# Pronunciation modeling for ASR -- knowledge-based and data-derived methods

ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Speech and Language 17 (2003) 69–85

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

# Pronunciation modeling for ASR – knowledge-based and data-derived methods

Mirjam Wester

*Department of Language and Speech, University of Nijmegen, Nijmegen, The Netherlands*

Received 23 August 2001; received in revised form 20 May 2002; accepted 22 July 2002

**Abstract**

This paper focuses on modeling pronunciation variation in two different ways: data-derived and knowledge-based. The knowledge-based approach consists of using phonological rules to generate variants. The data-derived approach consists of performing phone recognition, followed by smoothing using decision trees (D-trees) to alleviate some of the errors in the phone recognition. Using phonological rules led to a small improvement in WER; a data-derived approach in which the phone recognition was smoothed using D-trees prior to lexicon generation led to larger improvements compared to the baseline. The lexicon was employed in two different recognition systems: a hybrid HMM/ANN system and a HMM-based system, to ascertain whether pronunciation variation was truly being modeled. This proved to be the case as no significant differences were found between the results obtained with the two systems. A comparison between the knowledge-based and data-derived methods showed that 17% of variants generated by the phonological rules were also found using phone recognition, and this increases to 46% when the phone recognition output is smoothed by using D-trees.


## 1. Introduction

It is widely assumed that pronunciation variation is one of the factors which leads to less than optimal performance in automatic speech recognition (ASR) systems. Therefore, in the last few decades, effort has been put into finding solutions to deal with the difficulties linked to pronunciation variation. "Pronunciation variation" as a term could be used to describe most of the variation present in speech. However, this paper does not deal with the full scope of pronunciation

*E-mail address:* m.wester@let.kun.nl.

0885-2308/03/$ - see front matter 
doi:10.1016/S0885-2308(02)00030-X

variation, but rather with pronunciation variation that becomes apparent in a careful broad phonetic (phonemic) transcription of the speech, in the form of insertions, deletions or substitutions of phones relative to the canonical transcription of the words. This type of pronunciation variation can be said to occur at the segmental level.

Although it is assumed that pronunciation variation, in general, constitutes a problem for ASR, one may wonder if this assumption is correct, and whether modeling pronunciation variation at the segmental level has anything to offer towards the improvement of ASR performance. Studies by McAllaster et al. (1998) and Saraçlar et al. (2000) have shown that large improvements are feasible, if there is a match between the acoustic models used during recognition and the transcriptions in the lexicon. In other words, these experiments show that substantial improvements through pronunciation modeling are possible *in principle*.

In McAllaster et al. (1998) simulation experiments were carried out to determine the effect on recognition performance if all of the pronunciation variants encountered by the decoder were in fact contained in the lexicon. The simulation experiments show that when the data complies perfectly with the probability assumptions of the model (achieved by fabricating the data on the basis of the models) the WER drops from ca. 40% to less than 5%.

In Saraçlar et al. (2000) cheating experiments were conducted by carrying out an unconstrained phone recognition on the *test* speech. The phone string that resulted from this phone recognition was aligned with the reference word transcriptions for the test set and the *observed* pronunciation of each word in the test set was extracted. Next, the pronunciation dictionary was modified individually for each test utterance by including only the *observed* pronunciations for each of the words in the utterance. Using the modified lexicon to rescore a lattice obtained with the baseline ASR system led to a relative improvement of 43% in WER. Both these studies show that the performance can improve substantially if there is a close match between the acoustic models and the transcriptions, in other words, knowing the correct pronunciations can result in large gains.

Thus, it seems that the problem of modeling pronunciation variation lies in accurately predicting the word pronunciations that occur in the test material. In order to achieve this, the pronunciation variants must first be obtained in some way or other. Approaches that have been taken to modeling pronunciation variation can be roughly divided into pronunciation variants derived from a corpus of pronunciation data or from pre-specified phonological rules based on linguistic knowledge (Strik and Cucchiarini, 1999). Both have their pros and cons. For instance, the information from linguistic literature is not exhaustive; many processes that occur in real speech are yet to be described. On the other hand, the problem with an approach that employs data to obtain information is that it is extremely difficult to extract *reliable* information from the data.

Irrespective of how the pronunciations are obtained, choices must be made as to which variants to include in the lexicon, and/or to incorporate at other stages of the recognition process. Simply adding pronunciations en masse is futile; it is all too easy to increase the word error rates (WERs). It has been shown in many studies that simply adding variants to the lexicon does not lead to improvements, and in many cases even causes deteriorations in WER. For instance, in the studies of Yang and Martens (2000) and Kessens et al. (2001) it was shown that when the average number of variants per word in the lexicon exceeds roughly 2.5, the system with variants starts performing worse than the baseline system without multiple variants. Predicting which pronunciations will be the correct ones for recognition goes hand in hand with dealing with confusability in the lexicon, which increases when variants are added.

In a data-derived approach, a great deal of confusability is introduced by errors in automatic phonemic transcriptions. These phonemic transcriptions are used as the information source from which new variants are derived; consequently incorrect variants may be created. One commonly used procedure to alleviate this is to smooth the phonemic transcriptions – whether provided by linguists (Riley et al., 1999) or phone recognition (Fosler-Lussier, 1999) – by using decision trees to limit the observed pronunciation variation. Other approaches combat confusability by rejecting variants that are highly confusable on the basis of phoneme confusability matrices (Sloboda and Waibel, 1996; Torre et al., 1996). In Holter and Svendsen (1999), a maximum likelihood criterion was used to decide which variants to include in the lexicon. Amdall et al. (2000) propose log-likelihood-based rule pruning to limit confusability. Measures such as absolute or relative frequency of occurrence have also been employed to select rules or variants (Cremelie and Martens, 1999; Kessens et al., 2001). Finally, confidence measures have been employed to combat confusability by augmenting a lexicon with variants using a confidence-based evaluation of potential variants (Fosler-Lussier and Williams, 1999; Williams and Renals, 1998).

