/-deletion rule is according to Cucchiarini and van den Heuvel (1995). The descriptions given here are not exhaustive, but describe how we implemented the rules.

(1) /n/-deletion: In standard Dutch, syllable-final /n/ can be dropped after a schwa, except if that syllable is a verbal stem or if it is the indefinite article *een* /@n/ "a". For many speakers, in particular in the western part of the Netherlands, the deletion of /n/ is obligatory. For example:

reizen /rEiz@n/ → /rEiz@/

(2) /r/-deletion: The rule for /r/-deletion can be divided into three parts based on the type of vowel preceding the /r/. First, /r/-deletion may occur if it is in the coda, preceded by a schwa and followed by a consonant. For example:

Amsterdam /Amst@rdAm/ → /Amst@dAm/

Second, for the cases where /r/ follows a short vowel, Cucchiarini and van den Heuvel (1995) make a distinction between unstressed and stressed short vowels. They state that after a short, stressed vowel in coda position, /r/-weakening can take place, but /r/-deletion is not allowed. However, we decided to treat /r/-weakening in the same way as /r/-deletion because there is no intermediate phone model in our phone set which describes /r/-weakening. Thus, we created pronunciation variants which, based on the rules, might be improbable, but we decided to give the CSR the possibility to choose. For example:

stressed: *Arnhem* /ARnEm/ → /AnEm/

unstressed: *Leeuwarden*
/le:wARd@n/ → /le:wAd@n/

Third, /r/-deletion may occur if it is in the coda, preceded by a long vowel and followed by a consonant. For example:

Haarlem /ha:RlEm/ → /ha:lEm/

(3) /t/-deletion: The process of /t/-deletion is one of the processes that typically occurs in fast speech, but to a lesser extent in careful speech. If a /t/ in a coda is preceded by an obstruent, and followed by another consonant, the /t/ may be deleted. For example:

rechtstreeks /rExtstre:ks/ → /rExstre:ks/

If the preceding consonant is a sonorant, /t/-deletion is possible, but then the following consonant must be an obstruent (unless the obstruent is a /k/). For example:

's avonds /sa:vOnts/ → /sa:vOns/

Although Booij does not mention that in some regional variants /t/-deletion also occurs in word-final position, we decided to apply the /t/-deletion rule in word-final position following an obstruent (unless the obstruent is an /s/). For example:

Utrecht /ytrExt/ → /ytrEx/

(4) /@/-deletion: When a Dutch word has two consecutive syllables headed by a schwa, the first schwa may be deleted, provided that the resulting onset consonant cluster consists of an obstruent followed by a liquid. For example:

latere /la:t@r@/ → /la:tr@/

(5) /@/-insertion: In nonhomorganic consonant clusters in coda position schwa may be inserted. If the second of the two consonants involved is an /s/ or a /t/, or if the cluster is a nasal followed by a homorganic consonant, /@/-insertion is not possible. Example:

Delft /dELft/ → /dEl@ft/

Each of the rules described above was tested in isolation by adding the variants to the lexicon and carrying out a recognition test. Tests were also carried out for all five rules together. In this case, all the steps of the general procedure were carried out.

2.4. Modeling cross-word pronunciation variation

The two different methods we used to model cross-word pronunciation variation are explained below. The type of cross-word variation which we modeled concerns processes of cliticization, contraction and reduction (Booij, 1995).

2.4.1. Method 1 for modeling cross-word pronunciation variation

The first step in cross-word method 1 consisted of selecting the 50 most frequently occurring word sequences from our training material. Next, from those 50 word sequences we chose those words which are sensitive to the cross-word processes cliticization, contraction and reduction. This led to the selection of seven words which made up 9% of all the words in the training corpus (see Table 2). The variants of these words were added to the lexicon and the rest of the steps of the general procedure were carried out (see Section 2.2). Table 2 shows the selected words (column 1), the total number of times the word occurs in the training material (column 2), their baseline transcriptions (column 3) and their added cross-word variants (column 4).

2.4.2. Method 2 for modeling cross-word pronunciation variation

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 joined together and added as separate entities to the lexicon. 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 could be chosen. Thus, 22 multi-words were selected. Subsequently, these multi-words were added to the lexicon and the language model. It was necessary for us to also add the multi-words to the language model, because effectively, for our CSR they are "new" words. Next, the cross-word variants of the multi-words were also added to the lexicon, and the remaining steps of the general procedure were carried out (see Section 2.2).

