Improving recognition performance by modelling pronunciation variation

Judith Kessens and Mirjam Wester
Department of Language & Speech
University of Nijmegen
E-mail: {kessens|wester}@let.kun.nl

Abstract

This paper describes a method for improving the performance of a continuous speech recognizer by modelling pronunciation variation. Although the improvements obtained with this method are small, they are in line with those reported by other authors. A series of experiments was carried out to model pronunciation variation. In the first set of experiments word internal pronunciation variation was modelled by applying a set of four phonological rules to the words in the lexicon. In the second set of experiments, variation across word boundaries was also modelled. The results obtained with both methods are presented in detail. Furthermore, statistics are given on the application of the four phonological rules on the training database. We will explain why the improvements obtained with this method are small and how we intend to increase the improvements in our future research.

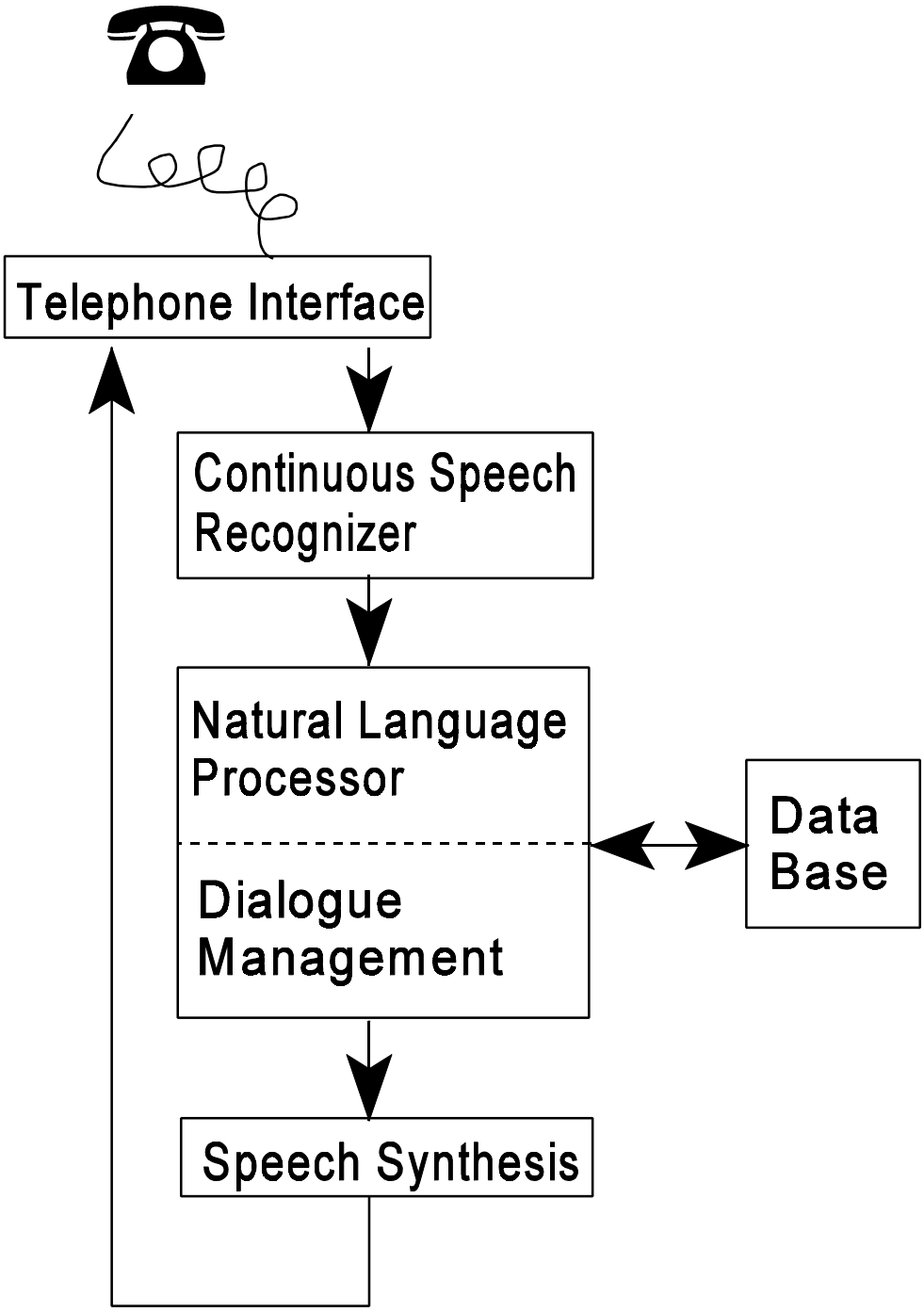

***Figure 1**: Architecture of Spoken Dialogue System*

Introduction

At the Department of Language and Speech at the University of Nijmegen, we are working on a Spoken Dialogue System (SDS) that will be employed to automate part of a public transport information service. This system was adapted from a German prototype (Steinbiss *et al.*, 1995) developed by Philips Research Labs (Aachen, Germany), and was further improved by means of a bootstrapping method (Strik *et al.* 1996 and 1997).

The architecture of the SDS is shown in Figure 1. The SDS consists of a telephone interface, a continuous speech recognizer (CSR), a natural language processor (NLP), a dialogue management (DM) module which is connected to a database, and a text-to-speech (TTS) synthesizer. The telephone interface is responsible for the interaction between the telephone network and the SDS. The incoming speech signal is converted into sequences of words by the CSR. The NLP searches for information in the sequences of recognized words, for example desired departure time or time of arrival. The DM module stores the information found by the NLP and checks whether or not it is complete. If information is missing, the system continues to ask questions until all necessary information is obtained. The DM looks up the answer in a database, formulates a reply in text form, and sends it on to the TTS synthesizer. The TTS synthesizer converts the text into a speech signal and passes this signal to the telephone interface, which in turn sends the message through the telephone line to the user.

