Making a difference

On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition

Making a difference
On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition

Judith Maria Kessens
Ph. D. Thesis, Nijmegen, 2002.
Keywords: Speech recognition, pronunciation variation, automatic phonetic transcription

Photography: Judith M. Kessens
Design: Maaike Maréchal
Printed and bound by Ponsen & Looijen B.V., Wageningen

ISBN: 90-9015829-4

Making a difference

On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition

Een wetenschappelijke proeve op het gebied van de Letteren

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen in het
openbaar te verdedigen op maandag 10 juni 2002
des namiddags om 1:30 uur precies

door

Judith Maria Kessens

geboren op 19 februari 1972
te Muiden

Promotor:	Prof. dr. L. W. J. Boves
Co-promotor:	Dr. W. A. J. Strik

Manuscriptcommissie:	Prof. dr. R. van Hout, Prof. dr. ir. J.-P. Martens (ELIS, University of Gent) Prof. dr. T. Svendsen (NTNU, Norwegian University of Science and Technology)

The research in this thesis was carried out within the framework of the Priority Programme Language and Speech Technology, supported by NWO (Dutch Organization for Scientific Research).

Voor mijn vader

Contents

Contents **vii**

Een woord van dank **ix**

List of publications **xi**

Making a difference **1**

1 Speech recognizer and speech material **3**

1.1 The OVIS spoken dialogue system 3
1.2 The speech recognition component 5
1.2.1 Acoustic pre-processing 5
1.2.2 Training 7
1.2.3 Recognition 9
1.3 The VIOS speech material 10

2 Pronunciation variation **13**

2.1 Sources of pronunciation variation 13
2.2 Why is pronunciation variation problematic for ASR? 14
2.3 Overview of methods to model pronunciation variation in ASR 16

3 Goals and methodology **18**

3.1 Goals 18
3.2 Methodology 18
3.2.1 Automatic phonetic transcription 18
3.2.2 Modeling pronunciation variation 19
3.2.3 Evaluation 20

4 Summaries of the articles **22**

4.1 Summary 1 22
4.2 Summary 2 24
4.3 Summary 3 26
4.4 Summary 4 28

5 Discussion **31**

5.1 Automatic phonetic transcription 31
5.1.1 Automatic versus manual phonetic transcription 31
5.1.2 Automatic transcription quality versus recognition performance 32
5.1.3 Application areas of automatic phonetic transcription 32

5.2 Modeling pronunciation variation 33
5.2.1 General method of modeling pronunciation variation 33
5.2.2 Why are the improvements so small? 35
5.2.3 Alternatives to phone level modeling of pronunciation variation 37

6 Conclusions and future work 39

6.1 Conclusions 39
6.2 Future work 39
6.2.1 Automatic phonetic transcription 39
6.2.2 Improving pronunciation variation modeling 40
6.2.3 Comparison of methods 41

References 43

Appendix A 48

The articles 49

1 Obtaining phonetic transcriptions: A comparison between expert listeners and a continuous speech recognizer, Language & Speech 44 (3), 377-403. 51

2 On automatic phonetic transcription quality: Lower WERs do not guarantee better transcriptions, submitted to Computer, Speech and Language. 81

3 Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation, Speech Communication 29, 193-207. 115