All of the selected multi-words have at least two pronunciations. If the parts of the multi-words are counted 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 that for cross-word method 1 due to the contextual constraints imposed by the multi-words. Table 3 shows the multi-words (column 1), the total number of times the multi-word occurs in the training material (column 2), their baseline transcriptions (column 3) and their added cross-word variants (column 4).

Table 2
The words selected for cross-word method 1, their counts in the training material, baseline transcriptions and added cross-word variants

Selected word	Count	Baseline	Variant(s)
ik	3578	Ik	k
dat	1207	dAt	dA
niet	1145	nit	ni
is	643	Is	s
de	415	d@	d
het	382	@t	hEt, t
dit	141	dIt	dI

2.5. Combination of the within-word and cross-word methods

In addition to testing the within-word method and the two cross-word methods in isolation, we also employed the general procedure to test the combination of the within-word method and cross-word method 1, and the combination of the within-word method and cross-word method 2. In these experiments the within-word pronunciation variants and the cross-word pronunciation variants were added to the lexica simultaneously.

For the combination of the within-word method with cross-word method 2, an extra set of experiments was carried out. This was necessary in order to be able to split the effect of adding multi-words from the effect of adding the multi-words' pronunciation variants. To achieve this, the experiments for the within-word method were repeated with the multi-words added to the lexica.

Table 3
The multi-words selected for cross-word method 2, their counts in the training material, baseline transcriptions and added cross-word variants

Multi-word	Count	Baseline	Variant(s)
ik_wil	2782	IkwIl	kwIl
dat_is	345	dAtIs	dAIs, dAs
ja_dat_klopt	228	ja:dAtklOpt	ja:dAklOpt
niet_nodig	224	nitno:d@x	nino:d@x
wil_ik	196	wIlIk	wIlk
dat_hoeft_niet	181	dAthuftnit	dAhuftnit, dAhuftni, dAthuftni
ik_heb	164	IkhEp	khEp
niet_naar	122	nitna:R	nina:R
het_is	74	@tIs	hEtIs, tIs
dit_is	74	dItIs	dIIs, dIs
niet_vanuit	72	nitvAn9yt	nivAn9yt
de_eerste	45	d@e:Rst@	de:Rst@
ik_zou	40	IkzAu	kzAu
ik_weet	38	Ikwe:t	kwe:t
ik_wilde	35	IkwIld@	kwIld@
niet_meer	31	nitme:R	nime:R
ik_hoef	31	Ikhuf	khuf
ik_moet	26	Ikmut	kmut
dit_was	25	dItwAs	dIwAs
ik_zei	24	IkzEi	kzEi
heb_ik	22	hEpIk	hEpk
is_het	20	Is@t	IshEt, Ist

The effect of the inclusion of multi-words in the language model and the lexica could then be measured by comparing these results to the results of the within-word method in isolation.

3. CSR and material

3.1. CSR

The main characteristics of the CSR are as follows. The input signals consist of 8 kHz, 8 bit A-law coded samples. Feature extraction is done every 10 ms for 16 ms frames. The first step in feature analysis is an FFT analysis to calculate the spectrum. In the following step, the energy in 14 mel-scaled filter bands between 350 and 3400 Hz is calculated. Next, a discrete cosine transformation is applied to the log filterband coefficients. The final processing stage is a running cepstral mean substraction. 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, word-based language models (unigram and bigram) and a lexicon. The acoustic models are continuous density hidden Markov models (HMMs) with 32 Gaussians per state. The topology of the HMMs is as follows: each HMM consists of six states, three parts of two identical states, one of which can be skipped (Steinbiss et al., 1993). In total, 39 HMMs were trained. For each of the phonemes /l/ and /r/, two models were trained, because a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/). For each of the other 33 phonemes context-independent models were trained. In addition, one model was trained for non-speech sounds and a model consisting of only one state was employed to model silence.

3.2. Material

Our training and test material, selected from the VIOS database (Strik et al., 1997), consisted of 25 104 utterances (81 090 words) and 6267 utter-

ances (21 106 words), respectively. Recordings with a high level of background noise were excluded.

The baseline training lexicon contains 1412 entries, which are all the words in the training material. Adding pronunciation variants generated by the five phonological rules (within-word method) increases the size of the lexicon to 2729 entries (an average of about 2 entries per word). The maximum number of variants that occurs for a single word is 16. For cross-word method 1, eight variants were added to the lexicon. For cross-word method 2, 22 multi-words and 28 variants of the multi-words were added to the lexicon.