In this work, in addition to smoothing the phone transcriptions by using decision trees, a metric is employed that calculates the confusability in a lexicon, given a set of training data. The purpose of this metric, which was first introduced in Wester and Fosler-Lussier (2000), is to measure the inherent confusability of a lexicon. Pronunciation variation modeling would benefit greatly from some kind of metric that could quantify the effect of adding a new variant for instance.This metric is a first step in this direction. In addition, it is used as a selection criterion to decide which variants to include in a lexicon and which to exclude on the basis of the degree of confusability caused by a certain word.

Besides the issue of confusability, we are interested in ascertaining the merit of a data-derived approach in other terms than WER reductions. The question that arises especially in the case of a data-derived approach is whether pronunciation variation is indeed being modeled or if the system is merely being tuned to its own idiosyncrasies. In a data-derived approach, the risk of a great deal of circularity exists: a certain recognizer is used to carry out phone recognition, the output of the phone recognition is subsequently used to generate variants, and then the same recognizer is used to test whether incorporating the variants in the recognition process leads to an improvement in WER.

The issue whether pronunciation variation is truly being modeled by the data-derived approach is approached from two different angles. First of all, the two main methods described in the pronunciation modeling literature are used to model pronunciation variation. A direct comparison between a lexicon generated using a knowledge-based approach and a lexicon generated using a data-derived approach sheds light on which (if any) variation is modeled by both approaches. Secondly, two different recognition systems were employed to compare the effect of one and the same lexicon in two different recognition systems: a hybrid ANN/HMM system and an HMM recognition system. By using the same data-derived lexicon in two different recognition systems it is possible to determine whether pronunciation variation is modeled or whether by using a data-derived approach the system is merely being made to model its own idiosyncrasies.

The rest of this paper is as follows. First, the type of speech material is described, followed by a short description of the continuous speech recognizers that were employed. Next, a description is given of how the various lexica pertaining to pronunciation modeling are created: the knowledge-based approach to generating new pronunciations and the data-derived approach to

pronunciation modeling. In Section 4, an extended description of the confusability metric, proposed in Wester and Fosler-Lussier (2000), is given. This is followed by the results of recognition experiments employing the different pronunciation lexica. In Section 6, comparisons are made as to which variants overlap in the different lexica. Finally, we end by discussing the implications of our results and shortly summarizing the most important findings of this research.

## 2. Material and recognizers

### *2.1. Speech material*

In this study, we focus on segmental (phonemic) variation within VIOS (Strik et al., 1997), a Dutch corpus composed of human-machine "dialogues" in the domain of train timetable information, conducted over the telephone. Our training and test material, selected from the VIOS database, consisted of 25,104 utterances (81,090 words) and 6267 utterances (20,489 words), respectively. This corresponds to 3531 dialogues, with a total duration of 10h48 speech (13h12 silence), consisting of approximately 60% male and 40% female speakers. Recordings with a high level of background noise were excluded.

Fig. 1 shows the cumulative frequency of occurrence of the words in the VIOS training material as a function of word frequency rank. This figure shows that 82% of the training material is covered by the 100 most frequently occurring words. In total, 1104 unique words occur in the training material. The 14 most frequently observed words are all one syllable long and cover 48%

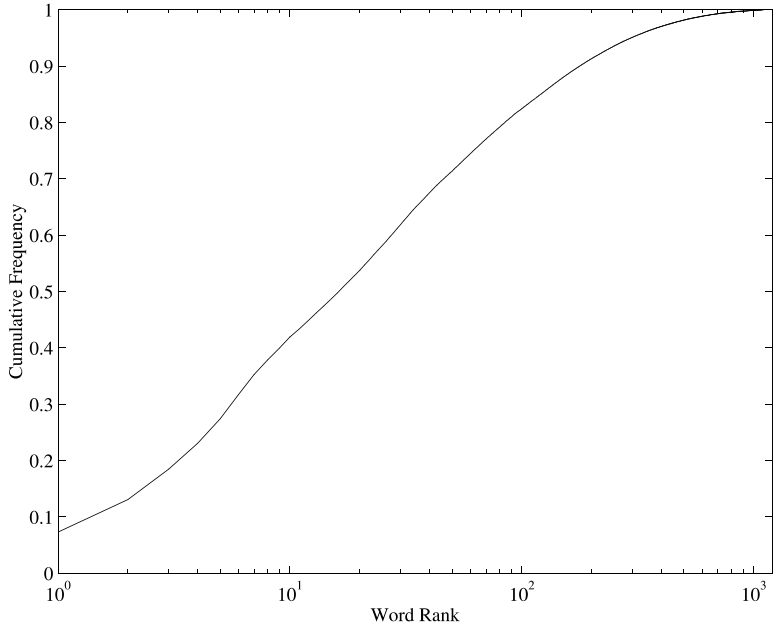


Fig. 1. Cumulative frequency of occurrence as a function of word frequency rank for the words in the VIOS training material.

of the training material. These findings are similar to results which were found for the Switchboard data (Greenberg, 1999). Furthermore, as the VIOS corpus comprises data collected from a train timetable information system, 43% of the words in the lexicon concern station names, which corresponds to 16% of the words in the training material.

### 2.2. CSRs

As was mentioned in Section 1, we are interested in comparing the effect of modeling pronunciation variation using two different recognizers, in order to find out if the results obtained with one system can be reproduced by another system, or if the results are possibly system dependent. To this end, two recognition systems were employed: the ICSI[1] hybrid ANN/HMM speech recognition system (Bourlard and Morgan, 1993) and the Phicos recognition system (Steinbiss et al., 1993). The main difference between the two CSRs is that in the ICSI system acoustic probabilities are estimated by a neural network instead of by mixtures of Gaussians, as is the case in the Phicos system.