In the present research, we are only concerned with the CSR component of the SDS. This CSR was gradually improved through a bootstrapping method, by adding more data. However, since a point was reached at which no further increase in performance could be obtained by adding more data, new methods of improving the system were sought. Given that the SDS is an interactive system the kind of speech callers use may be extremely varied. The manner in which people speak to a machine can vary from using very sloppy articulation to hyper articulation. Therefore, it is obvious to expect that the system's performance might be improved by modelling pronunciation variation.

Pronunciation variation can be divided into two kinds of variation. The first kind of variation is variation in the realized sequence of phones a word consists of, and the second kind is variation in the acoustic realization of sounds, the so called allophonic variation. Up till now, we have only studied

the first kind of pronunciation variation, because we expect the allophonic variation to be implicitly modelled in the HMMs.

In the next section, some of the difficulties caused by pronunciation variation are discussed for both the training and the recognition procedure. First, we explain how recognition works and how modelling pronunciation variation may improve it. Next, we explain the training procedure and how pronunciation variation can be modelled during training. In the following section, the method for modelling pronunciation variation is explained in detail. Subsequently, the results obtained are analysed in detail. In the last section, we discuss why the improvements obtained so far with this method are small, and how we intend to adapt the method to obtain considerably higher recognition performance.

How does a continuous speech recognizer work?

Recognition

The CSR component of the spoken dialogue system converts incoming speech signals into corresponding sequences of words. In the online SDS, the CSR passes a number of sequences of words on to the NLP, in the form of a wordgraph. From this wordgraph, it is possible to compute that sequence of words which best fits to the incoming speech signal, the so called Best Sentence (BS).

Nowadays, almost all CSRs are probabilistic machines. This means that the CSR calculates the probability that the incoming speech signal is the result of the production of a specific sequence of words. This probability is calculated for a lot of potential sequences of words. The sequence with the highest probability is “recognized”. Before a CSR can be used, it has to be trained. During training the recognizer “learns” the probability of observing a specific speech signal when a certain sequence of words is uttered.