The baseline test lexicon contains 1154 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, e.g. station names. The test corpus does not contain any out-of-vocabulary words. This is a somewhat artificial situation, but we did not want the CSR's 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 (within-word method) leads to a lexicon with 2273 entries (also an average of about 2 entries per word). For cross-word methods 1 and 2, the same variants were added to the test lexicon as those which were added to the training lexicon.

4. Results

The results in this section are presented as best sentence word error rates (WER). The percentage WER is determined by

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100,$$

where S is the number of substitutions, D the number of deletions, I the number of insertions and N is the total number of words. During the scoring procedure only the orthographic representation was used. Whether or not the correct pronunciation variant was recognized was not taken into account. Furthermore, before scoring took place, the multi-words were split into the separate words they consist of. The significance of differences in WER was calculated with a t-test for comparison of means ($p = 0.05$) for independent samples.

Table 4 shows the results for modeling pronunciation variation for all methods in isolation, and the various combinations of methods. In Section 4.1, the results for the within-word method are described, and in Section 4.2, this is done for the two cross-word methods. Subsequently, the results of combining the within-word method with each of the cross-word methods are described in Section 4.3. In Section 4.4, a comparison is made between testing the methods in isolation and in combination. Finally, the overall results are presented in Section 4.5.

4.1. Modeling within-word pronunciation variation

Row 2 in Table 4 (within) shows the results of modeling within-word pronunciation variation. In column 2, the WER for the baseline condition (SSS) is given. Adding pronunciation variants to the lexicon (MSS) leads to an improvement of 0.31% in WER compared to the baseline (SSS). When, in addition, retrained phone models are

Table 4
WER for the within-word method (within), cross-word method 1 (cross 1), cross-word method 2 (cross 2), the within-word method with multi-words added to the lexicon and language model (within + multi), and the combination of the within-word method with cross-word method 1 (within + cross 1) and cross-word method 2 (within + cross 2)

	SSS	MSS	MMS	MMM
within	12.75	12.44	12.22	12.07
cross 1	12.75	13.00	12.89	12.59
cross 2	12.41*	12.74	12.99	12.45
within + multi	12.41*	12.05	11.81	11.72
within + cross 1	12.75	12.70	12.58	12.14
within + cross 2	12.41*	12.37	12.30	11.63

* Multi-words added to the lexicon and the language model.

used (MMS), a further improvement of 0.22% is found compared to the MSS condition. Finally, incorporating variants into the language model leads to an improvement of 0.15% compared to the MMS condition. In total, a significant improvement of 0.68% was found (SSS → MMM) for modeling within-word pronunciation variation.

4.2. Modeling cross-word pronunciation variation

Rows 3 (cross 1) and 4 (cross 2) in Table 4 show the results for each of the cross-word methods tested in isolation. It is important to note that the SSS condition for cross-word method 2 is different from the SSS condition for cross-word method 1. This is due to adding multi-words to the lexicon and the language model, which is indicated by an asterisk in Table 4. Adding multi-words to the lexicon and language model leads to an improvement of 0.34% (SSS → SSS*).

In contrast to the within-word method, adding variants to the lexicon leads to deteriorations of 0.25% and 0.33% WER for cross-word methods 1 and 2, respectively (SSS → MSS, SSS* → MSS). Although for cross-word method 1, part of the deterioration is eliminated when retrained phone models are used (MMS), there is still an increase of 0.14% in WER compared to the baseline (SSS). Using retrained phone models for cross-word method 2 leads to a further deterioration in WER of 0.25% (MSS → MMS). Adding pronunciation variants to the language model (MMM) leads to improvements of 0.30% and 0.54% for cross-word method 1 and 2 respectively, compared to the MMS condition.

Compared to the baseline, the total improvement is 0.16% for cross-word method 1, and 0.30% for cross-word method 2 (SSS → MMM). However, when the result of cross-word method 2 is compared to the SSS* condition (multi-words included), a deterioration of 0.04% is found (SSS* → MMM).