The shared characteristics are the choice of phonemes, used to describe the continuous acoustic stream in terms of discrete units, and the language models that were employed. In both systems, 37 phonemes were employed. For the phonemes /l/ and /r/ a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/)[2]. The other 33 phonemes were context-independent. Models for non-speech sounds and silence were also incorporated in the two CSR systems. The systems use word-based unigram and bigram language models.

The lexicon is the same in both systems, in the sense that it contains the orthography of the words and phone transcriptions for the pronunciations. However, it is different in the sense that the ICSI lexicon contains prior probabilities for the variants of the words, whereas the Phicos lexicon does not. In the ICSI lexicon the prior probabilities are distributed over all variants for a word and add up to one for each word. In order to incorporate prior probabilities for variants in the Phicos system, the language model is based on the probability of the variants instead of on the words (cf. Kessens et al. (1999)).

In the Phicos recognition system Steinbiss et al. (1993), continuous density hidden Markov models (HMMs) with 32 Gaussians per state are used. Each HMM consists of six states, three parts of two identical states, one of which can be skipped. The neural network in the ICSI hybrid HMM/ANN speech recognition system (Bourlard and Morgan, 1993) was bootstrapped using segmentations of the training material obtained with the Phicos system. These segmentations were obtained by performing a Viterbi alignment using the baseline lexicon (Section 3.1) and Phicos baseline acoustic models, i.e., no pronunciation variation had been explicitly modeled. The neural net takes acoustic features plus additional context from eight surrounding frames of features at the input, and outputs phoneme posterior probability estimates. The neural network has a hidden layer size of 1000 units and the same network was employed in all experiments.

The feature descriptions which are used in the experiments are 12th-order PLP features and energy for the ICSI system. In the Phicos system, the first and second derivatives are added to the

[1] International Computer Science Institute.

[2] SAMPA-notation is used throughout this paper. Available from http://www.phon.ucl.ac.uk/home/sampa/dutch.htm.

feature vectors to ensure that the amount of context information that is employed in the two systems is as equal as possible. Because the ICSI system takes into account context frames, the Phicos system is given first and second derivatives, which roughly corresponds to eight surrounding frames, thus accounting for the extra context information.

## 3. Lexica generation

Using a knowledge-based approach and a data-derived approach to pronunciation modeling, we generated a number of new lexica. In all the new lexica, pronunciation variants were *added* to the baseline lexicon (Section 3.1). Section 3.2 describes the linguistically motivated approach to modeling pronunciation variation, followed by an explanation of how we derived pronunciations from the data in Section 3.3.

### *3.1. Baseline lexicon*

The baseline lexicon comprises 1198 words and contains *one* variant per word. The transcriptions were obtained using the transcription module of a Dutch Text-to-Speech system (Kerkhoff and Rietveld, 1994), which looks up the words in two lexica: CELEX (Baayen, 1991) and ONOMASTICA, which was used specifically for station names (Quazza and van den Heuvel, 2000). For those words for which no transcription was available a grapheme-to-phoneme converter was used, and all transcriptions were manually checked and corrected when necessary. In the ICSI baseline lexicon all prior probabilities are equal to one, as there is only one variant per word. The Phicos lexicon does not contain prior probabilities.

### *3.2. Knowledge-based lexicon*

In a knowledge-based approach, the information about pronunciations is derived from knowledge sources, for instance hand-crafted dictionaries or the linguistic literature. In this study, we selected five phonological processes, which are described in the literature, to formulate rules with which pronunciation variants were generated. The rules are context dependent and are applied to the words in the baseline lexicon. The resulting variants are unconditionally added to the lexicon. Table 1 shows the five phonological rules and their application contexts. For a more detailed description of the phonological rules, see Kessens et al. (1999).

Table 1
Phonological rules and context for application.

| Rule | Context for application |
|---|---|
| /n/-deletion | n → ø/ @ _ # |
| /r/-deletion | r → ø/ [+vowel] _ [+consonant] |
| /t/-deletion | t → ø/ [+obstruent] _ [+consonant] |
| schwa-deletion | @ → ø/ [+obstruent] _ [+liquid][@] |
| schwa-insertion | ø→ @/ [+liquid] _ [−coronal] |

In the ICSI recognizer, each pronunciation is assigned a prior probability which is usually estimated from the frequency count of the pronunciations seen in the training corpus. However, for the knowledge-based approach the priors were not determined on the basis of the training data, but the probability mass was distributed evenly over all the pronunciations of a word. This was done in order to be able to make the comparison with the same lexicon used in Phicos as fair as possible (recall Phicos does not contain priors in the lexicon).

### 3.3. Data-derived lexicon

In a data-derived approach, the information used to develop the lexicon is in some way distilled from the training data. The raw information that is used for data-derived generation of lexica is obtained by performing phone recognition of the training material using the ICSI recognizer. In this type of recognition task, the lexicon does not contain words, but a list of 37 phones, and a *phone* bigram grammar is used to provide phonotactic constraints. The output is a sequence of phones; no word boundaries are included. To obtain word boundaries, the phone recognition output is aligned to the reference transcription which does contain word boundaries. The reference transcription is obtained by looking up the transcriptions of the words in the baseline lexicon. These alignments are used as the basic information for generating the data-derived lexica.

However, pronunciation variants obtained from phone transcriptions are at once too many and too few. Thus, one would want to derive some kind of "rules" from the data. The approach we use is based on the decision-tree (D-tree) pronunciation modeling approach developed by Riley and Ljolje (1996) and which has been used by many others in the field (Fosler-Lussier, 1999; Riley et al., 1999; Robinson et al., 2001; Saraçlar et al., 2000) for pronunciation modeling of read and spontaneous English.

