ELSEVIER

Speech Communication 40 (2003) 517–534

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A data-driven method for modeling pronunciation variation

Judith M. Kessens *, Catia Cucchiarini, Helmer Strik

A²RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 5600 HD Nijmegen, The Netherlands

Received 2 July 2001; received in revised form 2 July 2002; accepted 29 July 2002

Abstract

This paper describes a rule-based data-driven (DD) method to model pronunciation variation in automatic speech recognition (ASR). The DD method consists of the following steps. First, the possible pronunciation variants are generated by making each phone in the canonical transcription of the word optional. Next, forced recognition is performed in order to determine which variant best matches the acoustic signal. Finally, the rules are derived by aligning the best matching variant with the canonical transcription of the variant. Error analysis is performed in order to gain insight into the process of pronunciation modeling. This analysis shows that although modeling pronunciation variation brings about improvements, deteriorations are also introduced. A strong correlation is found between the number of improvements and deteriorations per rule. This result indicates that it is not possible to improve ASR performance by excluding the rules that cause deteriorations, because these rules also produce a considerable number of improvements. Finally, we compare three different criteria for rule selection. This comparison indicates that the *absolute frequency of rule application* (F_{abs}) is the most suitable criterion for rule selection. For the best testing condition, a statistically significant reduction in word error rate (WER) of 1.4% absolutely, or 8% relatively, is found.

Résumé

Ce papier décrit une méthode à base de règles, guidée par les données, qui est destinée à modéliser les variations de prononciation en reconnaissance automatique de la parole (RAP). Les différentes phases de cette méthode guidée par les données sont les suivantes. Premièrement, les éventuelles variantes de prononciation sont générées en considérant que chaque phone de la transcription canonique peut être omis. Ensuite, nous effectuons une reconnaissance forcée afin de déterminer quelle variante correspond la mieux au signal acoustique. Enfin, les régles sont dérivées en alignant la meilleure variante avec la transcription canonique qui lui correspond. Une analyse des erreurs commises est effectuée afin de mieux comprendre le processus de modèlisation de prononciation. Cette analyse montre que, bien que cette modélisation apporte des améliorations, elle introduit également des détériations. Une forte corrélation entre le nombre d'améliorations et le nombre de détériations par régle a été trouveé. Ce résultat indique qu'il n'est pas possible d'améliorer les performances de la RAP en élimant les règles qui détériorent la reconnaissance pour certains candidats, étant donné que ces mêmes règles améliorent la reconnaissance de beaucoup d'autres candidats. Enfin, nous avons comparé trois critères visant à sélectionner les meilleures règles. La "fréquence absolute d'application d'une règle (F_{abs})"

* Corresponding author. Tel.: +31-0-24-3612055; fax: +31-0-24-3612907.

E-mail addresses: kessens@let.kun.nl, j.kessens@let.kun.nl (J.M. Kessens), c.cucchiarini@let.kun.nl (C. Cucchiarini), w.strik@let.kun.nl (H. Strik).

URL: http://lands.let.kun.nl/.

0167-6393/02/$ - see front matter
doi:10.1016/S0167-6393(02)00150-4

Nomenclature

Symbol list

T_{can}	canonical transcription
T_{dd}	data-driven transcription
F_{abs}	the number of times a rule is applied in T_{dd}
F_{cond}	the number of times the condition for a rule is met in T_{can}
F_{rel}	F_{abs}/F_{cond}
$F_{abs\text{-}rule\ set}$	summation of F_{abs} for each rule in the rule set
$F_{cond\text{-}rule\ set}$	summation of F_{cond} for each rule in the rule set
$F_{rel\text{-}rule\ set}$	$F_{abs\text{-}rule\ set}/F_{cond\text{-}rule\ set}$

est apparue comme le meilleur critère. Pour la meilleure condition de test, nous avons obtenu une réduction du taux d'erreur par mot statistiquement significative de 1.4% (taux absolu) ou de 8% (taux relatif).

Zusammenfassung

Dieser Artikel beschreibt eine regelbasierte datengesteuerte (DG) Methode zur Modellierung der Aussprachevariation in automatischer Spracherkennung. Die DG-Methode gliedert sich in drei Stufen. Zuerst werden die möglichen Aussprachevarianten eines Wortes erzeugt durch Auslassung eines jeden Phonems aus der kanonischen Transkription. Danach wird durch Zwangserkennung diejenige Aussprachevariante gewählt, die dem akustischen Signal am besten gerecht wird. Zuletzt werden durch Alinierung der kanonischen Transkription mit der in akustischer Hinsicht besten Variante die (frequentesten) Regeln abgeleitet. Eine Fehleranalyse erbrachte neue Einsichten in die Ergebnisse dieser Aussprachemodellierung. Es stellte sich heraus, dass Verbesserungen und Verschlimmerungen in der Worterkennung Hand in Hand gingen. Für jede Regel korrelierte die Anzahl der Verbesserungen stark mit der Anzahl der Verschlimmerungen. Dieses Ergebnis weist darauf hin, dass es nicht möglich ist, diejenige Regeln, die die Erkennung bestimmter Wörter verschlimmern, schlechthin zu streichen, weil diese Regeln auch Verbesserungen mit sich bringen. Wir haben, zum Schluss, drei Kriterien zur Regelauswahl verglichen. Die absolute Frequenz für die Anwendung einer Regel (F_{abs}) erwies sich als das geeignetste Kriterium. Für die beste Testbedingung wurde eine statistisch signifikante Reduzierung der Wortfehlerquote festgestellt (1.4% absolut, 8% relativ).

Keywords: Pronunciation variation; Data-driven; Rule-based; Speech recognition; Error analysis; Rule selection

1. Introduction

As has been widely recognized in the last two decades, the enormous variation in pronunciation among speakers of the same language or even the same language variety constitutes a serious challenge to automatic speech recognition (ASR) for an overview, see Strik and Cucchiarini (1999). For this reason, researchers have been looking for ways to model at least part of this variation in order to improve the performance of ASR systems.

In previous papers (Kessens et al., 1999; Wester et al., 1998), we reported on our attempts to model pronunciation variation on the basis of phonological knowledge. We showed that this kind of knowledge can indeed be used to improve the recognition performance of our Dutch continuous speech recogniser (CSR) significantly. However, comprehensive inventories of systematic pronunciation variation do not exist in the literature. In particular, this applies to the type of speech we are dealing with, i.e. extemporaneous/spontaneous speech. As is well known, spontaneous speech is still an under-researched area at the moment (Strik and Cucchiarini, 1999), with the result that the kind of information we would like to have cannot be found in the literature. For this reason, we have

been looking for alternative ways of obtaining information on pronunciation variation.

A method that we have investigated, and that has been used by other authors too (see e.g. Cremelie and Martens, 1999; Fukada et al., 1999; Williams and Renals, 1998; Schiel et al., 1998; Amdal et al., 2000), consists in trying to obtain this information directly from the speech signal, i.e. in a data-driven (DD) manner. As in most DD methods, we use the CSR to get a transcription of the speech signal. However, this is not straightforward. Of course it is possible to carry out unconstrained phone recognition by using the acoustic models alone, i.e. without the top-down constrictions of language model and lexicon, but phone accuracy appears to be only 50–70% in this case, and this is not enough for most purposes. For this reason, the results of free phone recognition are usually filtered or smoothed (see e.g. Riley et al., 1999; Fosler-Lussier, 1999). In the present study, however, we use another approach, namely forced recognition. Forced recognition means that the CSR has to decide for each word in each utterance which pronunciation variant best matches the acoustic signal. Usually, the number of variants that can be chosen during forced recognition is limited. For example, in our knowledge-based approach to modeling pronunciation variation, maximally 16 variants/word were obtained (Kessens et al., 1999). In the approach that we use in this study, however, the number of variants that can be chosen is much larger. By increasing the number of possible variants that can be chosen during forced recognition, the CSR is less constrained and forced recognition more and more resembles (free) phone recognition.