4 A data-driven method for modeling pronunciation variation, submitted to Speech Communication. 133

Samenvatting (summary in Dutch) 163

Curriculum Vitae 171

Een woord van dank

Nu dit boek voltooid is, kan ik dan eindelijk iedereen bedanken die een positieve bijdrage heeft geleverd aan de totstandkoming ervan. De eerste persoon die ik wil bedanken is mijn begeleider en copromotor Helmer Strik. Helmer, de manier waarop je me begeleid hebt in het onderzoek heb ik als zeer prettig ervaren. Je gaf me voldoende sturing, maar ook behoorlijk veel vrijheid, zodat ik zelf de richting van mijn onderzoek kon bepalen. Voor mij is weer eens bewezen dat promotie-onderzoek staat (of valt) bij een goede begeleiding. Wat ik het meest aan jou gewaardeerd heb, is dat je niet alleen geïnteresseerd was in mijn werk, maar ook in mij als persoon; jouw steun en begrip tijdens moeilijke momenten in mijn leven waren voor mij erg belangrijk. Verder ben ik mijn promotor Loe Boves veel dank verschuldigd. Loe, ik heb het enorm gewaardeerd dat je veel van je (kostbare) tijd hebt vrijgemaakt om mijn werk te kunnen voorzien van nuttig commentaar. Iedere keer bleek weer dat jij de vinger precies op de zere plek wist te leggen (ook al deed dat af en toe best een beetje pijn). Een derde persoon die ik wil bedanken, is Mirjam Wester. Mirjam, onderzoek doen met jou was absoluut niet saai! Op onze kamer was het nooit stil; ik denk terug aan vele gesprekken die vaak over het werk gingen, maar ook over dingen die niets met het werk te maken hadden. Bedankt voor de goede discussies en alle lol! Tenslotte heb ik veel samengewerkt met Catia Cucchiarini. Catia, de samenwerking met jou was altijd zeer prettig. De vele discussies over het onderzoek waren altijd inspirerend en leerzaam. Vooral op momenten dat ik minder blij was met mijn onderzoeksresultaten lukte het jou altijd weer om mij te inspireren door jouw enthousiasme.

Ik kijk terug op een leerzame en boeiende periode van onderzoek, aan de afdeling Taal & Spraak van de Katholieke Universiteit Nijmegen. De goede sfeer op de afdeling heeft zeker een positieve bijdrage geleverd aan mijn proefschrift! Daarom wil ik iedereen van de afdeling Taal & Spraak bedanken voor de gezellige koffiepauzes en de gezamenlijke lunches in “de Refter”. Toen ik voor het eerst in Nijmegen kwam, had ik nog nooit van een “hidden Markov model” gehoord. Nu ben ik voorzien van een grote bagage kennis over automatische spraakherkenning. Dit is zeker te danken aan de inspirerende werkomgeving en de mensen die er werken (en gewerkt hebben). De interdisciplinaire werkomgeving bij Taal & Spraak heeft ertoe geleid dat ik me breder heb kunnen ontwikkelen dan alleen op het gebied van de spraaktechnologie. Veel van mijn spraaktechnologische kennis heb ik te danken aan A^2RT. Daarom wil ik alle (ex) A^2RT-ers speciaal bedanken voor het lezen van de vele papers en artikelen, de discussies over mijn onderzoek en de nuttige feedback op proefpresentaties.

Verder wil ik NWO, Shell Nederland B.V. en “the International Speech Communication Association” (ISCA) bedanken voor de financiële bijdragen aan verschillende congresbezoeken.

Dan zijn er nog mensen die ik extra dank verschuldigd ben. Roel, voor alle geduld en steun die je mij hebt gegeven tijdens het merendeel van mijn promotie-onderzoek. David en Maike voor het bewijs dat liefde geen grenzen kent. Bernard, omdat je mijn soul-brother en ceremoniemeester bent, en Ruth voor de dubbele gezinsuitbreiding. Yke, omdat je er altijd voor me bent. Maaike voor je vriendschap en voor het ontwerp van de kaft van dit proefschrift. Al mijn vrienden, voor de nodige gezelligheid en ontspanning. Diana, Dorota, Janienke, Simo en Mieke, voor alle bijzondere trein-, fiets-, Bruna- en Etos momenten. Mijn kamergenoten Mirjam en Mieke, voor degezelligheid. “Janimfke” en “Di Para” omdat jullie mijn paranimfen willen zijn. Mirjam, Catia, Febe, Janienke, Johan en Henk, voor het lezen van het eerste deel van dit proefschrift.

Tenslotte wil ik mijn ouders bedanken, zonder hen was ik nooit zover gekomen als nu. Zij hebben me altijd in woord en daad gesteund en mij gemotiveerd en gestimuleerd. Lieve Josephine, bedankt voor je onvoorwaardelijke steun en liefde. Zelfs op momenten die voor jou heel moeilijk waren, speelde je het voor elkaar om er voor mij te zijn. Lieve Joop, ik ben er dankbaar voor dat je het eerste jaar van mijn promotie nog hebt mee kunnen maken. Ik weet zeker dat je absoluut trots op me zou zijn geweest; minstens zo trots als jóuw vader, omdat die tweede dr. Kessens er dan toch eindelijk is!