The CSR can only recognize words which are in the lexicon. The words are listed in the lexicon in two forms: an orthographic form and a transcription in basic sound units. These basic sounds are all Dutch phonemes and a few non-speech sounds. In this article, the basic recognition units are called “phones”. During recognition, the phone transcriptions for the words are looked up in the lexicon and the words are replaced by the corresponding

sequences of phones. For each phone there is a corresponding phone model, the so called hidden Markov model (HMM). The statistics of the corresponding phone are stored in this model.

During the recognition phase, the CSR attempts to recognize an unknown sequence of words. If all possible sequences of words had to be generated, the number of hypotheses would be vast. Fortunately, from the start all hypotheses are scored according to their probability. The probability of a word is calculated on the basis of the HMMs which correspond to the phones a word consists of. The majority of the hypotheses appear to be much less probable than the best hypotheses so that they can be removed from the list of possible solutions without consequences. For each possible sequence of words which remains, the optimal alignment and the corresponding probability are calculated. Finally, the sequence of words with the highest probability is recognized.

For each word in the lexicon there is only one phone transcription: this is the canonical phone transcription which represents the most probable pronunciation of the word, based on introspective linguistic knowledge. Using a lexicon with only one canonical phone transcription leads to the problem that a word which is pronounced differently from the pronunciation in the lexicon may be incorrectly recognized. An example is the pronunciation of *Delft* (Dutch city).

Suppose the canonical pronunciation in the lexicon is: /dɛlft/

Suppose the realized pronunciation is: /dɛləf/

In the realized pronunciation /ə/ is inserted and /t/ is deleted with respect to the canonical pronunciation in the lexicon. In this example, the calculated probability of the realized pronunciation of the word is lower than it would have been if the spoken phone sequence had been exactly equal to the phone transcription in the lexicon. The probabilities for /l/ and /f/ are lower, because /ə/ is inserted. For the phone /t/ the probability is also lower, because /t/ is not realized at all. In this situation, it is possible that an incorrect word has a higher probability than the uttered word, and consequently the incorrect word is recognized.

A possible solution for this problem is to allow for multiple pronunciations in the lexicon instead of a single pronunciation. During recognition, these added pronunciation variants function as additional hypotheses. It is then to be expected that the actual realized pronunciation will deviate less from the most probable variant than when a canonical lexicon is used.

Training

Before we are able to do any kind of recognition, the phone models need to be trained. To train the phone models, it is necessary to have a large amount of recorded speech material (corpus) with corresponding transcriptions.

The training procedure consists of the following steps:

- The phone transcriptions of the words are looked up in the lexicon. Each utterance is replaced by its phone transcription; in this way, the phone transcription for the whole utterance is obtained.
- The Viterbi algorithm is used to find the optimal alignment between the speech signal and the phone transcription. In fact, this alignment is a segmentation because the boundaries are determined for each phone unit in the transcription. For each alignment, the Viterbi algorithm calculates a probability. This probability can be interpreted as the chance that the phone transcription and the speech signal belong together. The optimal alignment is the alignment with the highest probability.
- After segmentation, all parts of the speech material which correspond to the same phone are statistically processed. This results in a stochastic model (HMM-model) for each basic recognition unit (phone). To obtain reliable estimates of the model's parameters, it is necessary to use a large number of realizations of each phone to train the models.

All steps are repeated a number of times in an iterative process. There is mathematical proof that the average probability of the transcribed words is improved each iteration.

In addition to the phone models, a language model is also trained. A language model predicts the probability of occurrence of a word (unigram), or of a sequence of words (bigram, trigram etc.). These language models play an important role in the recognition task. However, in this research, the language models remain unchanged, so we will pay no further attention to this topic here.

If a canonical lexicon is used during training, a similar difficulty arises as in the recognition procedure. The pronunciation of a word can differ from the pronunciation in the lexicon represented by the canonical phone transcription. If the phone models are trained on the basis of this wrong phone sequence, parts of the speech signal train the wrong model, and consequently the phone models become contaminated. If we look at the same

example of the Dutch city *Delft*, we see that in the spoken word /ə/ is inserted and /t/ is deleted with respect to the canonical pronunciation. So, in this case, the models for /l/ and /f/ are contaminated, because parts of the speech signal where /ə/ is spoken train the models for /l/ and /f/. The model for /t/ is also contaminated, because parts of the speech signal train the model, while /t/ is not realized at all.