D-trees are used to predict pronunciations based on the alignment between the reference transcription of the training material and a transcription obtained using phone recognition output. The D-trees are used to smooth the phone recognition output before generating a lexicon. We used the Weka package[3] (Witten and Frank, 2000) to generate relatively simple D-trees, only taking into account the left and right neighboring phone identity in order to match the type of contexts used in our knowledge-based phonological rules. According to Riley et al. (1999) most of the modeling gain for the pronunciation trees comes from the immediate $\pm 1$ phonemic context, lexical stress, and syllable boundary location information. Therefore, in a subsequent experiment we also added syllable position (onset, nucleus, coda) as a feature in designing the D-trees. We did not incorporate stress as work by van Kuijk and Boves (1999) showed that information contained in the abstract linguistic feature "lexical stress" deviates too much from realized stress patterns in Dutch data.

For each of the 37 phones a D-tree was built. The D-tree model is trying to predict:

$$P(realization|canonical,\ context) \tag{1}$$

by asking questions about the context. Using the distributions in the D-trees, finite state grammars (FSG) were built for the utterances in the training data. During this FSG construction,

[3] Weka is a Java-based collection of machine learning algorithms for solving real-world data mining problems. Available from http://www.cs.waikato.ac.nz/ml/weka/index.html.

transitions with a probability lower than 0.1 were disallowed. This results in fewer arcs in the FSG and consequently the possibility of creating spurious pronunciations is diminished. Subsequently, the FSG were realigned with the training data, and the resulting "smoothed" phone transcriptions were used to generate a new lexicon. For a more detailed description of this D-tree approach, see, for example Fosler-Lussier (1999).

Various lexica were generated using the techniques described above. In all cases the prior probabilities for the pronunciations were based on the combination of the phone recognition transcript and pronunciations in the baseline lexicon. The two "lexica" were merged as follows to generate prior probabilities:

$$P_{merged}(pron|word) = \frac{P_{ph.rec.}(pron|word) + P_{baseline}(pron|word)}{2}. \quad (2)$$

In the phone recognition lexicon, the probability of a pronunciation is estimated on the basis of the phone recognition transcript ($P_{ph.rec.}$). In the baseline lexicon the probability ($P_{baseline}$) of the pronunciation of a word is 1. Merging these two lexica according to Eq. (2) ensures that the baseline pronunciations are always present in the new lexicon and that the different lexica contain the same words. If the phone recognition output was taken as is, the result would be out-of-vocabulary words in the testing condition.

## 4. A measure of confusability

One of the problems that remains at the heart of every approach to modeling pronunciation variation is which variants to include in the lexicon and which to exclude. Some variants lead to improvements and others to deteriorations, and it is difficult to determine which will influence the WER most (Wester et al., 2000). Ideally, what one would want in designing a lexicon is being able to judge beforehand what the optimal set of variants will be for describing the variance in the corpus at the level of the different pronunciations. We took a step in this direction by creating a metric by which we could judge the confusability of individual variants, as well as the overall confusability of a lexicon, based on the lexicon containing variants and the training material (Wester and Fosler-Lussier, 2000).

The metric works as follows: first a forced alignment of the training data is carried out, using the pronunciations from the lexicon for which the confusability is to be determined. The forced alignment results in a phone transcription of the training material; it should be clear that the phone transcription depends on the variants contained in the lexicon and the acoustic signal. After the phone transcription is obtained, the set of variants that match any substring within the result of the forced alignment is calculated, producing a lattice of possible matching words. For example, in Fig. 2, we compute the forced alignment of the word sequence "ik wil de trein om uh" ("I would like to catch the train at uh") resulting in the phonemic string /I k w I L d @ t r Ei n O m @/. We can then find all variants in the lexicon that span any substrings, e.g., the word "wilde" ("wanted" or "wild") corresponding to the phone transcription /w I L d @/.

The confusability metric is calculated by adding up the number of words that correspond to each phone (as shown in Fig. 2 in the row marked "All confusions") divided by the total number of phones. Thus the score for this utterance would be: $\frac{29}{14} = 2.1$, as the total number of phones is

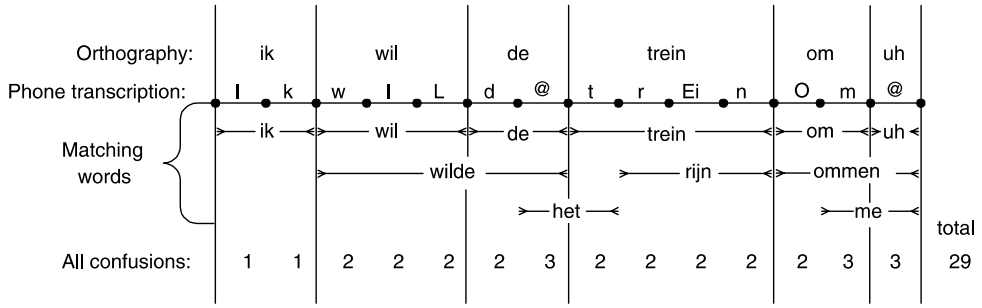


Fig. 2. Example of part of the lattice used to compute the average confusion.

14, and all confusions add up to 29. The average confusability for the lexicon is calculated by summing up the number of words that correspond to each phone in all utterances and dividing by the total number of phones in the training material.

In addition to the overall confusability of a lexicon given the training material, we were also interested in obtaining word level confusability scores in order to be able to discard highly confusable variants from the lexicon. The confusability count is defined as the number of times a variant of a certain word matches the phone transcription of a different word in the training material. In the experiments presented in Section 5.3, a variant of a word is discarded from the lexicon when its confusability count is $\geqslant 100$, unless the variant is the baseline variant.

## 5. Results

This section describes the results that were obtained using the various approaches to modeling pronunciation variation. The names of the lexica are indicated in the text and tables in *italics*. The tables show the word error rate (WER) results, the number of entries in the lexica (variants), the average number of variants per word (vars/word) and the confusability of the lexicon (conf), i.e., the average phone level confusion over all words in the training material. Results that differ significantly from the baseline result are indicated in bold. To establish significance a difference of proportions test was used, with a threshold of $p < 0.05$.