There are two main reasons why we chose only to focus on deletion processes. The first one is that we expect deletions (and insertions) to be more important than substitutions, since substitutions can implicitly be modelled in the phone models. The second reason for choosing deletions, as opposed to or in addition to insertions, is that we expect deletion processes to be more frequent in our speech material. A reason for expecting deletions to be more frequent is that we are dealing with extemporaneous/spontaneous speech. Furthermore, we started off with a lexicon containing a single canonical pronunciation for each word. This canonical pronunciation is a kind of citation form, which contains no deletions except deletions due to a number of obligatory deletion rules (e.g. degemination).

In many DD approaches, the new pronunciation variants found by the CSR are directly added to the lexicon. In some studies, the new information is implemented in terms of rules, which are subsequently used to generate pronunciation variants (e.g. Cremelie and Martens, 1999; Amdal et al., 2000). In the present study, we employ data-derived rules. The main reason for using rules instead of adding the variants directly to the lexicon is that it is easier to draw conclusions in terms of rules than in terms of the individual pronunciation variants, since there are more observations available per rule than per individual variant.

Since we are convinced that just reporting on decreased WERs does not contribute very much to our understanding of pronunciation variation modeling and the way this can improve CSR performance, we carried out a detailed error analysis. The goal of this error analysis is to determine how the changes in WER came about. Despite its indisputable importance for understanding what is really going on, this kind of analysis is seldom performed in pronunciation variation modeling research. Exceptions are Ravishankar and Eskenazi (1997); Kessens et al. (1999) and Wester et al. (2000). A limitation of the error analysis performed by Ravishankar and Eskenazi (1997) is that it was carried out manually, with the consequence that the amount of material that could be analysed was limited. A limitation of the error analyses carried out in our previous studies (Kessens et al., 1999; Wester et al., 2000) is that these analyses were performed at sentence level. The present error analysis is performed automatically, and at word level.

The aim of the present paper is threefold. First, we analyse whether the DD method of modeling pronunciation variation that we adopted leads to a reduction in WER. Our second goal is to find out how the changes in WER came about. This is done by performing a detailed error analysis on the recognition results. The third goal of this paper is to examine the adequacy of three criteria for rule

selection, with a view to establishing which one is the most promising. With this knowledge, it would be possible to make more sound choices about which rules (or which pronunciation variants) to model. The three goals described above will be dealt with in Sections 3–5 of this paper, preceded by Section 2, in which details are given about the speech material and the CSR we used. Section 6 contains a general discussion of the findings presented in this paper, while the main conclusions are drawn in Section 7.

2. Speech material and CSR

2.1. Speech material

Our speech material was selected from the VIOS database, which contains a large number of telephone calls recorded with the on-line version of a spoken dialogue system called OVIS (Strik et al., 1997). OVIS is employed to automate part of an existing Dutch public transport information service. The total VIOS material was divided into three non-overlapping corpora. Table 1 shows the statistics of these three corpora. The second column displays the number of utterances that are included in each corpus (# utterances). The third column shows the number of words (# words), and the last column displays the percentage of the total VIOS database (percentage).

2.2. CSR

The main characteristics of the CSR are as follows. The input signals were sampled at 8 kHz using 8 bit A-law coding. The front-end acoustic processing consists of calculating 14 MFCCs plus their deltas, every 10 ms for 16 ms frames. 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. For more details on the CSR, see Strik et al. (1997). The test and training lexica contain 1288 words and 1465 words, respectively, plus three entries; one for noise and two for filled pauses. The baseline lexicon contains one transcription per word. This so-called 'canonical transcription' is obtained using a text-to-speech system for Dutch (Kerkhoff and Rietveld, 1994) followed by a manual correction. The acoustic models and language models (unigram and bigram) are estimated on the training material.

Table 1
Statistics of the three corpora

Corpus	# Utterances	# Words	Percentage
Training	59,640	176,080	60
Test	19,880	58,647	20
Error analysis	19,880	58,630	20
Total	99,400	293,357	100

3. WER reduction through data driven modeling of pronunciation variation

The goal of the first phase of the research is to analyse whether the DD method of modeling pronunciation variation that we have adopted indeed leads to a reduction in WER. The pronunciation variants that we use in the recognition experiments are generated using rules that are derived on the basis of automatic transcriptions of the training data. In Section 3.1, the automatic rule extraction procedure and the procedure for selection of the candidate rules are described. This is followed by a description of the recognition experiments in Section 3.2. Subsequently, in Section 3.3, the results are presented. Finally, in Section 3.4 we discuss the results and we draw conclusions.

3.1. Obtaining the rules

3.1.1. Automatic extraction of candidate rules

The candidate rules were extracted from automatic transcriptions of all the utterances in the training corpus. The following five steps describe

the whole procedure of automatic extraction of the candidate rules:

1. For each word in an utterance, the canonical transcription (T_{can}) is looked up in the baseline lexicon.
2. Pronunciation variants are generated by making each phone in T_{can} optional, with the constraint that one phone per syllable should remain present. For example: Suppose T_{can} is '/wIL/' (want), then the following pronunciation variants were generated for this word: /wIL/, /wI/, /wL/, /IL/, /w/, /I/ and /L/.
3. With all the generated pronunciation variants, forced recognition is performed using the baseline phone models. During forced recognition, the CSR does not choose between all the words in the lexicon, instead, for each word in the utterance, it has to determine the pronunciation variant that best matches the acoustic signal (see Wester et al., 2001). In this way, DD transcriptions (T_{dd}) of all the utterances of the training corpus are obtained.
4. A dynamic programming algorithm is used to align T_{can} with T_{dd}. An example of the alignment of T_{can} with T_{dd} is the following:

T_{can} | d @ | v @ R b I n d I N | Y t r E x t |
('|' = word boundary)

T_{dd} | d @ | v @ – b I n – I N | Y t r E – – |
('–' = deletion)

5. Using the alignments obtained in step 4, we formulate candidate deletion rules. These rules are defined in the following manner:

$/\mathtt{L\ F\ R}/_{can} \rightarrow /\mathtt{L\ -\ R}/_{dd}$

This means that the focus phone F in T_{can} following the phone L (left context) and preceding the phone R (right context) is deleted in T_{dd}. The left and right context can be a phone or a word boundary. These kinds of rules are referred to as 'rewrite rules' in literature (see Strik and Cucchiarini, 1999). It should be noted that this rule formalism is different from the one that is normally adopted in knowledge-based studies. The most striking difference is that knowledge-based rules are usually more generally formulated. For example, L and R can be classes of phones, instead of one single phone.

For each candidate rule, we also calculate the following three frequency measures:

- F_{cond}: the number of times the condition for the rule ($/\mathtt{L\ F\ R}/$) is met in T_{can},
- F_{abs}: the number of times a rule is applied in T_{dd}, and
- F_{rel}: $F_{abs}/F_{cond} (0 < F_{rel} \leqslant 1)$.

3.1.2. Motivations for performing rule selection

Before using the rules to generate variants for the recognition experiments, we made a selection of the candidate rules. In the research on modeling of pronunciation variation, rule (or variant) selection forms a vital part of the research methodology (for an overview of rule selection procedures, see Strik, 2001). There are various motivations for performing rule/variant selection. First of all, the addition of pronunciation variants to the lexicon increases confusability, especially if the lexicon is large. This means that the more variants are included in the lexicon, the more lexical confusability increases due to the addition of variants. The large increase in confusability is probably the reason why usually only small improvements or even deteriorations are found if the number of variants that has been included in the lexicon is very large. By making an appropriate selection of the pronunciation variants, the balance between solving and introducing errors is probably more positive. A second reason for constraining the number of variants is to limit decoding time, since decoding time is directly related to the size of the lexicon. Third, in DD approaches, the data-derived variants are usually selected or filtered, as the variants might be based on artefacts of the CSR instead of being based on genuine pronunciation variation. In this paper, there are two extra reasons for performing rule selection. First of all, we carried out an error analysis procedure at rule level. In order to ensure that substantial changes in WER are measured, it is necessary to select the rules that are most 'promising' in this respect. Second, we estimate prior probabilities of

pronunciation variants based on automatic transcriptions of the training material (obtained through forced recognition). In order to reliably estimate the prior probabilities, the number of observed variants should not be too small.