A possible solution to this problem can be to annotate manually what has been said in the speech material and to train new, less contaminated phone models on the basis of this more accurately annotated training corpus. The main disadvantage of manually annotating is that it is time-consuming and therefore costly. For this reason, we propose a method in which the CSR is used to annotate the speech material automatically. In order to do so, a lexicon is needed with multiple pronunciation variants for each word. The new phone models are expected to be less contaminated than the original ones. Therefore, they allow for less pronunciation variation than the original phone models. In order to optimally use the new phone models during recognition, it is necessary to use a lexicon in which the pronunciation variation is explicitly modelled, i.e. a lexicon with multiple pronunciations for a word.

Method and Material

Method

The starting-point of the current research was a CSR in which a lexicon was used with only one, canonical pronunciation for each word. In order to model pronunciation both in training and in recognition, it is necessary to generate a lexicon with multiple pronunciations for each word. The method we used resembles those used previously with success by Cohen (1989) and Lamel & Adda (1996). In this approach, phonological rules are used to generate pronunciation variants automatically, i.e. to expand the lexicon. The expanded lexicon can then be used during training, recognition, or both. During recognition, the old recognition lexicon is simply replaced by the new one in order to make it possible to recognize pronunciation variants. During training, the pronunciation variants are used to obtain new phone models as follows:

1. Start off with a single pronunciation lexicon, containing canonical

pronunciation forms. Use the original corpus and this single pronunciation lexicon to calculate the first version of the phone models.

2. Choose a set of phonological rules.
3. Generate a multiple pronunciation lexicon by expanding the single pronunciation lexicon with pronunciation variants obtained with the set of phonological rules.
4. Do a forced recognition in order to determine which variants have been realized in the corpus. During this recognition, the CSR is forced to choose the pronunciation variant which is the best description of the speech signal. It is only possible to choose between different phone transcriptions of the same word, but not between different words. By substituting the initial transcriptions with those selected during forced recognition, new phone transcriptions for the training corpus are obtained automatically.
5. The new transcriptions are used to calculate updated phone models.

Steps 4 and 5 can be repeated a number of times. We expect the updated phone models to be less contaminated than the original ones. When these updated phone models are used during forced recognition the correct variant will be chosen more often than when the original phone models are used. Therefore, each iteration the newly updated phone models will be less contaminated. Steps 2 through 5 can be repeated for different sets of rules.

Our ultimate goal is to find the rules that are optimal in the sense that application of these rules gives the greatest increase in recognition performance. The goal of the current research was to test whether the method proposed above is suitable for our purposes. In order to do so, the method was first tested by using four phonological rules which were applied word internally, as will be explained below. Next, cross-word variation was modelled for a set of frequently occurring word sequences.

Phonological rules

In order to select the initial set of phonological rules, a number of criteria were followed. Variation occurs both within words and across words. Given the use of a lexicon in our CSR, it was obvious to begin with word internal variation. Therefore, the first criterion was to choose rules of word phonology. Second, we decided to start with rules concerning those phenomena that are known to be most detrimental to automatic speech recognition. Of the three possible recognition errors, i.e. insertions, deletions, and substitutions, we expect the first two to have the greatest consequences

for speech recognition, because they affect the number of segments present in different realizations of the same word. Therefore, starting with rules concerning insertions and deletions was the second criterion we adopted. The third criterion was to choose rules that are frequently applied. Frequently applied can be interpreted in two ways. First, a rule can be frequent because it is frequently applied whenever the context for its application is met. Second, the context in which a rule is applicable can be very frequent, even though the rule is not applied in most of the cases. Obviously, it is this latter case of “frequent occurrence” which is most interesting for automatic speech recognition, since it is difficult to predict in this case which variant should be selected as the canonical form, while in the former case the most frequent form would probably suffice as sole transcription. The fourth criterion (related to the previous one) was that the rules should regard phones that are relatively frequent in Dutch, since rules that concern frequent phones will influence the recognizer's performance to a larger extent. Finally, we decided to start with rules that have been extensively described in the literature, so as to avoid possible effects of overgeneration and undergeneration due to incorrect specification of the rules.