4.3. Modeling within-word and cross-word pronunciation variation

As was explained in Section 2.5, two processes play a role when using multi-words to model cross-word pronunciation variation, i.e., firstly, adding the multi-words and, secondly, adding variants of the multi-words. To measure the effect of only adding the multi-words (without variants), the experiments for within-word variation were repeated with the multi-words added to the lexicon and the language model. Row 5 in Table 4 (within + multi) shows the results of these experiments. The effect of the multi-words can be seen by comparing these results to the results of the within-word method (row 2 in Table 4). The comparison clearly shows that adding multi-words to the lexicon and the language model leads to improvements for all conditions. The improvements range from 0.34% to 0.41% for the different conditions.

In row 6 (within + cross 1) and row 7 (within + cross 2) of Table 4, the results of combining the within-word method with the two cross-word methods are shown. It can be seen that adding variants to the lexicon improves the CSR's performance by 0.05% and 0.04% for cross-word methods 1 and 2, respectively (SSS → MSS, SSS* → MSS). Using retrained phone models (MSS → MMM) improves the WER by another 0.12% for cross-word method 1, and 0.07% for cross-word method 2. Finally, the improvements are largest when the pronunciation variants are used in the language model too (MMM). For cross-word method 1, a further improvement of 0.44% is found compared to MMS, and for cross-word method 2, an even larger improvement of 0.67% is found.

For the combination of the within-word method with cross-word method 1, a total improvement of 0.61% is found for the test condition MMM compared to the baseline (SSS). For the same test condition, the combination of the within-word method with cross-word method 2 leads to a total improvement of 0.78% compared to the SSS* condition.

4.4. Comparing methods in isolation and in combination

In order to get a clearer picture of the differences in results obtained when modeling pronunciation variation in isolation and in combination,

the results presented in the previous sections were analyzed to a further extent.

First, the difference in WER (ΔWER) between each of the methods tested in isolation and the baseline was calculated. Next, the ΔWER for each of the cross-word methods in isolation was added to the ΔWER for the within-word method in isolation. The results of these summations are indicated by the "sum" bars in Figs. 1 and 2. The differences in WER between the baseline and the combinations of within-word and cross-word methods 1 and 2 were also calculated. These results are shown in Figs. 1 and 2 and are indicated by the "combi" bars. Fig. 1 shows the results for cross-word method 1, and Fig. 2 shows the results for cross-word method 2.

In these figures, it can be seen that the sum of the improvements for the two methods tested in isolation is not the same as the improvement obtained when testing the combinations of the methods. For cross-word method 1, the sum of the methods in isolation gives better results, whereas for cross-word method 2, the combination leads to higher improvements.

Fig. 3 shows the differences in WER between the results of adding variants of each of the five phonological rules to the lexicon separately, the summation of these results ("sum") and the result of the combination of all five rules ("combi"). The differences shown in Fig. 3 are all on the basis of the MSS condition, i.e. variants are only added to the lexicon. In isolation, the rule for /n/-deletion leads to an improvement. The variants generated by the rules for /r/-deletion and /@/-deletion seem to have almost no effect at all. The variants for /t/-deletion and /@/-insertion have some effect, but lead to a deterioration in WER compared to the baseline. The sum of these results is a deterioration

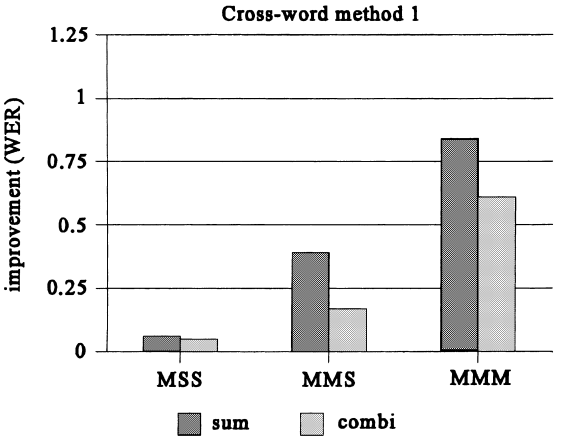

Fig. 1. Improvements (WER) for cross-word method 1 combined with the within-word method and the sum of the two methods in isolation.

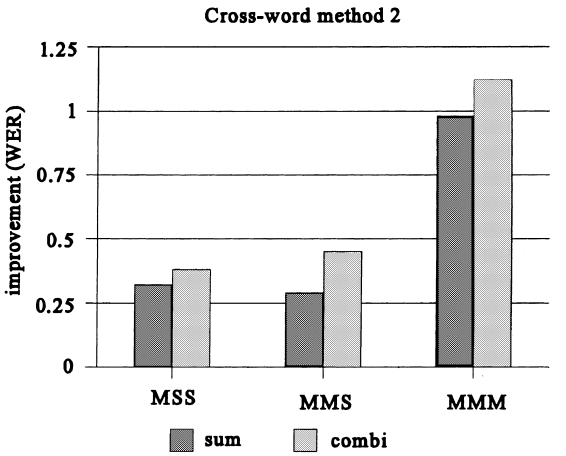

Fig. 2. Improvements (WER) for cross-word method 2 combined with the within-word method and the sum of the two methods in isolation.