Several measures have been used to select rules or variants, e.g.: confidence measures (e.g. Wiliams, 1999), a maximum likelihood criterion (e.g. Holter and Svendsen, 1999), and confusability measures (Wester and Fosler-Lussier, 2000) and entropy (Yang and Martens, 2000a). In this paper, we concentrate on frequency measures to select the rules. One is inclined to think that the most frequent rules should be selected, but rules can be frequent in three different ways: (1) because the condition for rule application occurs frequently (F_{cond} is large), (2) because the rule is frequently applied (F_{abs} is large), and (3) because the rule is frequently applied in relation to the number of times its condition for application is met (F_{rel} is large). Several other authors have used frequency measures for rule or variant selection, or have used frequency measures as part of the selection procedure. For instance, Riley et al. (1999) and Lehtinen and Safra (1998) use F_{rel} to select variants. Others, like Williams and Renals (1998), use F_{abs} as part of their variant selection method. Furthermore, a combination of F_{rel} and F_{abs} is also used as a criterion to select variants (Schiel et al., 1998; Ravishankar and Eskenazi, 1997). For rule selection, F_{rel} is probably used more often (see e.g. Cremelie and Martens, 1999; Amdal et al., 2000).

3.1.3. Details on the rule selection procedure

The first criterion we applied was to select the rules for which $F_{\text{abs}} > 100$. This was done for various reasons. First, the DD transcriptions may contain errors due to artefacts of the CSR. Since it can be expected that transcription errors occur randomly, the rules that are based on transcription errors are probably not as frequent as the rules that are based on genuine deletion processes. For this reason, we expect them to be filtered out if the threshold for F_{abs} is set to 100. Furthermore, we expect that a minimum number of occurrences of 100 is enough to ensure substantial changes in WER and to reliably estimate the probabilities of the pronunciation variants. The second criterion we applied was to exclude the rules for which either the left or the right context was deleted, or in other words, we excluded the rules based on transcriptions with two or more deletions in a row. This is done because these deletions occur probably less often, and the occurrence of two deletions in a row might be an indication of an error.

After applying the automatic rule extraction procedure to the training corpus, in total 2951 candidate rules were obtained, which together describe the deletions of 8.5% of the total number of 686,909 phones in the training corpus. If the two selection criteria are applied simultaneously, about half of the deletions are covered, whereas the size of the rule set is reduced to 3% of the original size. The first selection criterion ($F_{\text{abs}} > 100$) appears to be the strictest pruning measure, since it excludes 20% more rules than the second selection criterion (L and R not deleted). By applying the two selection criteria simultaneously, 91 of the 2951 rules are selected. In Appendix A, the statistics of the 91 selected rules are given. A number of the rules that are found are related to phonological processes described in the literature. For example, rule 9 (word final deletion of /n/ after /@/) is very similar to the process of /n/-deletion (Booij, 1995). More examples of plausible deletion rules are described in Kessens et al. (2000).

3.2. Recognition experiments

The 91 selected rules are tested in recognition experiments by composing various sets of rules. At this point of the research, we had no certainty about the optimal criterion for rule selection. As F_{rel} is probably used most often for rule selection, we used F_{rel} for selection of the various rule sets. Seven sets of rules were selected by varying the threshold for F_{rel}. These threshold values are shown in the second column of Table 2 ($F_{\text{rel}} >$). Next, we applied the selected rules to the transcriptions in the baseline test lexicon in order to generate pronunciation variants. By adding these variants to the baseline test lexicon, different multiple pronunciation lexica were obtained. In Table 2, the statistics of the multiple pronunciation lexica are given. The third column displays the number of rules that were selected (# Rules). The

Table 2
Statistics of the multiple pronunciation lexica

	F_{rel} >	# Rules	# Added variants	< Variants/ word >	Max.
1	0.50	7	81	1.06	4
2	0.40	10	322	1.25	8
3	0.30	16	466	1.36	12
4	0.20	25	702	1.54	12
5	0.15	38	993	1.77	12
6	0.10	53	1896	2.47	64
7	0	91	3528	3.73	128

fourth column shows the number of added variants (# Added variants), and column five displays the average number of pronunciation variants per word present in the recognition lexicon (<Variants/word>). Finally, in the last column, the maximum number of pronunciation variants per word is given (Max.).

The selected sets of rules were tested in recognition experiments. As in Kessens et al. (1999) three other testing conditions were used in addition to the baseline testing condition (SSS). In short, these testing conditions imply incorporating the pronunciation variants at all three levels of the CSR: the lexicon, the phone models and the language model.

Testing condition MSS: The *lexicon* is expanded by adding pronunciation variants to it, thus creating a multiple pronunciation lexicon. The only difference with the baseline testing condition SSS is that in testing condition MSS the baseline lexicon is replaced by a multiple pronunciation lexicon.

For the other two testing conditions, an extra step is needed. In this step, pronunciation variants are automatically transcribed in the training corpus. This is accomplished by performing forced recognition with the baseline phone models and the set of variants which have been automatically generated with the selected set of rules.

Testing condition MMS: The *phone models* are retrained on the basis of the new transcription of the training corpus. The only difference with testing condition MSS is that in testing condition MMS the baseline phone models are replaced by the retrained phone models.

Testing condition MMM: A new *language model* is calculated on the basis of automatic transcriptions of the pronunciation variants in the training corpus. In the baseline language model all pronunciation variants of the same word are assigned equal prior probabilities. In the new language model, however, different variants of the same word are assigned their own specific prior probabilities. These prior probabilities are calculated on the basis of the automatic transcriptions of the pronunciation variants in the training corpus. The only difference with testing condition MMS is that in testing condition MMM the baseline language model is replaced by the new language model.

3.3. Results of recognition experiments

The WER is defined as follows:

$$\mathrm{WER} = \frac{S + D + I}{N} \tag{1}$$

where S is the number of substitutions, D the number of deletions, I the number of insertions, and N the total number of words. The WER of 16.94% for our baseline system (SSS) is indicated by the symbol '•' in Fig. 1. Furthermore, the WERs for the three testing conditions are plotted as a function of the average number of variants per word in the lexicon (for the correspondence between the average number of variants per word and the number of rules, see Table 2). The reason for using the average number of variants per word is that this measure is directly related to the size of the lexicon, and thus to decoding time.

Fig. 1 shows the following trends when going from using 1 variant/word to 3.7 variants/word:

(1) Testing condition MSS: The WER first decreases, but if more than 1.5 variants/word (25 rules) are used the WER increases until the level of the baseline system is reached for 2.5 variants/word (53 rules). When 3.7 variants/word are used (91 rules), a large increase in WER is measured compared to the baseline (SSS or 1 variant/word).
(2) Testing condition MMS: The same trend is observed as for testing condition MSS, but the absolute values of the WERs are somewhat lower.
(3) Testing condition MMM: As opposed to the previous testing conditions, the WERs are always

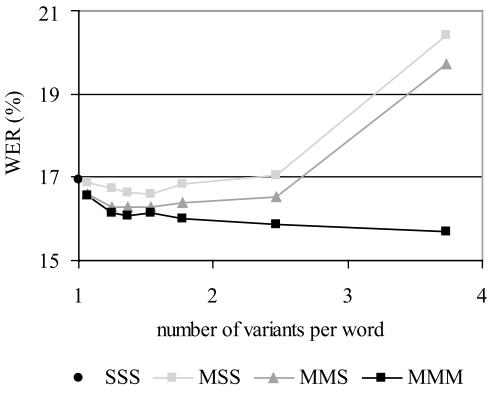

Fig. 1. WERs for the different testing conditions.

lower than the WER for the baseline testing condition. The WER reduction is significant (t-test, $\alpha = 0.05$) for 1.25 or more variants/word (or: 10 or more rules). Furthermore, it can be seen that the decrease in WER becomes smaller with an increasing average number of variants per word. This means that a similar gain in performance will cost more and more in terms of decoding time.