### 5.1. Knowledge-based approach

Phonological rules were used to generate variants, all of which were added to the baseline lexicon to create a new lexicon (*Phon_Rules*). In Kessens et al. (1999), we found that modeling pronunciation variation at all three levels in the recognizer, i.e. the lexicon, the language model and the phone models, led to the largest decrease in error rates within the Phicos recognition system using 14 MFCCs plus deltas. We repeated these experiments for the Phicos system using 12th-order PLP features, with their first and second derivatives, and energy. To discover whether including pronunciation variation at all three levels is also beneficial to the performance of the ICSI system, we incorporated pronunciation variation in the language model by adding probabilities for the pronunciation variants instead of for the words (*Phon_Rules+LM*) and retrained

the neural networks on the basis of a new alignment containing the pronunciation variants of the five phonological rules (*Phon_Rules* + *LM* + *PM*).

Recall that priors for the variants in the ICSI lexicon are all equal in these experiments to make the comparison with the Phicos system as fair as possible (Section 3.2). Just to make sure the ICSI system is not being penalized by this choice, we also measured the effect of estimating the priors on the training data. We found that using priors estimated on the basis of the training material led to the same WER as using a uniform distribution.

Table 2 shows the WER results for the ICSI and Phicos systems when five phonological rules are employed in the recognition system. These results show that modeling pronunciation variation using the five phonological rules has little effect on WERs in the ICSI system, whereas when linguistically motivated pronunciation variation is modeled at all three levels in the Phicos system an improvement is found at each step. On the basis of these results, we decided, for the ICSI system, not to include variants in the language model and not to retrain the neural nets in subsequent experiments. For the Phicos system, we continued carrying out pronunciation modeling at all three levels, as these results also indicate that the best result is obtained for the Phicos recognizer when pronunciation variation modeling is performed at all three levels.

### 5.2. Data-derived approach

A number of lexica were created using the data-derived approach (Section 3.3). First, a lexicon was generated on the basis of the "raw" phone recognition output (*Phone_Rec*). Next, a lexicon was generated using D-trees that were created using the phone recognition transcripts and a context consisting of left and right neighboring phones (*D-tree*); and finally a lexicon was created using D-trees which incorporated syllable information in addition to left and right neighboring phones (*D-tree_Syl*).

The *D-tree_Syl* lexicon was used to determine whether a data-derived lexicon generated within the ICSI system would lead to similar results when tested in the Phicos system. To ascertain the effect of priors for the variants, an experiment was carried out in which the priors in the lexicon were ignored during decoding (*D-tree_Syl (no priors)*). This situation is comparable to the Phicos testing condition in which variants are only added to the lexicon, but not included in the language model. Subsequently, for the Phicos system pronunciation variants were incorporated in the language model (*D-tree_Syl* + *LM*), which is comparable to (*D-tree_Syl* + *priors*) for ICSI. Finally, the phone models were retrained (*D-tree_Syl* + *LM* + *PM*) for the Phicos system, as

Table 2
Results for the baseline lexicon and lexica generated using the linguistic approach, for the ICSI and Phicos systems

| Lexicon | WER ICSI | WER Phicos | Variants | Vars/word | Conf |
|---|---|---|---|---|---|
| *Baseline* | 10.7 | 10.4 | 1198 | 1 | 1.5 |
| *Phon_Rules* | 10.5 | 10.3 | 2066 | 1.7 | 1.7 |
| *Phon_Rules* + *LM* | 10.6 | 10.2 | 2066 | 1.7 | 1.7 |
| *Phon_Rules* + *LM* + *PM* | 10.7 | 10.1 | 2066 | 1.7 | 1.7 |

WER, word error rate; variants, the number of entries in the lexica; vars/word, the average number of variants per word; and conf, the average phone level confusion for all words in training material.

previous experiments with Phicos have shown that the best way of incorporating pronunciation variation is to do it at all three levels. For the ICSI system, this last testing condition was not carried out.

Table 3 shows the WERs for the ICSI and Phicos systems using the different data-derived lexica. Adding all the variants from the raw phone recognition leads to a deterioration in performance. The deterioration is not as large as one might expect, but it should be kept in mind that the lexicon does not only contain variants from the phone recognition, because, like all other lexica, it was merged with the baseline lexicon and the priors for the baseline variants are higher than the priors for other variants. In any case, the decoding time does increase substantially, which is in line with expectations.

The results in Table 3 further show that modeling pronunciation variation using D-trees leads to a significant improvement in the ICSI system. A relative improvement of 7.5% compared to the baseline result is found. Including syllable information in the D-trees in addition to left and right neighboring phone identity does not further improve the performance.

Simply employing the *D-tree_Syl* lexicon in the Phicos system leads to a significant deterioration in WER compared to the baseline result. Ignoring the priors in the ICSI lexicon leads to a deterioration of the same magnitude. When the variants are added to the language models the performance of the Phicos system improves dramatically, although the improvement is not significant compared to the baseline result. Incorporating pronunciation variation in the recognition process by retraining the phone models leads to a slight degradation compared to only incorporating it in the language models. This is a slightly surprising result as in previous experiments retraining has always led to improvements in WER.

Inspection of the lexical confusability scores in Tables 2 and 3 show that the highest degree of confusability is clearly found in the phone recognition lexica; this is followed by the D-trees lexica, and the least amount of confusability is contained in the phonological rule lexica. However, there is no straightforward relationship between the confusability score and the WER performance. Consequently, it is not clear how the confusability score could be used to predict which lexicon is "better". In addition, there is no relationship between the number of entries in the lexicon (or the number of variants per word) and the WER. However, decoding time increases dramatically with a higher number of entries in the lexicon, which is an extra reason to sparingly add variants to the lexicon. In the following section, we employ the confusability metric to discard confusable variants instead of only measuring the confusability in a lexicon.