Nijmegen, april 2002,

Judith Kessens.

List of publications

This thesis consists of the following publications:

- Wester, M., Kessens, J.M., Cucchiarini, C. and Strik, H. (2001). Obtaining phonetic transcriptions: A comparison between expert listeners and a continuous speech recognizer, *Language & Speech* **44 (3)**, 377-403.
- Kessens, J.M. and Strik, H. On automatic phonetic transcription quality: Lower WERs do not guarantee better transcriptions, *submitted to Computer Speech and Language*.
- Kessens, J.M., Wester, M. and Strik, H. (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation, *Speech Communication* **29**, 193-207.
- Kessens, J.M., Cucchiarini, C. and Strik, H. A data-driven method for modeling pronunciation variation. *submitted to Speech Communication.*

Other publications not included in this thesis:

- Kessens, J. M. and Strik, H. (2001). Lower WERs do not guarantee better transcriptions. In: *Proceedings of Eurospeech*, Aalborg, Denmark, 1721-1724.
- Strik, H., Cucchiarini, C. and Kessens, J. M. (2001). Comparing the performance of two CSRs: How to determine the significance level of the differences. In: *Proceedings of Eurospeech*, Aalborg, Denmark, 2091-2094.
- Kessens, J. M., Wester, M. and Strik, H. (2000). Automatic detection and verification of Dutch phonological rules. In: *PHONUS 5: The workshop on "Phonetics and Phonology in ASR"*, Saarbrücken, Germany, 117-128.
- Kessens, J. M., Strik, H. and Cucchiarini, C. (2000). A bottom-up method for obtaining information about pronunciation variation. In: *Proceedings of ICSLP*, Beijing, China, 274-277.
- Strik, H., Cucchiarini, C. and Kessens, J. M. (2000). Comparing the recognition performance of CSRs: In search of an adequate metric and statistical significance test. In: *Proceedings of ICSLP*, Beijing, China, 740-743.
- Wester, M., Kessens, J. M. and Strik, H. (2000). Pronunciation variation in ASR: Which variation to model? In: *Proceedings of ICSLP*, Beijing, China, 488-491.
- Wester, M., Kessens, J. M. and Strik, H. (2000). Using Dutch phonological rules to model pronunciation variation in ASR. In *PHONUS 5*: *The workshop on "Phonetics and Phonology in ASR"*, Saarbrücken, Germany, 105-116.
- Kessens, J. M., Wester, M. and Strik, H. (1999). Modeling within-word and cross-word pronunciation variation to improve the performance of a Dutch CSR. In: *Proceedings of ICPhS*, San Francisco, USA, 1665-1668.

- Wester, M. and Kessens, J. M. (1999). Comparison between expert listeners and continuous speech recognizers in selecting pronunciation variants. In: *Proceedings of ICPhS*, San Francisco, USA, 723 - 726.
- Kessens, J. M., Wester, M. and Strik, H. (1998). The selection of pronunciation variants: comparing the performance of man and machine. In: *Proceedings of ICSLP*, Sydney, Australia, 2715-2718.
- Wester, M., Kessens, J. M. and Strik, H. (1998). Improving the Performance of a Dutch CSR by Modeling Pronunciation Variation. In: *Proceedings of the ESCA workshop "Modeling Pronunciation Variation for Automatic Speech Recognition"*, Kerkrade, The Netherlands, 145-150.
- Wester, M., Kessens, J. M. and Strik, H. (1998). Modeling pronunciation variation for a Dutch CSR: testing three methods. In: *Proceedings of ICSLP*, Sydney, Australia, 2535-2538.
- Wester, M., Kessens, J. M. and Strik, H. (1998). Selection of pronunciation variants in spontaneous speech: Comparing the performance of man and machine. In: *Proceedings of the ESCA workshop "on the Sound Patterns of Spontaneous Speech: Production and Perception"*, Aix-en-Provence, France, 157-160.
- Wester, M., Kessens, J. M. and Strik, H. (1998). Two automatic approaches for analyzing the frequency of connected speech processes in Dutch. In: *Proceedings of ICSLP*, Sydney, Australia, 3351-3356.
- Kessens, J. M., Wester, M., Cucchiarini, C. and Strik, H. (1997). Testing a method for modelling pronunciation variation. In: *Proceedings of the COST workshop "Speech Technology in the Telephone Network: Where are we today?"* Rhodes, Greece, 37-40.
- Kessens, J. M. and Wester, M. (1997). Improving recognition performance by modelling pronunciation variation. In: *Proceedings of the CLS opening Academic Year '97/'98*, Nijmegen, The Netherlands, 1-20.
- Wester, M., Kessens, J. M., Cucchiarini, C. and Strik, H. (1997). Modelling pronunciation variation: some preliminary results. In: *Proceedings of the Department of Language & Speech*, Nijmegen, The Netherlands, 127-137.