3.4. Discussion and conclusions

The recognition experiments demonstrated that the DD rules can be used effectively to improve recognition performance. Our results showed that only adding variants to the lexicon (MSS) does not always lead to a reduction in WER. The WERs were only slightly lower when also retrained phone models were used (MMS). The best results were obtained when, in addition to the new lexicon and phone models, variant-specific probabilities were used in the language model (MMM). The difference in recognition result between testing condition MMM on the one hand and testing condition MSS and MMS on the other hand was largest for the set of 91 rules; without using variant-specific probabilities in the language model (MSS and MMS), a significant deterioration in recognition result is obtained, whereas the opposite is true if each variant is associated with its corresponding probability in the language model (MMM). In previous research in which we used knowledge-based rules, we also found that testing condition MMM yields the best results (Kessens et al., 1999), but we did not find significant deteriorations for the other two testing conditions. Yang and Martens (2000b) have reported on recognition experiments in which the probabilities of the variants were removed. They found that recognition performance rapidly decreases with an increasing number of variants per word in the lexicon. With more than 3 variants/word, the system with variants performed even worse than the baseline system. These results are very comparable to our results, since we found a decrease in recognition performance if more than 2.5 variants/word are used in the lexicon.

For the best testing condition (MMM, 91 rules), we measured a significant improvement in WER of 1.2% absolutely or 7.3% relatively compared to the baseline (SSS). However, at this point it is not clear whether an even larger improvement could be obtained by using more rules. Since we are not only interested in reducing WER, we do not try to further improve recognition performance. At this moment, we first try to understand how exactly the changes in WER came about. In this way we hope to gain insight that might be used to further improve recognition performance.

4. Analysis of the reduction in WER

The goal of this phase of the research is to find out how exactly the reduction in WER came about. This is accomplished by carrying out an error analysis at word level. In Section 4.1, the method of error analysis is described and compared with a method used in a previous study (Kessens et al., 1999). In Section 4.2, the results of the error analysis are presented. Finally, in Section 4.3 we discuss the results and summarize our conclusions.

4.1. Method of error analysis

During error analysis, we analysed the changes in recognition result by comparing the recognition result of testing condition MMM to the baseline testing condition SSS. The following four steps describe the automatic error analysis procedure:

Table 3
Examples of changes in recognition result between testing condition MMM and SSS

	SPOKEN	SSS	MMM	Type	Category	Rule
1	Ik	ik<Ik>	ik<Ik>	No-change	–	–
2	**wil**	**wil<wIL>**	–	**Deterioration**	**No-variant**	–
3	–	**ik<Ik>**	–	**Improvement**	**No-variant**	–
4	**naar**	**Maarn<ma:Rn>**	**naar<na:>**	**Improvement**	**Variant**	**64:{na:R }**
5	Utrecht	Delft<dELft>	Ede<e:d@>	Different error	–	–

1. *Automatic alignment:* The recognition results of MMM and SSS were aligned with the spoken utterance. This step is necessary in order to determine whether a word is recognized correctly or not (and thus to calculate the WER). An example of the alignment result is given in Table 3. The first column indicates the word number, whereas the second column shows the word that is spoken (SPOKEN). The third column displays the recognized word in the baseline testing condition (SSS), and the fourth column shows the recognized word in testing condition MMM. Between '< >' the transcription of the recognized word is given.
2. *Type of change:* Each change was labelled as 'improvement' (SSS = incorrect, MMM = correct), 'deterioration' (SSS = correct, MMM = correct), 'no-change' (SSS = correct, MMM = correct), or as 'different error' (SSS = incorrect, MMM = incorrect). An example of this labelling is shown in column 5 of Table 3.
3. *Category of change:* Since we are only interested in changes in recognition result that have a direct consequence on the WER, we excluded the different errors from further analysis. Each change (improvement or deterioration) was classified in one out of two categories: the change was labelled as 'variant' if the recognized word was a variant, or 'no-variant' if this was not the case, e.g. word 4 was labelled as variant, whereas words 2 and 3 were labelled as no-variant (see column 6 of Table 3).
4. *Contributions per rule:* For each change that is labelled as variant it was determined by which rule the variant was generated. For example, the variant 'naar < na :>' was generated by applying rule 64 to the word 'naar < na : R >' (see last column of Table 3). In this way, we were able to count the number of times that an improvement or deterioration in recognition result was caused by a specific rule. If more than one rule was applied, then the count was equally distributed over the rules: if N rules were applied to the recognized word, then each of these rules was assigned a score of $1/N$.

4.1.1. Comparison with previous error analysis

In Kessens et al. (1999), we also reported on an error analysis that was carried out to analyse the effect of modeling pronunciation variation. The error analysis that we perform in the present study is different from the previous one in various ways. A first difference is that error analysis was performed at sentence level, whereas in this study it is done at word level. In Kessens et al. (1999) we noted that error analysis should not be carried out on the test corpus, because then the test corpus is no longer an independent test set. Therefore, error analysis is now performed on an independent error analysis corpus. Furthermore, in Kessens et al. (1999) we concluded that due to interaction between pronunciation variants it will not suffice to study rules in isolation. For this reason, in this study we analyse the results of different combinations of rules, and we determine the contribution per rule. Finally, in the current error analysis, we analyse changes in recognition result for the best testing condition 'MMM' instead of for the sub-optimal testing condition 'MSS', as we did in the previous study.

4.2. Results of error analysis

In Section 4.2.1, we present the WERs measured on the error analysis corpus and compare them to the results measured on the test corpus. Next, in the three following sections, the results

are given for the four different steps of the error analysis procedure.

4.2.1. Automatic alignment: WERs

The WER for the baseline testing condition measured on the error analysis corpus is 16.49%. In Fig. 2, the WERs are given for testing condition MMM measured on the test and error analysis corpus, plotted as a function of the average number of variants in the lexicon. It can be seen that the WERs are in general somewhat lower for the error analysis corpus compared to the test corpus. However, in general the same trend is observed: For an increasing number of variants per word the WER decreases, but the decrease in WER becomes smaller if the average number of variants per word is increased.

4.2.2. Type of change

WERs only reflect the net result of the changes in recognition result. To gain more insight, we analysed the different types of changes that actually occur. Fig. 3 shows the different types of changes. Furthermore, the 'total net result' is shown, which is defined as the difference between the number of improvements and the number of deteriorations. The total net result is directly related to the WER reduction:

$$\text{reduction in WER} = \text{WER}_{\text{SSS}} - \text{WER}_{\text{MMM}} = 100\% \times \frac{\text{total net result}}{\text{total number of words}} \quad (2)$$

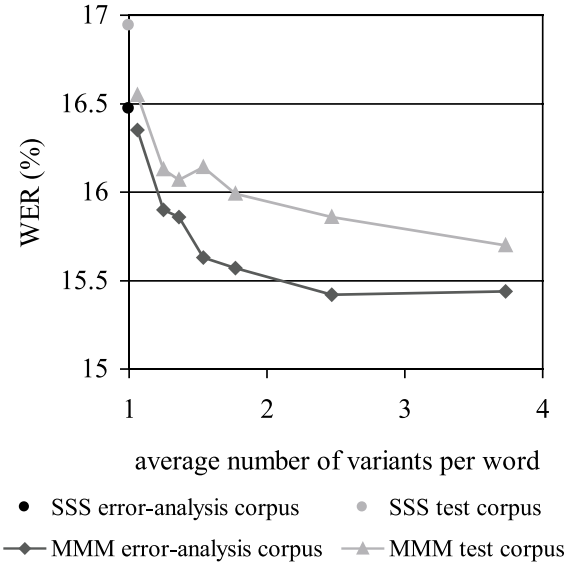

Fig. 2. WERs for testing condition MMM measured on the test and error analysis corpus.

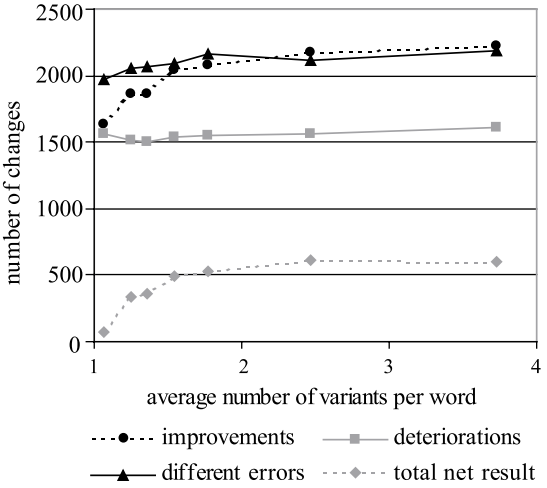

Fig. 3. Different types of changes for testing condition MMM compared to SSS measured on the error analysis corpus.