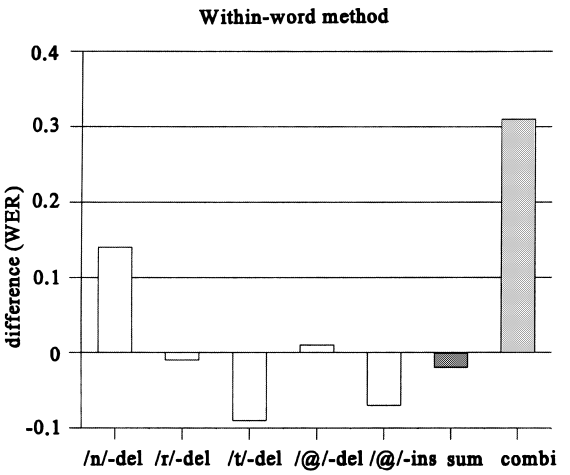

Fig. 3. Difference in WER between the baseline result and results of adding variants of separate rules to the lexicon, sum of those results, and combination result of all rules.

in WER of 0.02%. However, combining all methods, leads to an improvement of 0.31% compared to the baseline.

4.5. Overall results

For all methods, the best results are obtained when pronunciation variants are used during training and recognition, and when they are added to the language model (MMM). All methods lead to an improvement in the CSR's performance when their results are compared to the result of the baseline (SSS). These improvements are summed up in Table 5. Modeling within-word variation in isolation gives a significant improvement of 0.68%, and in combination with cross-word method 2, the improvement is also significant.

Up until now we have only presented our results in terms of WER (as is done in most studies). WERs give an indication of the net change in the performance of one CSR compared to another one. However, they do not provide more detailed information on how the recognition results of the two CSRs differ. Since this kind of detailed information is needed to gain more insight, we carried out a partial error analysis. To this end, we compared the utterances recognized with the baseline test to those recognized with our best test (MMM for within + cross 2 in Table 4). For the moment, we have restricted our error analysis to the level of the whole utterance, mainly for practical reasons. In the near future, we plan to do it at the word level too.

The results in Table 6 show how many utterances in the test corpus are actually recognized correctly or incorrectly in the two tests. These results show that 75.7% of the utterances are recognized correctly in both conditions (baseline test correct, final test correct), and 17.3% of the utterances are recognized incorrectly in both conditions. Improvements are found for 4.3% of the utterances (baseline test incorrect, final test correct), and deteriorations are found for 2.9% of the utterances (baseline test correct, final test incorrect).

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 7. The group of 1083 utterances (17.3%) which are recognized incorrectly in both tests (see Table 6) consist of 609 utterances (9.7%) for which both tests produce the same incorrect recognition results and 474 utterances ($17.3 - 9.7 = 7.6\%$) with different mistakes. In addition, improvements were found for 267 utterances (4.3%) and deteriorations for 183 utterances (2.9%), as was already mentioned above. Consequently, the net result is an improvement for only 84 utterances ($267 - 183$), whereas in total the recognition result changes for 924 utterances ($474 + 267 + 183$). These changes are a consequence of our methods of modeling pronunciation variation, but they cannot be seen in the WER.

Table 5
ΔWER for condition MMM compared to the baseline (SSS) for all methods

Method	ΔWER
within	0.68*
cross 1	0.16
cross 2	0.30
within + cross 1	0.61
within + cross 2	1.12*

* Significant improvements.

Table 6
Comparison between baseline test and final test condition: number of correct utterances, incorrect utterances, improvements and deteriorations (percentages between brackets)

		Baseline test	
		Correct	Incorrect
Final test	Correct	4743(75.7%)	267 (4.3%)
	Incorrect	183 (2.9%)	1083(17.3%)

Table 7
Type of change in utterances going from baseline condition to final test condition (percentages between brackets)

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%)

The WER only reflects the net result obtained, and our error analysis has shown that this is only a fraction of what actually happens due to applying our methods.