Making a difference

For human beings, speech constitutes a very efficient means of communication. This has induced many people to think that speech might also be a very efficient means of communication between human beings and machines. For this reason, attempts have been made to use speech as input to computers. The term *Automatic Speech Recognition* (ASR) is used for the technology that is required to transform 'speech' into 'text'. Since the emergence of the first automatic speech recognizer in 1952 (Davis et al., 1952), substantial progress has been made in the field of ASR. What started as recognition of ten digits spoken in isolation by a single speaker has now evolved to speaker-independent, large-vocabulary recognition of fluent, extemporaneous speech. In spite of the progress that has been made, a gap still exists between the performance of human beings and machines on speech recognition. For instance, Lippmann (1997) showed that the performance of present-day speech recognizers is at best one order of magnitude worse than human speech recognition on similar tasks.

There are a number of differences in the way speech is decoded by human beings and by machines that could explain why ASR performance has not yet reached the same level of performance as human speech recognition. One of the main differences between human and machine speech recognition is that human beings use much more information for speech decoding than machines do. For instance, most human beings use two ears for hearing, whereas speech recognizers usually process a single stream of speech. Furthermore, a speech recognizer can only recognize the words that are contained in its vocabulary. Another difference is that human beings have certain expectations on the kind of speech that is likely to be produced. These expectations can be flexibly and quickly adjusted, depending on the speaker who is talking and the topic of the conversation. This kind of quick adaptation is hardly used in ASR systems. Other examples of information that machines can use only to a limited extent compared to human beings is information on intonation, stress, speaking rate, and *pronunciation variation*.

Pronunciation variation refers to the fact that words can be pronounced in many different ways. Differences exist in the way speech is pronounced by various speakers, but even if the same speaker utters a word more than once, it will never be pronounced in exactly the same way. Humans usually have no difficulties in processing different pronunciation variants of the same word, since they have knowledge of pronunciation variation. However, for speech recognizers, pronunciation variation forms a problem, because, in general, speech recognizers do not explicitly take into account the different ways in which words can be pronounced. In the beginning of ASR research, the amount of variation in pronunciation was limited by using only isolated words. Since then, the type of speech that can be processed has evolved from isolated words to spontaneous speech. Especially in spontaneous speech the amount of pronunciation variation is very large. Words are more connected to each other in spontaneous speech. As a consequence, the pronunciation of one word is influenced by that of adjacent words. Furthermore, words are usually articulated less carefully in spontaneous

speech. Modeling pronunciation variation is seen as a possible way of improving the performance of ASR systems that handle spontaneous speech. The research described in this thesis constitutes an attempt to find an adequate way of explicitly modeling Dutch pronunciation variation in order to improve the performance of ASR.

The body of this thesis consists of four articles describing research related to modeling pronunciation variation. The articles are preceded by six chapters that provide the context for the research reported on in this thesis. The organization of these chapters is as follows. Chapter 1 explains the operation and architecture of the speech recognizer that is used in this research. Chapter 2 deals with the various sources of pronunciation variation, and explains why pronunciation variation is problematic for ASR. Chapter 3 describes the goals and the general research methodology. Subsequently, the articles are summarized in Chapter 4, and the results are discussed in Chapter 5. Finally, the major conclusions are given in Chapter 6, together with some recommendations for future research.