Fig. 3 shows that many changes occur, whereas the total net result or the WER reduction is very small. To give an example: For the set of 91 rules, 2219 words improve, 1613 deteriorate, and 2185 different errors occur. The improvements correspond to an absolute WER reduction of 3.8%, and the deteriorations to an increase in WER of 2.8%. The total net result or the WER reduction is (3.8% − 2.8% =) 1%. These results show that it should be possible to obtain a larger gain in recognition performance if one could find a way to make the balance between solving and introducing errors more positive.

4.2.3. Category of change

The next step in the error analysis procedure is a further analysis of the total net result. This was done by dividing all changes into the two categories of changes: variant and no-variant. The net result for each category of changes was obtained by subtracting the number of deteriorations from the number of improvements for that category.

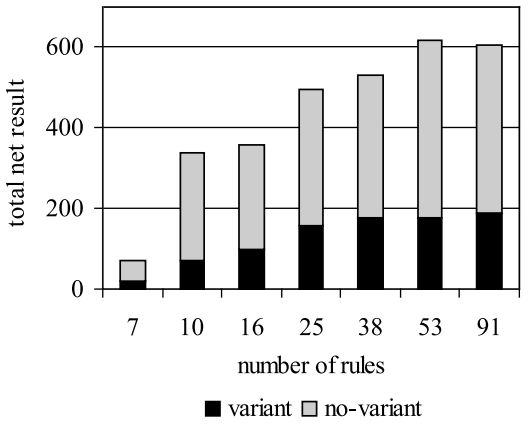

Fig. 4. Distribution of the total net result over the two categories of changes for all rule sets.

The distribution of the total net result over the two categories of changes is given in Fig. 4.

The category changes with the label variant (black bars in Fig. 4) contribute for 21–33% to the total net result. For this category of changes we can determine the contributions per rule, whereas this cannot be done for the no-variant category of changes. The net result of the category of changes with the label variant will be referred to as net result of variants in the rest of this paper. Fig. 5 shows the regression line between the net result of variants and the total net result. Such a strong correlation (0.98) indicates that the total net result (or WER, see (2)) can be predicted quite well on the basis of the net result of variants.

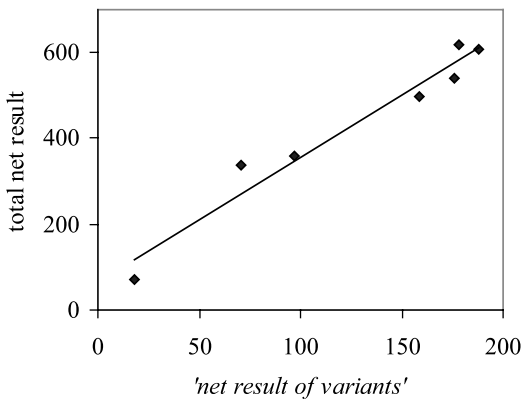

Fig. 5. Regression line for the correlation between net result of variants at rule set level and the total net result.

4.2.4. Contributions per rule

We further analysed the contributions of the different rules to the net result of variants. To this end, we took the changes that were labelled as variant. Next, we counted for each rule (in each of the 7 rule sets) how many deteriorations and improvements the rule caused. Finally, the net result per rule was determined by subtracting the number of deteriorations from the number of improvements.

Fig. 6 displays the number of improvements as a function of the number of deteriorations for each rule in each of the seven rule sets (240 data points). There exists a high correlation between the number of improvements and deteriorations caused by a specific rule (Pearson's correlation is 0.98). The regression line in Fig. 6 might give the impression that the high correlation between deteriorations and improvements is mainly determined by a small number of points, namely the six data-points in the right upper half of Fig. 6. This is not the case, since Pearson's correlation is still fairly high (0.77) if these six data-points are excluded. Fig. 6 also shows that, in general, more improvements are introduced than deteriorations, which means that the net result per rule is in general an improvement (thus a WER reduction, see (2)).

In Fig. 7, the contributions to the net result are plotted for each specific rule in each of the seven rule sets. In order to make it easier to interpret this figure, we only plotted the rules for which the absolute value of the net result is $\geqslant 5$ in one of the

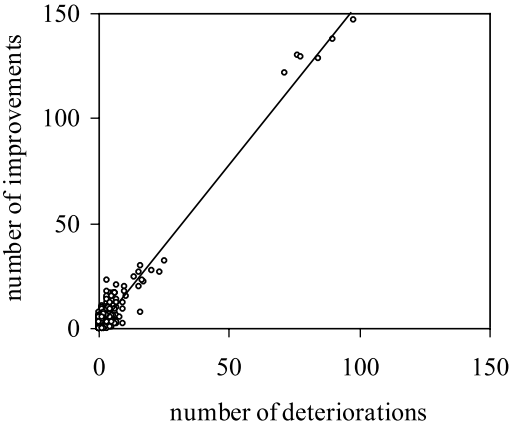

Fig. 6. Correlation between the number of improvements and the number of deteriorations.

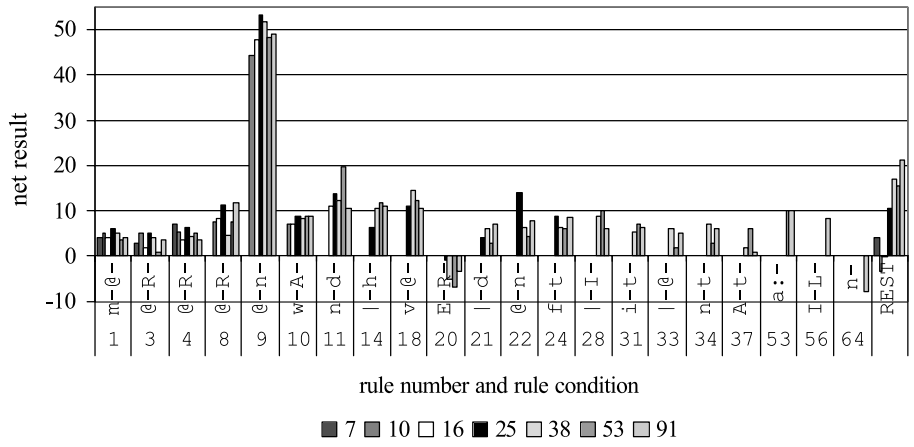

Fig. 7. Contributions of each individual rule to the net result.

rule sets (this was the case for 21 rules). On the horizontal axis, the rules are plotted together with the rule number and the context. On the vertical axis the change in net result is plotted ('+' = improvement, '−' = deterioration).

In Fig. 7, it can be observed that not all rules contribute equally to the net result as the total net result is mainly determined by about 1/4 of the rules (plotted in Fig. 7). Among these rules, rule 9 (@ n |) makes the largest positive contribution. Rules 20 (E R t) and 64 (n a: R) are the only rules that have a negative net result of more than 5 deteriorations.

4.3. Discussion and conclusions of error analysis

The error analysis that we performed in this study clearly has some advantages compared to the error analysis that we performed in our previous study (Kessens et al., 1999). The present error analysis revealed some differences and commonalities with the previous one, but also some new results. In our previous study we found that the results for the various rules tested in isolation cannot predict the results for the rules tested in combination. In this study, we tested different combinations of rules and for each rule we determined the contribution to the total net result. These results show that indeed the contribution in WER reduction per rule is different in each set of rules, but the differences are not very large. Three remarks concerning this apparent discrepancy in result have to be made. First of all, another study (Wester et al., 2000) revealed that the differences in sentence error rate (SER = number of incorrect sentences) for rules tested in isolation and in combination are corpus dependent. Second, one has to take into account that SERs/WERs cannot be simply added. Different rules can solve or introduce exactly the same errors when they are tested in isolation, whereas when the same rules are tested in combination, the error can be solved or introduced only once. Second, as we already mentioned in the previous study, interaction between pronunciation variants can occur, whereas this interaction is not possible when the rules are tested in isolation.

A commonality between the results of the two error analyses is that besides improvements, also deteriorations are introduced through the modeling of pronunciation variation. These deteriorations substantially eliminate the improvements, resulting in a small total net improvement in SER/WER. The results are also in line with the error analysis results of Ravishankar and Eskenazi (1997). These authors found that the number of errors corrected through modeling pronunciation variation are quite significant, but at the same time also new errors were introduced, substantially or completely balancing off the gains.