On the basis of the above-mentioned criteria, the following four rules were selected. The description of the four rules is after Booij (1995):

1. /ə/-deletion:

When a Dutch word has two consecutive syllables headed by /ə/, the first /ə/ may be deleted, provided that the resulting onset consonant cluster is an obstruent + liquid cluster. Example:

latere /latərə/ → /latrə/ ‘later’

2. /t/-deletion:

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

Example:

snelstmogelijk /snɛlstmoxələk/ → /snɛlsmoxələk/ ‘fastest’

If the preceding consonant is a sonorant, /t/-deletion is possible, but then the following consonant must be an obstruent. If the obstruent following the

sonorant + /t/ cluster is a /k/, deletion does not apply. If a /t/ is preceded by a sonorant, and also followed by a sonorant, deletion is impossible.

Example:

's avonds	/savɔnts/ → /savɔns/	'in the evening'

Booij does not mention that in some Dutch dialects /t/-deletion also occur in word-final position. We decided to apply the /t/-deletion rule in word-final position following an obstruent (unless the obstruent is an /s/).

Example:

Utrecht	/ytrɛxt/ → /ytrɛx/	'Dutch city'

3. /n/-deletion:

In standard Dutch, syllable-final /n/ can be dropped after /ə/, except in the indefinite article *een* /ən/ 'a'. For many speakers, in particular in the western part of the Netherlands, the deletion of /n/ is obligatory. An /n/ is deleted if it is the final /n/ of a syllable after /ə/ and if that syllable is not a verbal stem. Example:

reizen	/rɛizən/ → /rɛizə/	'to travel'

4. /ə/-epenthesis:

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

Delft	/dɛlft/ → /dɛləft/	'Dutch city'

Material

The speech material was collected with an online version of the SDS connected to an ISDN line. The training and test material consisted of 25,104 utterances (83,890 words) and 6,276 utterances (21,108 words), respectively.

The most important characteristics of the CSR are the following. The input signals consist of 8 kHz, 8-bit A-law coded samples. Feature extraction is done every 10 ms for frames with a width of 16 ms. The first step in feature analysis is an FFT analysis to calculate the spectrum. Next, the energy is calculated in 14 mel-scaled filter bands between 350 and 3400 Hz. The final processing stage is the application of a discrete cosine transformation on the log filterbank coefficients. Besides these 14 cepstral coefficients, the 14 delta coefficients are also used. This makes a total of 28 feature coefficients. The CSR uses acoustic models (HMMs), language models (unigram and bigram), and a lexicon. The HMMs consist of three segments of two identical states, of which one state per segment can be skipped.

The canonical training lexicon contains 1,412 entries, which are all the words in the training corpus, while the recognition lexicon contains 1,154 entries. There were no out-of-vocabulary (OOV) words in the test corpus, which is a slightly artificial condition. The reason for this is that we wanted to measure the effect of modelling pronunciation variation and to avoid a situation in which a lot of errors would be caused by OOV words.

The four phonological rules selected for investigation affect 38% of the words in each lexicon. In a number of cases, more than one rule could be applied to one word. On average, 1.3 variants were generated for each word.

Results

Forced recognition

As forced recognition is an essential part of our method, a small-scale experiment was conducted to check whether the forced recognition procedure worked correctly. Listeners were asked to perform the same task as the forced recognition, which was to assess which pronunciation variant had been spoken. Their results were then compared to the results of the forced recognition. From this experiment we could conclude that the correct variant was chosen in 90% of the 711 cases.