Table 3
Results for lexica generated using a data-derived approach, for the ICSI and Phicos systems

| Lexicon | WER ICSI | WER Phicos | Variants | Vars/word | Conf |
|---|---|---|---|---|---|
| *Baseline* | 10.7 | 10.4 | 1198 | 1 | 1.5 |
| *Phone_Rec* | 10.9 | – | 20347 | 17.7 | 65.9 |
| *D-tree* | **9.9** | – | 5880 | 4.9 | 9.3 |
| *D-tree_Syl (no priors)* | **17.0** | **17.0** | 5912 | 4.9 | 9.0 |
| *D-tree_Syl + priors/LM* | **9.9** | 10.0 | 5912 | 4.9 | 9.0 |
| *D-tree_Syl + LM + PM* | – | 10.3 | 5912 | 4.9 | 9.0 |

Table 4
Results of using confusability metric to remove variants from lexica for the ICSI system

| Lexicon | Without pruning | With pruning | | | |
|---|---|---|---|---|---|
| | WER ICSI | WER ICSI | Variants | Vars/word | Conf |
| *Phon_Rules Conf* $\geqslant 100$ | 10.5 | 10.5 | 2054 | 1.7 | 1.6 |
| *D-tree_Syl Conf* $\geqslant 100$ | **9.9** | **10.0** | 5474 | 4.6 | 2.1 |
| *Phone_Rec Conf* $\geqslant 100$ | 10.9 | **10.1** | 15424 | 12.9 | 3.2 |
| *Phone_Rec Conf* $\geqslant 0$ | 10.9 | **10.1** | 9222 | 7.7 | 1.7 |

### 5.3. Confusability measure for pruning

The confusability metric was used to prune variants with a confusability count of 100 or higher. For the phone recognition lexicon we also applied a threshold of 0. A threshold of 0 means all confusable variants are removed except for the baseline variants, as in all cases the baseline variants are present in the lexica. The pruning was applied to the lexica: *Phon_Rules*, *D-trees_Syl*, and *Phone_Rec*. Table 4, column 2 shows the original WERs for the ICSI system prior to pruning with the confusability metric. The remaining columns show results for lexica after pruning had been carried out.

For the *Phon_Rules* lexicon and the *D-tree_Syl* lexicon, pruning the most confusable variants has no effect on the WERs compared to the same testing condition without using the confusability metric to prune variants. This is in contrast to the results found for the "raw" phone recognition lexicon (*Phon_Rec Conf*), where using the confusability metric to prune the most confusable variants leads to a significant improvement.

The difference in number of variants present in the phone recognition lexica also deserves some attention. Even when the confusability count for confusable words is set to 0, the *Phone_Rec* lexicon contains almost twice as many variants as the *D_tree* lexicon. This is due to the fact that many of the variants that are generated on the basis of phone recognition are so different from pronunciations chosen during forced alignment that they do not form a match with any of the forced alignment transcriptions. Some other way of pruning these "strange" pronunciations should be employed, as they do not seem to affect the WERs, but they do increase decoding times. It may seem strange that the confusability score for *Phone_Rec Conf* $\geqslant 0$ is not 1.5 as it is for the *Baseline* lexicon, but this is due to the fact that after all the confusable variants have been removed, a forced alignment of the training data is carried out again using the new lexicon. As the set of variants is different, the alignments also turn out differently and consequently other variants may be confused with each other.

## 6. Analysis of lexica

An analysis was carried out to determine how much overlap there is between lexica generated using the phonological rule method for generating variants and the data-derived approaches to generating variants. The *Phon_Rules* lexicon was used as the starting point for the comparison of the different lexica. This lexicon was chosen because the variants generated by the five phonological rules are valid variants, from a linguistic point of view. From an ASR point of view, the

Table 5
Overlap between variants generated using five phonological rules which truly occur in the training material and variants generated using phone recognition or variants generated by the D-trees

| Rules | Lexicon | | | | |
|---|---|---|---|---|---|
| | *Phon_Rules* | *Phone_Rec* | | *D-tree* | |
| | #vars | #vars | % | #vars | % |
| /n/-deletion | 195 | 34 | 17 | 100 | 51 |
| /r/-deletion | 141 | 30 | 21 | 77 | 55 |
| /t/-deletion | 37 | 9 | 24 | 20 | 54 |
| schwa-deletion | 13 | 1 | 8 | 4 | 31 |
| schwa-insertion | 36 | 1 | 3 | 2 | 6 |
| combination | 68 | 10 | 15 | 23 | 34 |
| Total | 490 | 85 | 17 | 226 | 46 |

validity of the variants depends on whether the variants actually occur in the data. Therefore, we made comparisons using all variants generated by the phonological rules as a starting point, but only investigating those variants that actually occur in the training material. A forced alignment of the training material was carried out using the *Phon_Rules* lexicon to find out which variants actually occur

For each of the phonological rules (see Table 1) lists of variants were made. The extra category "combination" in Table 5 refers to the variants that are the result of more than one rule applying to a word. None of the variants were included in more than one list and baseline variants were not included. The overlap between the lexica was calculated by enumerating the variants (#vars) that occur in both the *Phon_Rules* and *Phone_Rec* lexicon, as well as in the *Phon_Rules* and the *D-tree* lexicon. The percentages indicate the proportion of variants in the *Phon_Rules* lexicon that is covered by the other lexica.

Table 5 shows that the D-tree approach "finds" about half of the variants that occur in the forced recognition when a lexicon with variants generated with the five phonological rules is used. Compared to the variants that are found as a result of the phone recognition (17%) this is a large gain. Besides underlining that the D-tree approach is generalizing beyond what is observed in the training material, the method also generalizes to true pronunciation variation. This is a clear indication that the data-derived method is not merely tuning the recognizer to it's own idiosyncrasies but it is indeed picking up on pronunciation variation.