The current error analysis also revealed some new results. We found that about 1/3 of the WER reduction was obtained because a variant was recognized. For this category of changes we can di-

rectly determine which rules caused the changes. For the other 2/3 of the WER reduction we cannot directly determine which rules caused the changes. At rule set level, a high correlation was found between the net result of the category changes that were labelled as variant and the total net result (Pearson's correlation is 0.98). This finding is encouraging, since it suggests that the total recognition result can be predicted on the basis of the recognition result of the category of changes labelled as variant.

Furthermore, analysis of changes labelled as variant revealed that the contribution to the total net result differs per rule: In total, the net improvement was mainly determined by only 1/4 of the rules, the other 3/4 of the rules had a very small effect on the total net result. Furthermore, it turned out that the number of improvements and the number of deteriorations per individual rule are highly correlated. This result is somewhat disappointing, since it means that by leaving out a rule that causes many deteriorations, the number of improvements is also reduced. However, the positive message is that most of the time there are more improvements than deteriorations, which means that the total net result is an improvement.

Since the results of error-analysis indicate that the number of improvements and deteriorations are highly correlated, excluding rules that cause many deteriorations is not a solution for obtaining maximal WER reduction. The question that remains is what criteria are most suitable for selecting an optimal set of rules, since there is a practical constraint on the number of variants that can be included in the lexicon as decoding time is increased if the lexicon is expanded. This question will be addressed in the following section.

5. Criteria for optimal rule selection

5.1. The three selection criteria

In Section 4.2.2, we saw that the correlation between the net result of variants and the total net result at rule set level is very high (Pearson's correlation is 0.98). Since the total net result is directly related to the WER reduction (see (2)), this indicates that the net result of variants could be used to predict the WER reduction. For this reason, the first obvious criterion to select the rules seems to be net result of variants.

A disadvantage of using net result of variants as a selection criterion is that it is always necessary to perform error analysis to be able to select the optimal set of rules, while it would be better to have a measure that does not require the two extra steps of performing a recognition experiment and error analysis. We used two rule-related frequency measures, namely F_{rel} and F_{abs}, to select the rules (see Section 3.1.2). These two measures were determined directly from the DD transcriptions obtained during automatic extraction of the candidate rules (see step 3 described in Section 3.1.1). Since it is to be expected that the frequency of application of a rule is related to the WER reduction, we investigated the adequacy of the two frequency measures F_{abs} and F_{rel} as selection criteria for the rules.

We examined the adequacy of the three criteria in the following way: rules are selected on the basis of different criteria and for each set of rules the WER is calculated. In Section 5.2, we first present the results of the recognition experiments. Subsequently, the relation between the WER reduction and each investigated criterion is presented. Next, in Section 5.3, we compare the results and we will draw conclusions on the adequacy of each criterion investigated.

5.2. Results

5.2.1. Recognition experiments

The net result of variants was determined on the basis of the recognition experiment carried out with all 91 rules (see Fig. 7 for the values of the net result of variants per rule). Rule selection was performed by including those rules for which the net result of variants was larger than the threshold value. First, we selected the rule with the largest net result (rule 9) and then we added rules by lowering the threshold for the net result. The following values of net result of variants were used as a threshold: 45, 10, 5, 1, 0, -1. To investigate the adequacy of F_{abs}, we composed different rule sets by varying the threshold for F_{abs}. The following

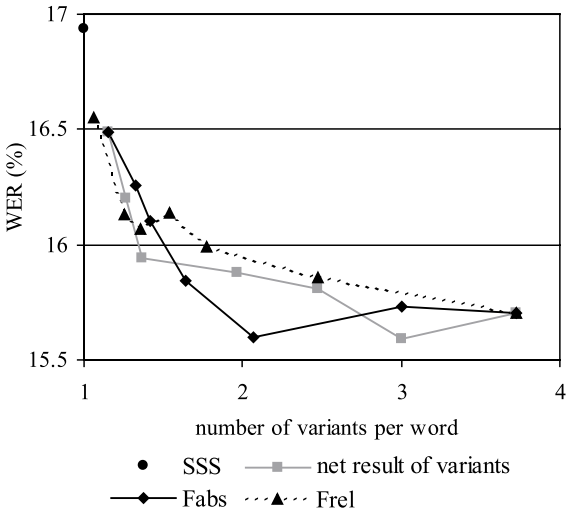

Fig. 8. WERs for rules selected on the basis of F_{abs}, F_{rel} and net result of variants.

values of F_{abs} were used as a threshold: 5000, 500, 400, 300, 200, 140. Since we already used F_{rel} as a selection criterion, we did not repeat the recognition experiments, and simply used the results reported in Section 3.3.

Fig. 8 presents the WERs measured for the rule sets obtained by selecting the rules on the basis of the three different selection criteria. It can be seen that for all selection criteria, apart from slight fluctuations, the WER decreases when the average number of variants per word is increased. The best results are obtained for $F_{abs} > 200$ (2 variants/word) and net result of variants >0 (3 variants/word): A reduction in WER of 1.4% absolutely, or 8% relatively is obtained (see Fig. 8).

5.2.2. Correlations at rule set level

The WER reduction was calculated by subtracting all the WERs plotted in Fig. 8 from the WER measured for the baseline (16.94%). In total, 19 rule sets were selected: 6 rule sets based on net result of variants, 6 rule sets based on F_{abs}, and 7 rule sets based on F_{rel}. For each of the 19 rule sets, the values of the three selection criteria at rule set level were determined in the following manner. The net result of variants at rule set level ('net result of variants$_{\text{-rule set}}$') was obtained by summing the net result of all rules in the set. F_{abs} at rule set level ($F_{\text{abs-rule set}}$) was obtained by summing the values of F_{abs} for all the rules in the set. F_{rel} at rule set level ($F_{\text{rel-rule set}}$) was obtained by dividing $F_{\text{abs-rule set}}$ by $F_{\text{cond-rule set}}$. F_{cond} (see Section 3.1.1, step 6) at rule set level ($F_{\text{cond-rule set}}$) was obtained by summing the values of F_{cond} for all the rules in the set.

Fig. 9 shows the values of the WER reduction and the corresponding measures at rule set level, together with the regression lines based on all 19 data points. In Fig. 9, '▲' indicates the rule sets selected based on net result of variants, '◆' indicates the rule sets that are selected based on F_{abs} and '■' indicates the rule sets selected based on F_{rel}. In Fig. 9, going from left to right means that the number of rules in the set is increased. The regression lines of all selection criteria show the trend that the WER reduction increases as the number of rules is increased.

In Fig. 9a, it can be seen that if net result of variants$_{\text{-rule set}}$ is increased, the WER reduction becomes larger, and the correlation is high (0.86). Fig. 9b shows that if $F_{\text{abs-rule set}}$ is increased, the WER reduction is also larger, and the correlation is even higher (0.93). The strong correlation between $F_{\text{abs-rule set}}$ and the WER reduction can be explained by the results that we found earlier. Error analysis revealed that the improvements and deteriorations per rule are highly correlated, but the net result is an improvement (see Fig. 6). This means that the more rules are used, and thus the higher $F_{\text{abs-rule set}}$, the larger is the total net improvement.

In Fig. 9c, it can be seen that the WER reduction is increased if $F_{\text{rel-rule set}}$ becomes smaller (Pearson's correlation is -0.83). This is against expectation, as one would expect the WER reduction to be larger if the relative frequency of application of the rules in the set is increased. A possible explanation for this result is that two criteria play a role: $F_{\text{rel-rule set}}$ and $F_{\text{abs-rule set}}$. Let us try to understand why F_{abs} is probably a better predictor of the WER reduction than F_{rel}. A specific value of F_{rel} could be the result of two completely different situations. To illustrate, an F_{rel} value of 50% could be obtained in the following two situations:

1. $F_{abs} = 1$ and $F_{cond} = 2$,
2. $F_{abs} = 10,000$ and $F_{cond} = 20,000$.