Within word pronunciation variation

In the first experiments, the set of four phonological rules were applied *word internally* to all words in the lexicon. The effects of adding pronunciation

variants were measured in Sentence Error Rates (SER = percentage of incorrectly recognized sentences). As a baseline system, we used a CSR with a canonical lexicon which contains one variant for each word. This "most probable pronunciation" can be a variant of which one of the four phonological rules has already been applied, e.g. the canonical form for *reizen* 'to travel' is /rɛizə/ (n-deletion). The SER for the baseline system is 21.48%.

The multiple lexicon was used during training and/or recognition, which results a combination of four testing conditions. For training, stages 4 and 5 of our method were repeated in iteration. A gradual improvement in recognition performance was observed during the first 3 iterations. The results in SER, for all four CSRs after 3 iterations, are given in Table 1. In this table "original" means that the original corpus was used to train the phone models, "updated 3" that the updated corpus obtained after 3 iterations was used to train the new phone models, "single" means that a canonical pronunciation lexicon was used during recognition, and "multiple" means that the multiple pronunciation lexicon was used during recognition.

Table 1. SER for four different test conditions

training	original	original	updated 3	updated 3
recognition	single	multiple	single	multiple
SER (%)	21.48	21.06	22.05	20.81

Using the multiple lexicon during training, but not during recognition, causes a deterioration in SER of 0.57% (compare column 4 to 2). This result is in line with our expectations. The updated phone models allow for less pronunciation variation than the original phone models, so in order to benefit from the less contaminated phone models the pronunciation variation has to be modelled in the lexicon. Using the multiple pronunciation lexicon during recognition alone led to an improvement in SER of 0.42% (compare column 3 to 2). Performance improved by another 0.25% (0.67% in total) when the multiple pronunciation lexicon was used during both training and recognition (compare column 5 to 3). It thus appears that the multiple pronunciation lexicon has more effect when used during recognition than during training. However, combining the two produces the best results. Although the

improvements in SER are not significant, the trends are in line with Lamel & Adda (1996).

Detailed analysis of changes in SER

The largest improvements in performance result from the use of a multiple pronunciation lexicon. In order to get more insight into the effects of this method, the results obtained with the two lexicons (single and multiple) were analysed in further detail. For instance, we counted the number of incorrect sentences obtained with the single and the multiple pronunciation lexicon. These results are shown in Table 2. In column 2 (SER unchanged) the errors which do not affect the SER are given, while in column 4 (SER changed), the changes in SER are shown which decrease (improvements) or increase (deteriorations) SER.

Table 2. Details of changes in SER due to recognition with a multiple pronunciation lexicon compared to recognition with a single lexicon

	SER unchanged		SER changed
same sentence, same error	1129	improvements	+56
same sentence, different error	163	deteriorations	-30
NETT RESULT	1292	NETT RESULT	+26

Table 2 shows that a considerable number of incorrectly recognized sentences remain incorrect (1292) when using a multiple pronunciation lexicon for recognition. There are cases in which a better solution is chosen (56), but, since in a number of cases a worse solution is chosen (30), the two effects cancel each other out, and the nett result (26) is small. This neutralization effect explains why no significant improvements in the SERs were observed in Table 1.

Improvements and deteriorations for each rule

Next, we studied the improvements and deteriorations in SER distributed over the different phonological rules. In one sentence, more than one phonological rule can cause (deterioration) or solve (improvement) errors. However, a sentence is either correct or incorrect. If an error occurs due to

two phonological rules, both rules get a count of 0.5 deterioration. Table 3 shows the improvements and deteriorations caused by each rule. In this Table, “n-del.” means n-deletion rule, “ə-epe.” means ə-epenthesis rule, “ə-del.” means ə-deletion, “t-del.” means t-deletion rule, and “unknown” means that it is unclear which rule caused or solved an error. From this Table it can be concluded that pronunciation modelling due to the n-deletion rule has the largest effect on recognition performance.

About 96% of the deteriorations can be explained by confusability. This means that a word which was correctly recognized before is now incorrectly recognized, because it is confused with a pronunciation variant which has been added to the multiple pronunciation lexicon.