5. Discussion

In this research, we attempted to model two types of variation: within-word variation and cross-word variation. To this end, we used a general procedure in which pronunciation variation was modeled at the three different levels in the CSR: the lexicon, the phone models and the language model. We found that the best results were obtained when all of the steps of the general procedure were carried out, i.e. when pronunciation variants were incorporated at all three levels. Below, the results of incorporating pronunciation variants at all three levels are successively discussed.

In the first step, variants were only incorporated at the level of the *lexicon*. Compared to the baseline (SSS → MSS), an improvement was found for the within-word method and for the within-word method in combination with each of the two cross-word methods. However, a deterioration was found for the two cross-word methods in isolation. A possible explanation for the deterioration for cross-word method 1 is related to the fact that the pronunciation variants of cross-word method 1 are very short (see Table 2); some of them consist of only one phone. Such short variants can easily be inserted; for instance, the plosives /k/ and /t/ might occasionally be inserted at places where clicks in the signal occur. Furthermore, this effect is facilitated by the high frequency of occurrence of the words involved, i.e. they are favored by the language model. Similar things might happen for cross-word method 2. Let us give an example to illustrate this: A possible variant of the multi-word "ik_wil" /IkwIl/ is /kwIl/. The latter might occasionally be confused with the word "wil" /wIl/. This confusion leads to a substitution, but effectively it is the insertion of the phone /k/. Consequently, insertion of /k/ and other phones is also possible in cross-word method 2, and this could explain the deterioration found for cross-word method 2.

When, in the second step, pronunciation variation is also incorporated at the level of the *phone models* (MSS → MMS), the CSR's performance improved in all cases, except in the case of cross-word method 2. A possible cause of this deterioration in performance could be that the phone models 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 was used to train the surrounding phone models, which results in contaminated phone models. Error-analysis revealed that in 5% of the cases a pause was indeed present within the multi-words in our training material. Further research will have to show whether this was the only cause of the deterioration in performance or whether there are other reasons why retraining phone models using multi-words did not lead to improvements.

In the third step, pronunciation variants were also incorporated at the level of the *language model* (MMS → MMM), which is beneficial to all methods. Moreover, the effect of adding variants to the language model is much larger for the cross-word methods than for the within-word method. This is probably due to the fact that many recognition errors introduced in the first step (see above) are corrected when variants are also included in the language model. When cross-word variants are added to the lexicon (step 1), short sequences of only one or two phones long (like e.g. the phone /k/) can easily be inserted, as was argued above. The output of forced recognition reveals that the cross-word variants occur less frequently than the canonical pronunciations present in the baseline lexicon: on average in about 13% of the cases for cross-word method 1, and 9% for cross-word method 2. In the language model with cross-word variants included, the probability of these cross-word variants is thus lower than in the original language model and, consequently, it is most likely that they will be inserted less often.

One of the questions we posed in the introduction was what the best way of modeling cross-word variation is. On the basis of our results we

can conclude that when cross-word variation is modeled in isolation, cross-word method 2 performs better than cross-word method 1, but the difference is non-significant. In combination with the within-word method, cross-word method 2 leads to an improvement compared to the within-word method in isolation. This is not the case for cross-word method 1, which leads to a degradation in WER. Therefore, it seems that cross-word method 2 is more suitable for modeling cross-word pronunciation variation. It should be noted, however, that most of the improvements gained with cross-word method 2 are due to adding the multi-words to the lexicon and the language model. An explanation for these improvements is that by adding multi-words to the language model the span of the unigram and bigram increases for the most frequent word sequences in the training corpus. Thus, more context information can be used during the recognition process. Furthermore, it should also be noted that only a small amount of data was involved in the cross-word processes which were studied; only 6–9% of the words in the training corpus were affected by these processes. Therefore, we plan to test cross-word methods 1 and 2 for a larger amount of data and a larger number of cross-word processes.

In Section 4.4, it was shown that testing the within-word method and cross-word method 2 in combination leads to better results than the sum of the results of testing the two methods in isolation. For cross-word method 1 the opposite is true, the within-word method in isolation leads to better results. 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 leads to a degradation in WER (compared to the baseline), whereas the combination leads to an improvement. It is clear that the principle of superposition does not apply here, neither for the five rules of the within-word method nor for the within-word method in combination with each of the two cross-word methods. This is due to a number of factors. First of all, different rules can apply to the same words. Consequently, when the five rules are used in combination, pronunciation variants are generated which are not generated for any of the rules in isolation. Furthermore, when methods are employed in combination, confusion can occur between pronunciation variants of each of the different methods. It is obvious that this confusion cannot occur when methods are tested in isolation. Finally, during decoding, the words in the utterances are not recognized independently of each other, and thus, interaction between pronunciation variants can occur. The implication of these findings is 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.