The main part of this thesis consists of four articles that are published in or submitted to scientific journals. Our general method of modeling pronunciation variation requires information on the occurrence of the pronunciation variation to be modeled. In order to obtain this information, we use our speech recognizer to make transcriptions of large amounts of speech material. This procedure is called *automatic transcription*. During automatic transcription, the CSR decides which of a number of possible variants best matches the actual pronunciation. The first two articles of this thesis are concerned with this kind of automatic transcription. The goal of the first article is to assess the quality of the automatic transcriptions made by the speech recognizer by comparing them with transcriptions made by expert linguists. In the second article, some of the properties of the speech recognizer that influence the quality of automatic transcriptions are investigated in order to obtain better quality automatic transcriptions. Both articles show that our method of obtaining automatic transcriptions can be used meaningfully in the research on modeling pronunciation variation. The automatic transcription procedure is used as part of a general method of modeling pronunciation variation that is employed in the last two articles. In the third article, pronunciation variation is modeled in a knowledge-based manner. To this end, we selected five frequently occurring phonological processes to be modeled in our speech recognizer. However, not all pronunciation variation that is present in our speech material is described in the literature. For this reason, in addition to our knowledge-based approach, we also adopted a data-driven approach to model pronunciation variation. In the data-driven approach, the speech recognizer is used in order to obtain transcriptions of the pronunciation variation that is present in our speech material. The work on data-driven modeling of pronunciation variation is described in the fourth article.

The research on modeling pronunciation variation showed that both our knowledge-based and our data-driven approaches for modeling pronunciation variation lead to improvements in recognition performance. In other words: *Making a difference* (differentiating) between various pronunciation variants does indeed *make a difference* in the performance of automatic speech recognition.

1 Speech recognizer and speech material

The research described in this thesis was carried out within the framework of the Priority Programme Language and Speech Technology (PP-TST[1]) of the Dutch Organization for Scientific Research (NWO[2]). The PP-TST started in 1995 and finished in 2000. The programme was carried out at the University of Nijmegen (KUN), the Center for Research on User-System Interaction (IPO), the University of Amsterdam (UvA), and the University of Groningen (RUG), in close collaboration with Philips Corporate Research and KPN Research. The goal of the PP-TST was to conduct fundamental and applied research in the context of a spoken dialogue system. The spoken dialogue system that was developed provides information on train timetables in the Netherlands over the telephone, and is called OVIS. OVIS is an acronym for 'Openbaar Vervoer Informatie Systeem' ('Public Transportation Information System'). The OVIS spoken dialogue system is briefly described in section 1.1. The research in this thesis is only concerned with the speech recognition component of OVIS. The architecture and operation of the speech recognition component are explained in more detail in section 1.2. Finally, section 1.3 describes the speech material that has been used for the experiments in this thesis.

1.1 The OVIS spoken dialogue system

The architecture of OVIS is shown in Figure 1. To illustrate how the system operates, I will use an example. A person (the user) calls OVIS to obtain a travel advice. First, the system has to detect that there is a telephone call coming in. This interaction between the telephone line and OVIS is handled by the *Telephone Interface*. When the call is established OVIS replies with a welcome message, and asks the following question:

OVIS: "From which station to which station would you like to travel?"

The user responds to the system by giving information on the desired connection, e.g.:

user: "I want to travel from Utrecht to Nijmegen"

The user's utterance is processed by the *Speech Recognition* module. This component converts the incoming speech signal into a sequence of words. The recognized sequence of words are passed to the *Natural Language Processing* module (Bonnema et al., 1997; Van Noord et al., 1999), which searches for relevant information in the sequence of recognized words. Not all words contain relevant information. In the example, *'from Utrecht'*, *'to Nijmegen'* are the relevant words in the sentence, because it can be inferred from these words that *'Utrecht'* is the departure station and that *'Nijmegen'* is the destination station. The *Dialogue Management* module (Veldhuijzen van Zanten, 1998) checks whether the information provided by the user is complete. The departure time is unknown in the example. Thus, the *Dialog Management* module passes a message to the *Natural Language Generation* module (Theune, 2000),

[1] Prioriteitprogramma Taal- en Spraak Technologie
[2] Nederlandse organisatie voor Wetenschappelijk Onderzoek

containing information about the data that is still missing. The *Natural Language Generation* module then formulates a question (text), which is converted into a speech signal by the *Speech Synthesis* module (Klabbers, 2000). Finally, the speech is sent to the *Telephone Interface* and the user will hear the following question:

OVIS: "At what time do you want to travel from Utrecht to Nijmegen?"