Table 3. Improvements and deteriorations for each phonological rule

	n-del.	ə-del.	ə-epe.	t-del.	unknown	TOTAL
improvements	+35	+1	+10	+4	+6	+56
deteriorations	-15	0	-6.5	-3.5	-5	-30
NETT RESULT	+20	+1	+3.5	+0.5	+1	+26

How different selection criteria for variants in lexicon affect SER
The presence of multiple pronunciations in the training corpus makes it possible to study how the frequency of the variants included in the recognition lexicon affects SER. We performed a test in which we used a single-variant lexicon containing the least frequent pronunciation of each word, and a second test in which a single-variant lexicon containing the most frequent variant in the training corpus was used. The results are shown in Table 4.

Table 4. SER for different variants in the recognition lexicon

variant	least frequent	canonical	all	most frequent
SER (%)	22.94	21.48	20.81	20.70

When the single-variant lexicon containing the least frequent pronunciations is used, SER is 22.94%. However, when the variants are replaced by the most probable variants, SER drops to 20.70%. For the canonical lexicon SER is 21.48%, and for the multiple lexicon (all) SER is 20.81%. In other words, using a lexicon with the most probable variants led to a significantly better performance in SER, than using a lexicon containing the least frequent variants. Using a canonical lexicon and a multiple pronunciation lexicon leads to a recognition performance somewhere in between.

These results show that selecting the right variants is crucial and that it is difficult to determine whether the method under study improves recognition performance to a sufficient extent. Since the decision is usually made by comparing performance before and after applying the method, it follows that the better the pretest performance, the smaller the improvement will be. If we had started with a lexicon containing the least probable variants, we would have concluded that modelling pronunciation variation leads to a considerable improvement. On the other hand, if we had started with a lexicon containing the most probable variants, we would have found no improvement at all. Clearly, our current results are somewhere in between.

Application of the four phonological rules in spontaneous speech

Using the forced recognition procedure it is possible to investigate which pronunciation variant is spoken in the speech database we used in our experiments. The results give some insight into how the four phonological rules we used are applied in spontaneous speech. We used the training corpus, consisting of 83,890 words, for the forced recognition. Of these words 14,950 words have one or more pronunciation variants. After forced recognition it is possible to count in how many cases a specific rule has been applied. In this case, application of the rule is independent of the canonical form, thus if the canonical form for the Dutch word *reizen* 'to travel' is realized as /rɛizə/, then the count for the n-deletion rule is raised by one. Because the forced recognition works correctly in 90% of the cases, the estimated error is 10%.

Table 5. Application of each rule in the training corpus

	n-del.	ə-del.	ə-epe.	t-del.
# times rule applied	7256	161	792	896
% rule applied	40%	59%	42%	25%

These results are not independent of the domain of the SDS. For instance, this material includes quite a lot of station names ending in “en”. In only 34% of the 2909 cases of the station names, the n-deletion rule is applied. Stations are important information for this SDS application. Therefore, people may tend to articulate the station names more precisely. Excluding the station names from the speech material would have led to a relatively higher occurrence of the n-deletion rule (45%).

Cross-word pronunciation variation

In order to test *cross-word* pronunciation variation, we defined a set of multi-words. A multi-word is a sequence of words joined together to create a new entry in the lexicon. In our experiments, we selected nine frequent multi-words, on the basis of the following criteria: they had to occur frequently in the corpus, they had to form a logical unit, and only sequences of words for which we expected cross-word pronunciation variation to occur were selected. Thus, a training corpus was created in which 8% of the words are multi-words. We applied the four phonological rules to all words in the lexicon, including the multi-words. Furthermore, we used reduction rules to obtain pronunciation variants for the nine multi-words. The phone models which gave best performance in the last experiments were used for recognition (updated 3). The recognition results are given in Table 4 (column 2 and 3). The results for recognition with phone models obtained after a new iteration are also given in Table 4 (updated 4, column 4 and 5).