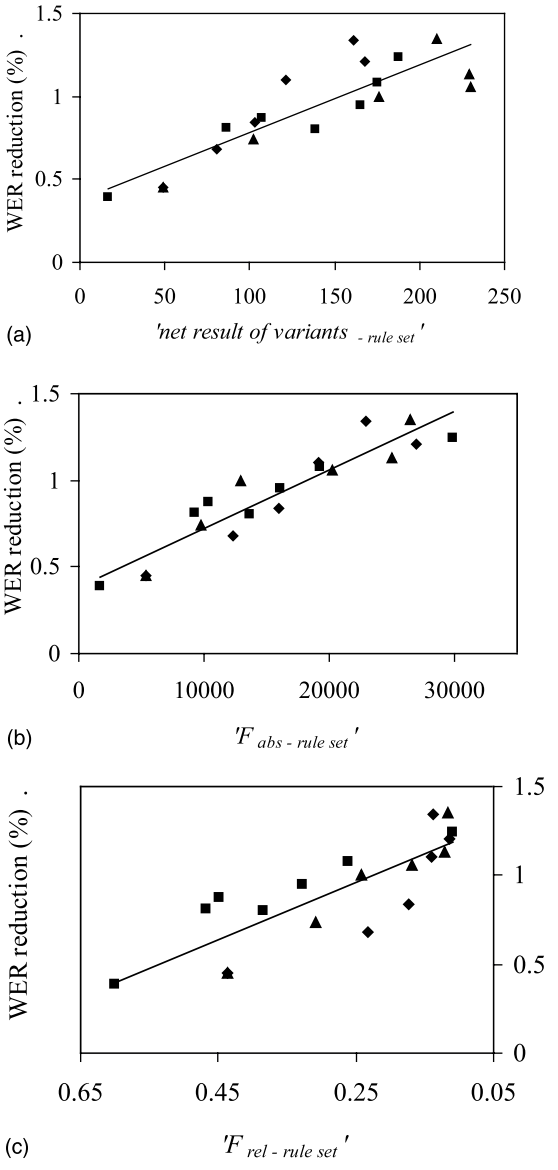

Fig. 9. Relation between (a) net result of variants$_{\text{-rule set}}$ and WER reduction, (b) $F_{\text{abs-rule set}}$ and WER reduction and (c) $F_{\text{rel-rule set}}$ and WER reduction.

It is easy to imagine that in relation to the total amount of material, situation 2 is bound to have a much greater effect on recognition performance than situation 1. While this difference clearly emerges from F_{abs}, it is completely blotted out in F_{rel}, which in turn explains why F_{abs} appeared to be a better predictor of the WER reduction.

5.3. Discussion and conclusions on rule selection criteria

Our results indicate that F_{abs} and net result of variants are better criteria for selecting the rules than F_{rel}. The question that remains is which of the two measures F_{abs} and net result of variants is the better criterion. Let us compare the results of the two criteria. First of all, the correlation with the WER reduction is higher for F_{abs} (0.93) than for net result of variants (0.86). Second, the net result of variants clearly has the disadvantage that it can only be used after performing a recognition experiment and carrying out an error analysis. F_{abs}, on the other hand, can be directly determined on the basis of the transcriptions used for automatic rule extraction. Third, for F_{abs} the optimal WER is obtained using an average of two variants/word in the lexicon, whereas three variants/word are needed to obtain optimal WER when net result of variants is used as a selection criterion (see Fig. 8). Since decoding time is correlated with the number of entries in the lexicon, this means that decoding time is shorter when the optimal rule set is obtained by selecting the rules on the basis of F_{abs} than on the basis of net result of variants. For all of these reasons, F_{abs} seems to be the most suitable criterion for rule selection.

6. General discussion

The results presented in this paper indicate that F_{abs} is an adequate predictor of recognition performance, and can therefore be used to select pronunciation rules. The question arises whether recognition performance could be further improved by using more rules. If indeed a linear relation exists between $F_{\text{abs-rule set}}$ and WER reduction, as plotted in Fig. 9a, then recognition performance could be further improved by increasing $F_{\text{abs-rule set}}$. Two remarks should be made about this point. The first concerns the linear relationship between $F_{\text{abs-rule set}}$ and the WER reduction. We expect that the relationship between $F_{\text{abs-rule set}}$ and WER reduction cannot be modelled by a simple straight line. For higher values of $F_{\text{abs-rule set}}$ we expect the straight line to flatten out. It might even be the case that recognition performance decreases for very high $F_{\text{abs-rule set}}$ values. A first

reason for expecting that the gain in recognition performance will be limited is that probably more unreliable rules are introduced by lowering the threshold for F_{abs}, as we expect that the rules based on transcription errors will have a low F_{abs}. A second reason is that, if the threshold for F_{abs} is lowered, the probabilities of the variants are estimated on the basis of smaller numbers, and the risk of not properly estimating the variant probabilities increases.

The second remark that should be made is that the relation between $F_{\text{abs-rule set}}$ and the average number of variants per word in the lexicon is not linear for our material, as is shown in Fig. 10. As a consequence, although we have indications that including more variants (by lowering the threshold for F_{abs}) can lower the WER, we know that the gain in performance will cost more and more in terms of decoding time.

For all these reasons, only a limited further improvement in recognition performance can be expected. The optimal value of F_{abs} will clearly be database and language specific, and for this reason, information concerning the values of F_{abs} can probably not be generalized to other contexts. In this connection, it would be interesting to devise a relative measure that can be more easily interpreted in other situations. Examples of such measures are: F_{abs} divided by the total number of deleted phones (e.g. for $F_{abs} > 100$, this measure would have the value 0.51), F_{abs} divided by the total number of phones (e.g. for $F_{abs} > 100$, this measure would have the value 0.04). An interesting research question would be to investigate whether more general conclusions can be drawn on the basis of this kind of relative measures by calculating them for different kinds of speech material, and comparing the values to each other.

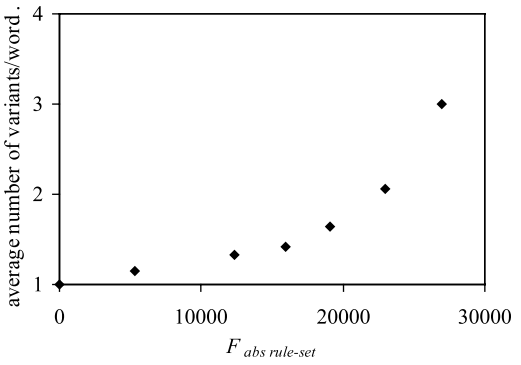

Fig. 10. Relation between $F_{\text{abs-rule set}}$ and the average number of variants per word in the lexicon. $F_{\text{abs-rule set}} = 0$ corresponds to the baseline system (SSS).

7. General conclusions

As mentioned in the introduction, the aim of the present paper was threefold. First, we analysed whether the DD method of modeling pronunciation we adopted does indeed lead to improvements in recognition performance. Since we found a total, statistically significant improvement of 1.4% WER absolutely, or 8% relatively for the best testing condition compared to the baseline testing condition, we conclude that the DD method of modeling pronunciation we adopted does indeed lead to improvements in recognition performance. Furthermore, we conclude that in order to ensure improvements in recognition performance, prior probabilities of the pronunciation variants need to be incorporated in the decoding process.

The second goal was to determine how exactly the reduction in WER came about. We found that besides improvements, also deteriorations were introduced through the modeling of pronunciation variation. These deteriorations substantially balance off the improvements, resulting in a small total net improvement in WER. These results show that it should be possible to obtain a larger gain in recognition performance if one could find a way to make the balance between solving and introducing errors more positive. Furthermore, we showed that about 1/3 of the WER reduction can be directly assigned to the rules, since the recognized words are variants, whereas for the other 2/3 of the changes, we could not determine which rule caused the change. However, since we found a high correlation between the number of changes labelled as variant and the total number of changes, it might be possible to predict the WER reduction on the basis of the changes labeled as variant. For this category of changes, the contribution to the net result differs per rule. Unfortunately, the number of improvements and the number of deteriorations per rule are highly correlated, but the positive message is that the net result per rule is, in general, an improvement.

Finally, the third goal was to find criteria that could be used for optimal rule selection. On the basis of our results, F_{abs} seems to be a more suitable criterion for optimal rule selection than F_{rel} and net result of variants.

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 authors would like to thank several members of the research group A^2RT and three anonymous reviewers for their useful comments on previous versions of this paper.