## 7. Discussion

In this paper, we reported on two different approaches to dealing with pronunciation variation; a knowledge-based and data-derived approach. One of the issues we set out to address was to compare these two approaches to modeling pronunciation variation. The approaches differ in the way that information on pronunciation variation is obtained. The knowledge-based approach consists of generating variants by using phonological rules for Dutch. The data-derived approach consists of performing phone recognition to obtain information on the pronunciation variation in the data, followed by smoothing with D-trees to alleviate some of the *unreliable* data introduced

by shortcomings of the recognition system. Both approaches lead to improvements, but of differing magnitudes.

Improvements due to modeling pronunciation variation using phonological rules are reported in quite a number of studies (Cohen, 1989; Ferreiros and Pardo, 1999; Flach, 1995; Lamel and Adda, 1996; Safra et al., 1998; Wiseman and Downey, 1998) for different types of speech, different languages, and employing different CSR systems. Unfortunately, relating the findings in those studies to each other and to the results found in this work is exceedingly difficult because there are factors that may have influenced the findings, but which have not been described in the studies, or which have not been investigated individually. Furthermore, as was stated in Strik and Cucchiarini (1999): "It is wrong to take the change in WER as the only criterion for evaluation, because this change is dependent on at least three different factors: (1) the corpora, (2) the ASR system, and (3) the baseline system. This means that improvements in WER can be compared with each other only if in the methods under study these three elements were identical or at least similar". As there is not much else but WERs to go by it should be clear it is extremely difficult to compare the different studies with each other.

In Kessens et al. (1999) and this study, the exact same training and test data were used, the only difference is the sets of acoustic features that were used. In Kessens et al. (1999) 14 MFCCs plus deltas were used, and in this study 12th-order PLP features plus their first and second derivatives and energy was used. In contrast to the results in Kessens et al. (1999), a significant improvement using the knowledge-based approach in Phicos was not found in this study. The difference between the experiments carried out using Phicos is the acoustic features that were employed. In this study, the starting point WER is significantly lower than in Kessens et al. (1999). Our results show that even though the trends are the same, pronunciation modeling through phonological rules has less effect when the starting-point WER is lower. In this case, it seems that the mistakes that were previously solved by modeling pronunciation variation are now being taken care of by improved acoustic modeling. This type of effect is also found in Ma et al. (1998) and Holter and Svendsen (1999). However, there are examples in the literature that this does not necessarily need to be the case. For instance, Riley et al. (1999) reports that reductions in WER due to modeling pronunciation variation persist after the baseline systems are improved by coarticulation sensitive acoustic modeling and improved language modeling.

One of the disadvantages of using a knowledge-based approach, i.e. not all of the variation that occurs in spontaneous speech has been described, is in part alleviated by using a data-derived approach. The challenge that is introduced when a data-derived approach is taken, is that the information which is used to generate variants is not always reliable. Results pertaining to the data-derived approach showed that simply adding all the variants from the raw phone recognition leads to a deterioration in performance. However, when subsequently D-trees were used to smooth the phone recognition, significant improvements in the ICSI system were found. A relative improvement of 7.5% was found compared to the baseline result. This is similar to findings reported for English (e.g., Fosler-Lussier, 1999; Riley et al., 1999; Robinson et al., 2001; Saraçlar et al., 2000) in the sense that improvements are found when D-trees are used to model pronunciation variation.

One of the other questions we were interested in answering: "Is pronunciation variation indeed being modeled, or are idiosyncrasies of the system simply being modeled?" can be answered by considering the following. First of all, the similar results obtained using two quite different recognition systems indicates that pronunciation variation is indeed being modeled. Second,

analysis of the lexica showed that the D-trees are learning phonological rules. We found that 17% of variants generated by the phonological rules were also found using phone recognition, and this increased to 46% when the phone recognition output was smoothed by using D-trees. This is a further indication that pronunciation variation is indeed being modeled.

Confusability is intuitively an extremely important point to address in pronunciation modeling. The confusability metric which we introduced is useful as a method for pruning variants. The results show that simply pruning highly confusable variants from the phone recognition lexicon leads to a significant improvement compared to the baseline. In other words, the confusability metric is a very simple and easy way of obtaining a result which is comparable to the result obtained using methods such as phonological rules or D-trees. However, we also intended to use the confusability metric to assign a score to a lexicon which could then be used to predict how well a lexicon would perform. The results in Table 4 quite conclusively demonstrate that the confusability score is not suited for this purpose as different confusability scores lead to roughly the same WER scores. The metric should be extended to include substitutions of phones in the alignments and in some way language model probabilities should also be incorporated, possibly then the confusability metric will be a better estimator of the confusability of a lexicon, given a set of training data.

Many studies (e.g., Cohen, 1989; Ma et al., 1998; Yang and Martens, 2000) have found that probabilities of the variants (or probabilities of rules) play an important role in whether an approach to modeling pronunciation variation is successful or not. In this study, this was once again shown by comparing results between Phicos and the ICSI system in Section 5.2. Not including priors in the ICSI system and not incorporating variants in the language model for Phicos showed significant deteriorations, whereas including probabilities showed significant improvements over the baseline. Yet if we are to relate this to the findings of McAllaster et al. (1998) and Saraçlar et al. (2000): if one can accurately predict word pronunciations in a certain test utterance the performance should improve substantially, we must conclude that estimating the priors for a whole lexicon is not optimal. The point is that a good estimate of priors is probably a conditional probability, with speaker, speaking mode, speaking rate, subject, etc. as conditionals. Some of these factors can be dealt with in a two-pass scheme by rescoring $n$-best lists as the pronunciation models in Fosler-Lussier (1999) showed; however, the gains found in this study remain small as it is extremely difficult to accurately estimate the conditionals. One of the explanations given in Fosler-Lussier (1999) as to why including extra-segmental features did not improve recognition results was that these features were not robust enough for accurate prediction of pronunciation probabilities in an automatic learning system (Fosler-Lussier, 1999, p. 150). This is the crux of the matter. It is of the utmost importance, if we are to incorporate extra features into the process of pronunciation modeling, that these features are robust. Therefore, finding methods of robust estimation of, for example, speaking rate and word predictability, must also be included in future research within the field of pronunciation modeling.