Table 6. SER for four different test conditions, using nine multi-words

training	updated 3	updated 3	updated 4	updated 4
recognition	single	multiple	single	multiple
SER (%)	21.54	20.76	21.48	20.65

Comparing column 2 of Table 6 with column 4 of Table 1 shows that the use of multi-words alone already improves SER by 0.51%. When multiple pronunciations for all nine multi-words are added to the recognition lexicon, SER remains almost unchanged (compare column 3 Table 6 with column 5 Table 1). Table 6 also shows that an extra iteration of the phone models further improves SER (column 4 and 5).

Discussion and conclusions

In the training corpus, in 40-50% of the cases in which an alternative pronunciation could be chosen, the forced recognition did choose an alternative variant. For the test material, it can be seen that also in 40-50% of the cases an alternative pronunciation variant was chosen, as appears from the output of recognition with a multiple pronunciation lexicon. Considering the fact that during forced recognition for 90% of the cases the correct variant is chosen, it can be concluded that pronunciation variation is indeed being modelled.

However, the improvements in recognition performance are not significant. As mentioned before, to judge the results it is important to take into account the performance of the CSR we started with. This CSR performs quite well (see Table 5), so it is to be expected that no big improvements will be found. Nevertheless we think there is room for improvement.

First of all, the improvements due to the use of updated phone models are small. This is possibly due to the fact that the number of phones which change after updating the training corpus is very small. About 8.4% of the words in the training corpus are replaced by an alternative variant after forced recognition. Most of the words differ in only one phone from the canonical form, so effectively 2.2% of the phones are changed after the first

iteration. Each iteration, the number of phones which are affected increases, but the increase is smaller in every iteration.

Second, when analysing in detail the changes in SER, it can be seen that the use of the multiple pronunciation lexicon has a larger effect on recognition performance than the use of updated phone models. However, these improvements are also small.

First of all, this is because the improvements are partly balanced off by the deteriorations which arise when using a multiple pronunciation lexicon. These deteriorations can be explained by confusability between words which were correctly recognized before and pronunciation variants that are added to the lexicon. The confusability might be reduced by taking into account the frequency of occurrence of a pronunciation variant. This can be done by lowering the probability of less frequent variants. In order to make a reliable estimate of the frequency of occurrence of a pronunciation variant, it is necessary to have enough realizations of each pronunciation variant. We think that the corpus we used to train our models is not large enough to make reliable estimates. Therefore, these probabilities are now being calculated on the basis of a substantially enlarged training data base.

Another reason for the small improvements we found is that we only used four phonological rules. From the experiment in which the nine most frequent multi-words were used it can be concluded that modelling between-word pronunciation variation might improve recognition performance. Cremelie & Martens (1995) reported that modelling pronunciation variation across word boundaries is even more important than modelling word internal variation. Therefore, we expect that expanding the rule set and also applying the rules across word boundaries will give larger improvements in recognition performance.

Acknowledgments

This work was funded by the Netherlands Organization for Scientific Research (NWO) as part of the NWO Priority Programme Language and Speech Technology.

References

G. Booij (1995) *The Phonology of Dutch* Oxford: Clarendon press.
M.H. Cohen (1989). ‘Phonological structures for speech recognition’ *Ph.D. dissertation,*

Univ. of California, Berkeley.

N. Cremelie & J.P. Martens (1995) 'On the use of pronunciation rules for improved word recognition'. *Proceedings EUROSPEECH'95, Madrid*: 1747-1750.

L.F. Lamel & G. Adda (1996) 'On designing pronunciation lexicons for large vocabulary, continuous speech recognition'. *Proceedings ICSLP'96, Philadelphia*: 6-9.

H. Strik, A. Russel, H. van den Heuvel, C. Cucchiarini & L. Boves (1996) 'Localizing an automatic inquiry system for public transport information'. *Proc. ICSLP'96, Philadelphia*: 853-856.

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

V. Steinbiss, H. Ney, X. Aubert, S. Besling, C. Dugast, U. Essen, D. Geller, R. Haeb-Umbach, R. Knesser, H.-G. Meier, M. Oerder & B.-H. Tran (1995) 'The Philips Research System for continuous-speech recognition' *Philips Journal of Research*, Vol. 49, No. 4, pp. 317-352.