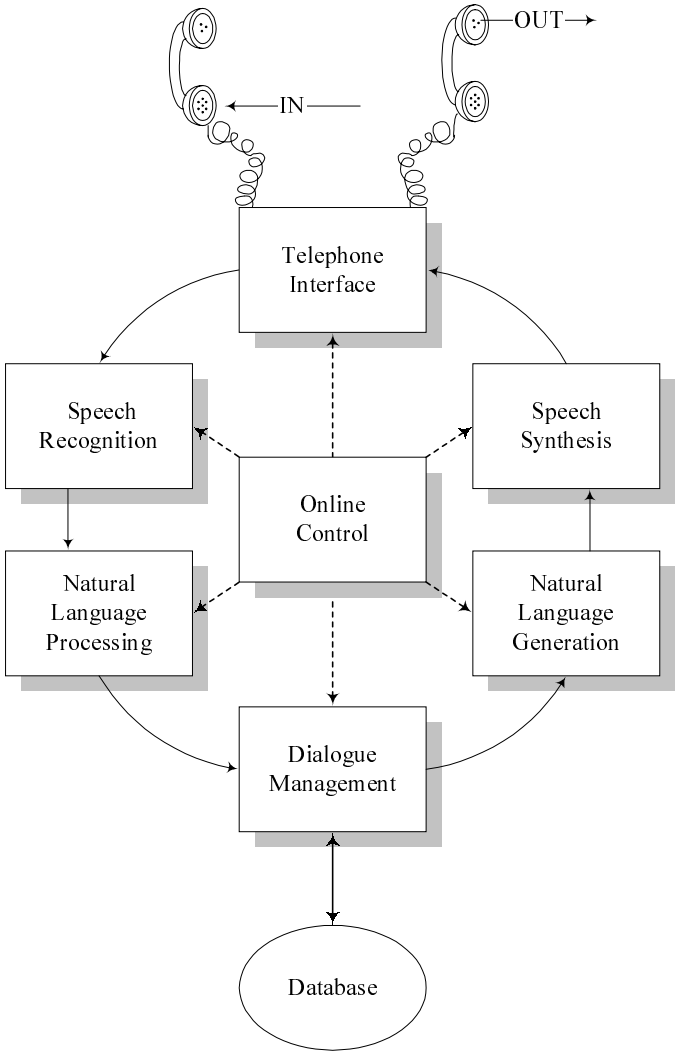

Figure1: Architecture of the OVIS Spoken Dialogue System

Suppose the user answers this question as follows:

user: "I want to depart tomorrow, at eight o'clock in the morning"

Now, the whole process is repeated: the *Speech Recognition* module recognizes the words, the *Natural Language Processing* module searches for the relevant information, and the *Dialogue Management* module checks whether the travel inquiry of the user is completely specified. Since the user has now provided all information, the timetable information is looked up in the *Database*, and the travel advice is formulated (text) by the *Natural Language Generation* module. Finally, the speech

signal generated by the *Speech Synthesis* module is passed to the *Telephone Interface*, resulting in the following travel advice:

> OVIS: "The train from Utrecht to Nijmegen departs at ten to eight from platform eleven. It arrives in Nijmegen at a quarter to nine on platform one."

After giving the travel advice, the system will enquire whether the user wants additional or other information. If the user does not want more information, the system will thank the user and the connection will be closed. Otherwise, the whole process starts all over again.

1.2 The speech recognition component

The research described in this thesis is only concerned with the *Speech Recognition* component of the OVIS system, i.e., a *Continuous Speech Recognizer* (CSR) that converts an incoming speech signal into a corresponding sequence of words (text). Before a CSR can process a speech signal, the signal needs to be converted into a representation that is suitable for automatic speech recognition. Section 1.2.1 describes this conversion, which is called *acoustic pre-processing*. Furthermore, a speech recognizer can only be used if it is trained. The *training* procedure is described in section 1.2.2. Finally, in section 1.2.3, the whole *recognition* process is described in more detail.