In Sections 4.1–4.4, various methods and their combinations were tested. This was done by calculating the WER after a method had been applied, and comparing this number to the WER of the baseline system. This amount of reduction in WER is a measure which is used in many studies about modeling pronunciation variation (see Strik and Cucchiarini, 1998). Although this measure gives a global idea of the merits of a method, it certainly does not reveal all details of the effect a method has. This became clear through the error analysis which we conducted (see Section 4.4). This error analysis showed that 14.7% of the recognized utterances changed, whereas a net improvement of only 1.3% in the sentence error rate was found (and 1.12% in the WER). Therefore, it is clear that a more detailed error analysis is necessary to obtain real insight into the effect of a certain method.

That is why we intend to carry out more detailed error analyses in the near future. Such a detailed error analysis should not be carried out on the test corpus, because then the test corpus is no longer an independent test set. Therefore, we will be using a development test set to do error analysis. Furthermore, instead of analyzing errors at the level of the whole utterance, we will be looking at the word level, and if necessary at the level of the phones. Through an error analysis, the effect of testing methods in isolation and in combination can be analyzed. It is hoped that this will yield the tools which are needed to decide beforehand which types of pronunciation variation should be modeled and how they should be tested.

To summarize, we obtained the best results when within-word pronunciation variation and cross-word pronunciation variation using multi-words were modeled in combination, and all the steps of the general procedure had been carried out. Using only five phonological rules and 22 multi-words a relative improvement of 8.8% was found (12.75%–11.63%).

Acknowledgements

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

References

Baayen, H., 1991. De CELEX lexicale databank. Forum der Letteren 32 (3), 221–231.

Booij, G., 1995. The Phonology of Dutch. Clarendon Press, Oxford.

Cohen, M.H., 1989. Phonological structures for speech recognition. Ph.D. dissertation. University of California, Berkeley.

Cohen, P.S., Mercer, R.L., 1974. The phonological component of an automatic speech-recognition system. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, 15–19 April 1974, pp. 177–187.

Cremelie, N., Martens, J.-P., 1998. In search of pronunciation rules. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 23–27.

Cucchiarini, C., van den Heuvel, H., 1995. /r/ deletion in standard Dutch. In: Strik et al. (Eds.), Proceedings of the Department of Language and Speech, University of Nijmegen, Vol. 19, pp. 59–65.

Kerkhoff, J., Rietveld, T., 1994. Prosody in Niros with Fonpars and Alfeios. In: de Haan, Oostdijk (Eds.), Proceedings of the Department of Language and Speech, University of Nijmegen, Vol. 18, pp. 107–119.

Kessens, J.M., Wester, M., 1997. Improving recognition performance by modeling pronunciation variation. In: Proceedings of the CLS opening Academic Year '97–'98, pp. 1–19. http://lands.let.kun.nl/literature/kessens.1997.1.html.

Lamel, L.F., Adda, G., 1996. On designing pronunciation lexica for large vocabulary continuous speech recognition. In: Proceedings of ICSLP-96, Philadelphia, pp. 6–9.

Perennou, G., Brieussel-Pousse, L., 1998. Phonological component in automatic speech recognition. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 91–96.

Steinbiss, V., Ney, H., Haeb-Umbach, R., Tran, B.-H., Essen, U., Kneser, R., Oerder, M., Meier, H.-G., Aubert, X., Dugast, C., Geller, D., 1993. The philips research system for large-vocabulary continuous-speech recognition. In: Proceedings of the ESCA Third European Conference on Speech Communication and Technology: EUROSPEECH '93, Berlin, pp. 2125–2128.

Strik, H., Cucchiarini, C., 1998. Modeling pronunciation variation for ASR: Overview and comparison of methods. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 137–144.

Strik, H., Russel, A., van den Heuvel, H., Cucchiarini, C., Boves, L., 1997. A spoken dialogue system for the Dutch public transport information service. Internat. J. Speech Technol. 2 (2), 119–129.

Wiseman, R., Downey, S., 1998. Dynamic and static improvements to lexical baseforms. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 157–162.