Appendix A

Statistics of the 91 selected rules, ordered according to descending F_{rel}

	Context	F_{rel}	F_{abs}		Context	F_{rel}	F_{abs}		Context	F_{rel}	F_{abs}
1	m @ r	0.88	225	31	i t |	0.18	416	61	e:n |	0.08	106
2	n d I	0.66	174	32	| n i	0.18	442	62	w I L	0.08	404
3	@ R m	0.61	272	33	| @ t	0.18	102	63	n d A	0.07	118
4	@ R t	0.57	638	34	n t s	0.17	165	64	n a:R	0.07	678
5	@ n v	0.53	131	35	t A S	0.16	186	65	o:R |	0.07	101
6	A L s	0.53	110	36	| w E	0.15	196	66	O R x	0.07	145
7	@ R b	0.51	151	37	A t |	0.15	310	67	| O m	0.07	300
8	@ R d	0.48	2031	38	m a:R	0.15	117	68	s E n	0.07	136
9	@ n |	0.43	5339	39	s t A	0.14	173	69	x @ n	0.07	328
10	w A R	0.42	234	40	p t |	0.14	118	70	a:x |	0.06	237
11	n d @	0.34	417	41	| r O	0.14	175	71	i n |	0.06	187
12	x @ v	0.34	109	42	x t |	0.13	498	72	E n t	0.06	118
13	@ R s	0.33	158	43	n t @	0.13	187	73	| d A	0.06	276
14	| h E	0.32	266	44	R t |	0.13	209	74	y R |	0.06	490
15	r y w	0.31	147	45	E n |	0.13	310	75	| O p	0.06	123
16	d @ r	0.30	333	46	v @ n	0.12	212	76	| I k	0.06	390
17	s t @	0.29	777	47	n t |	0.11	128	77	d A N	0.06	159
18	v @ r	0.28	555	48	w I n	0.11	149	78	a:L |	0.05	108
19	R n |	0.27	131	49	n I N	0.11	124	79	v A n	0.05	463
20	E R t	0.27	272	50	t @ x	0.11	221	80	| w I	0.04	233
21	| d @	0.26	205	51	| s t	0.10	147	81	| v A	0.04	370
22	@ n t	0.25	528	52	o:n I	0.10	104	82	| n a:	0.04	379
23	@ n s	0.23	106	53	a:R |	0.10	1089	83	N k |	0.04	106
24	f t |	0.22	235	54	O n |	0.09	117	84	| d E	0.04	129
25	| h u	0.22	156	55	A n |	0.09	736	85	A N k	0.04	108
26	R d @	0.19	137	56	I L |	0.09	481	86	A x |	0.03	130
27	@ R |	0.19	244	57	d A t	0.09	160	87	O m |	0.03	130
28	| I s	0.19	186	58	t @ r	0.09	378	88	I k |	0.03	199
29	d @ |	0.19	317	59	R x @	0.08	177	89	d A x	0.03	142
30	t w I	0.18	226	60	@ x |	0.08	194	90	| n e:	0.01	155
								91	j a: |	0.01	150

In the column 'context', the rule context is given (/L F $R/_{can}$, see Section 3.1.1 step 5). Furthermore, the relative (F_{rel}) and absolute (F_{abs}) frequencies of rule application are given for each rule.

References

Amdal, I., Korkmazskiy, F., Surendran, A.C., 2000. Joint pronunciation modeling of non-native speakers using data-driven methods. In: Proc. ICSLP00, Beijing, China, 16–20 October 2000, Vol. 3. pp. 622–625.

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

Cremelie, N., Martens, J.-P., 1999. In search of better pronunciation models for speech recognition. Speech Comm. 29, 115–136.

Fosler-Lussier, E., 1999. Dynamic Pronunciation Models for Automatic Speech Recognition, Ph.D. thesis, ICSI, University of California, Berkeley, USA.

Fukada, T., Yoshimura, T., Sagisaka, Y., 1999. Automatic generation of multiple pronunciations based on neural networks. Speech Comm. 27, 63–73.

Holter, T., Svendsen, T., 1999. Maximum likelihood modelling of pronunciation variation. Speech Comm. 29, 177–191.

Kerkhoff, J., Rietveld, T., 1994. Prosody in Niros with Fonpars and Alfeios. In: Proc. Dept. of Language & Speech, University of Nijmegen, Vol. 18. pp. 107–119.

Kessens, J.M., Wester, M., Strik, H., 1999. Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation. Speech Comm. 29, 193–207.

Kessens, J.M., Wester, M., Strik, H., 2000. Automatic detection and verification of dutch phonological rules. In: PHONUS5 Proc. "Workshop on Phonetics and Phonology in ASR", Saarbrücken: Institute of Phonetics, University of the Saarland, December 2000, pp. 117–128.

Lehtinen, G., Safra, S., 1998. Generation and selection of pronunciation variants for a flexible word recognizer. In: Proc. ESCA Workshop "Modeling Pronunciation Variation for Automatic Speech Recognition", Rolduc, Kerkrade, The Netherlands, 4–6 May 1998, A^{2}RT, University of Nijmegen, pp. 67–72.

Ravishankar, M., Eskenazi, M., 1997. Automatic generation of context-dependent pronunciations. In: Proc. Eurospeech'97, Rhodes, Greece, 22–25 September 1997, Vol. 5. pp. 467–470.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, J., McDonough, J., Nock, H., Saraçlar, M., Wooters, C., Zavaliagkos, G., 1999. Stochastic pronunciation modeling from hand-labelled phonetic corpora. Speech Comm. 29, 209–224.

Schiel, F., Kipp, A., Tillmann, H.G., 1998. Statistical modeling of pronunciation: it's not the model, it's the data. In: Proc. ESCA Workshop "Modeling Pronunciation Variation for Automatic Speech Recognition", Rolduc, Kerkrade, The Netherlands, 4–6 May 1998, A^{2}RT, University of Nijmegen, pp. 131–136.

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: Proc. ESCA Third European Conference on Speech Communication and Technology: Eurospeech'93, Berlin, pp. 2125–2128.

Strik, H., 2001. Pronunciation adaptation at the lexical level. In: Proc. ITRW Adaptation Methods for Speech Recognition, Sophia-Antopolis, France, pp. 123–130.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. Speech Comm., 225–246.

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.

Wester, M., Fosler-Lussier, E., 2000. A comparison of data-derived and knowledge-based modeling of pronunciation variation. In: Proc. ICSLP00, Beijing, China, 16–20 October 2000, Vol. 4. pp. 270–273.

Wester, M., Kessens, J.M., Strik, H., 1998. Improving the performance of a Dutch CSR by modeling pronunciation variation. In: Proc. ESCA Workshop "Modeling Pronunciation Variation for Automatic Speech Recognition", Rolduc, Kerkrade, The Netherlands, 4–6 May 1998, A^{2}RT, University of Nijmegen, pp. 145–150.

Wester, M., Kessens, J.M., Strik, H., 2000. Pronunciation variation in ASR: which variation to model? In: Proc. ICSLP00, Beijing, China, 16–20 October 2000, Vol. 4. pp. 488–491.

Wester, M., Kessens, J.M., Cucchiarini, C., Strik, H., 2001. Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer. Language & Speech 44 (3), 377–403.

Wiliams, G., 1999. Knowing what you don't know: roles for confidence measures in automatic speech recognition, Ph.D. thesis, Department of Computer Sciences, University of Sheffield, Sheffield, United Kingdom.

Williams, G., Renals, S., 1998. Confidence measures for evaluating pronunciation models. In: Proc. ESCA Workshop "Modeling Pronunciation Variation for Automatic Speech Recognition", Rolduc, Kerkrade, The Netherlands, 4–6 May 1998, A^{2}RT, University of Nijmegen. pp. 151–156.

Yang, Q., Martens, J.-P., 2000a. Data driven lexical modeling of pronunciation variation in ASR. In: Proc. ICSLP00, Beijing, China, 16–20 October 2000, Vol. 1. pp. 417–420.

Yang, Q., Martens, J.-P., 2000b. On the importance of exception and cross-word rules for the data-driven creation of Lexica for ASR. In: Proc. 11th ProRisc Workshop, 29 November–1 December, Veldhoven, The Netherlands. pp. 589–593.