1.2.1 Acoustic pre-processing

The most common approach to acoustic pre-processing is to convert the speech waveform into a sequence of acoustic feature vectors, which together form a compact representation of the spectral characteristics of the speech. Figure 2 shows an overview of the acoustic pre-processing that is used in our speech recognizer. The acoustic features that we used are Mel Frequency Cepstral Coefficients (MFCCs).

The first step is to convert the analog speech signal into a digital representation. To this end, the pressure (or voltage) value of the speech waveform is determined at equally spaced time points. The telephone speech that enters the CSR component in OVIS is already digitized. A sample is taken 8 times per ms; thus the sample frequency is 8 kHz (step ① in Figure 2). The second step is to extract the speech waveform for (overlapping) short time intervals; this is called *Time Windowing* (step ② in Figure 2). In our CSR, the acoustic features are calculated every 10 ms for time intervals of 16 ms. The speech signal is also pre-emphasized by applying high-frequency amplification to compensate for the attenuation caused by the radiation from the lips. The next step is the calculation of the spectral characteristics of the speech signal. To this end, a *Fast Fourier Transform* (FFT) of the windowed speech signal is calculated to obtain the FFT-based spectrum (step ③ in Figure 2). Next, *Mel-Scaled Filters* are applied (step ④ in Figure 2). Mel-scaling approximates the frequency resolution of the human ear. In order to make the statistics of the speech power spectrum approximately Gaussian, log compression is applied (step ⑤ in Figure 2). A Discrete Cosine Transform (DCT) is applied to the filterbank outputs in order to decorrelate the spectral representation of the speech signal (step ⑥ in Figure 2). By

using the DCT, the number of spectral parameters representing the speech is reduced, but as much as possible of the relevant information is retained. For historical reasons the result of the DCT is called *cepstral coefficients*. Finally, the first 14 cepstral coefficients and the 14 corresponding time differentials (⑦ in Figure 2) are retained, thus obtaining a 28-element acoustic vector.