## Acknowledgements

I extend my appreciation and gratitude to Eric Fosler-Lussier for all his help in working with the ICSI recognizer and for developing the confusability metric. I would also like to thank Johan de Veth, for his help incorporating PLP features into Phicos. Grateful appreciation is

extended to members of $A^2RT$ who gave useful comments on previous versions of this paper, especially, Helmer Strik, Loe Boves, Judith Kessens and Febe de Wet. Furthermore, this work would not have been possible without the hospitality of ICSI – where I was given the opportunity to visit for a number of months – and by a grant from the University of Nijmegen (Frye stipend) and a NWO travel scholarship.

## References

Amdall, I., Korkmazskiy, F., Surendran, A., 2000. Joint pronunciation modelling of non-native speakers using data-driven methods. In: Proc. ICSLP '00, Beijing, vol. III, pp. 622–625.

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

Bourlard, H., Morgan, N., 1993. Connectionist Speech Recognition: A Hybrid Approach. Kluwer Academic Publishers, Dodrecht.

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

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

Ferreiros, J., Pardo, J., 1999. Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations. Speech Commun. 29, 65–76.

Flach, G., 1995. Modelling pronunciation variability for special domains. In: Proc. EUROSPEECH '95, Madrid, pp. 1743–1746.

Fosler-Lussier, E., 1999. Dynamic pronunciation models for automatic speech recognition. Ph.D. thesis, University of California, Berkeley, CA.

Fosler-Lussier, E., Williams, G., 1999. Not just what, but also when: guided automatic pronunciation modeling for Broadcast News. In: DARPA Broadcast News Workshop, Herndon, VA., pp. 171–174.

Greenberg, S., 1999. Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation. Speech Commun. 29, 159–176.

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

Kerkhoff, J., Rietveld, T., 1994. Prosody in NIROS with FONPARS and ALFEIOS. In: de Haan, P., Oostdijk, N. (Eds.), Proc. Dept. Lang. Speech, Univ. Nijmegen, vol. 18, pp. 107–119.

Kessens, J., Cucchiarini, C., Strik, H., 2001. A data-driven method for modeling pronunciation variation. Speech Commun. (submitted).

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

Lamel, L., Adda, G., 1996. On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In: Proc. ICSLP '96, Philadelphia, PA, pp. 6–9.

Ma, K., Zavaliagkos, G., Iyer, R., 1998. Pronunciation modeling for large vocabulary conversational speech recognition. In: Proc. ICSLP '98, Sydney, pp. 2455–2458.

McAllaster, D., Gillick, L., Scattone, F., Newman, M., 1998. Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. In: Proc. ICSLP '98, Sydney, pp. 1847–1850.

Quazza, S., van den Heuvel, H., 2000. The use of lexicons in text-to-speech-systems. In: van Eynde, F., Gibbon, D. (Eds.), Lexicon Development for Speech and Language Processing. Kluwer Academic Publishers, Dodrecht, pp. 207–233 (Chapter 7).

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

Riley, M., Ljolje, A., 1996. Automatic generation of detailed pronunciation lexicons. In: Lee, C.-H., Soong, F., Paliwal, K. (Eds.), Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic Publishers, Dodrecht, pp. 285–302, Chapter 12.

Robinson, A., Cook, G., Ellis, D., Fosler-Lussier, E., Renals, S., Williams, D., 2001. Connectionist speech recognition of Broadcast News. Speech Commun. 37, 27–45.

Safra, S., Lehtinen, G., Huber, K., 1998. Modeling pronunciation variations and coarticulation with finite-state transducers in CSR. In: Proc. ESCA Workshop Model. Pronunciation Variation Autom. Speech Recognit., Kerkrade, pp. 125–130.

Saraçlar, M., Nock, H., Khudanpur, S., 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. Comput. Speech Lang. 14, 137–160.

Sloboda, T., Waibel, A., 1996. Dictionary learning for spontaneous speech recognition. In: Proc. ICSLP '96, Philadelphia, PA, pp. 2328–2331.

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. EUROSPEECH '93, Berlin, pp. 2125–2128.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. Speech Commun. 29, 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. Int. J. Speech Technol. 2 (2), 119–129.

Torre, D., Villarrubia, L., Hernández, L., Elvira, J., 1996. Automatic alternative transcription generation and vocabulary selection for flexible word recognizers. In: Proc. ICASSP '96, Munich, pp. 1463–1466.

van Kuijk, D., Boves, L., 1999. Acoustic characteristics of lexical stress in continuous telephone speech. Speech Commun. 27, 95–111.

Wester, M., Fosler-Lussier, E., 2000. A comparison of data-derived and knowledge-based modeling of pronunciation variation. In: Proc. ICSLP '00, Beijing, vol. I, pp. 270–273.

Wester, M., Kessens, J., Strik, H., 2000. Pronunciation variation in ASR: Which variation to model? In: Proc. ICSLP '00, Beijing, vol. IV, pp. 488–491.

Williams, G., Renals, S., 1998. Confidence measures for evaluating pronunciation models. In: Proc. ESCA Workshop Model. Pronunciation Variation Autom. Speech Recognit., Kerkrade, pp. 151–155.

Wiseman, R., Downey, S., 1998. Dymanic and static improvements to lexical baseforms. In: Proc. ESCA Workshop Model. Pronunciation Variation Autom. Speech Recognit., Kerkrade, pp. 157–162.

Witten, I., Frank, E., 2000. Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, Los Altos, CA.

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