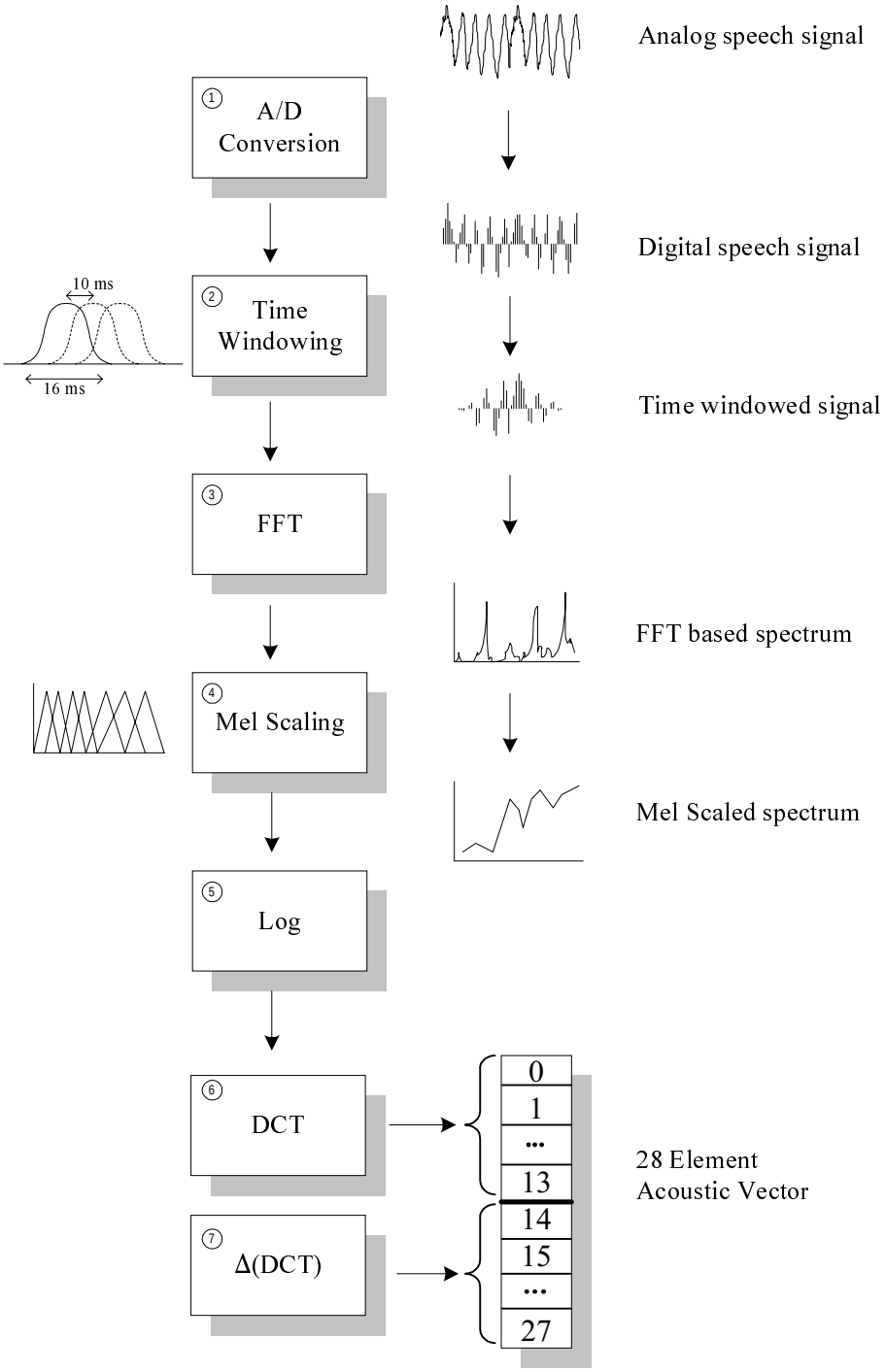

Figure 2: Acoustic pre-processing for obtaining MFCC -based feature vectors.

1.2.2 Training

Nowadays, CSRs are probabilistic engines. This means that a CSR calculates the probability of a word sequence *W* given the acoustic signal *X: P(W|X).* From among all possible word sequences, the word sequence $\hat{W}$ with the highest probability is the one that is recognized:

$$\hat{W} = \underset{W}{argmax}\ P(W \mid X) \tag{1}$$

According to Bayes' theorem, this can be written as follows:

$$\hat{W} = \underset{W}{argmax}\ \frac{P(X \mid W)P(W)}{P(X)} \tag{2}$$

Since $P(X)$ is independent of W:

$$\hat{W} = \underset{W}{argmax}\ P(X \mid W)P(W) \tag{3}$$

This means that maximizing *P(W|X)* is equal to maximizing the product of the following two probabilities:

- $P(X|W)$*:* The probability of observing a sequence of acoustic vectors given the hypothesized sequence of words. This term can only be computed after the fact, i.e., after the observation of a specific speech signal.
- $P(W)$*:* The probability of observing the hypothesized sequence of words. This probability is independent of the observed acoustic vectors. Therefore, this term represents the *prior* probability.

The statistical model that we use to estimate the acoustic probability $P(X|W)$ is called the *acoustic model*. Acoustic models are trained for all basic sound units of Dutch and for some non-speech sounds (see Appendix A). The basic sound units are very similar to what linguists call the *phonemes* of a language. As the Dutch phonemes /l/ and /r/ have different acoustic properties depending on their position in the syllable (post- or pre-vocalic), we distinguish between two types of /l/ and /r/. Different realizations of the same phoneme are also called *allophones*. The term *phones* is used to refer to the basic sound units in this thesis, as it covers both allophones and phonemes. The procedure for training the acoustic models is schematically presented in Figure 3.

In order to estimate the parameters of the acoustic models, it is necessary to have a large amount of recorded speech material with corresponding *orthographic transcriptions*. Orthographic transcriptions describe the words that are spoken in each utterance (text). In addition to the training material, a training lexicon is needed, which lists all words occurring in the training material together with a *phone transcription*.

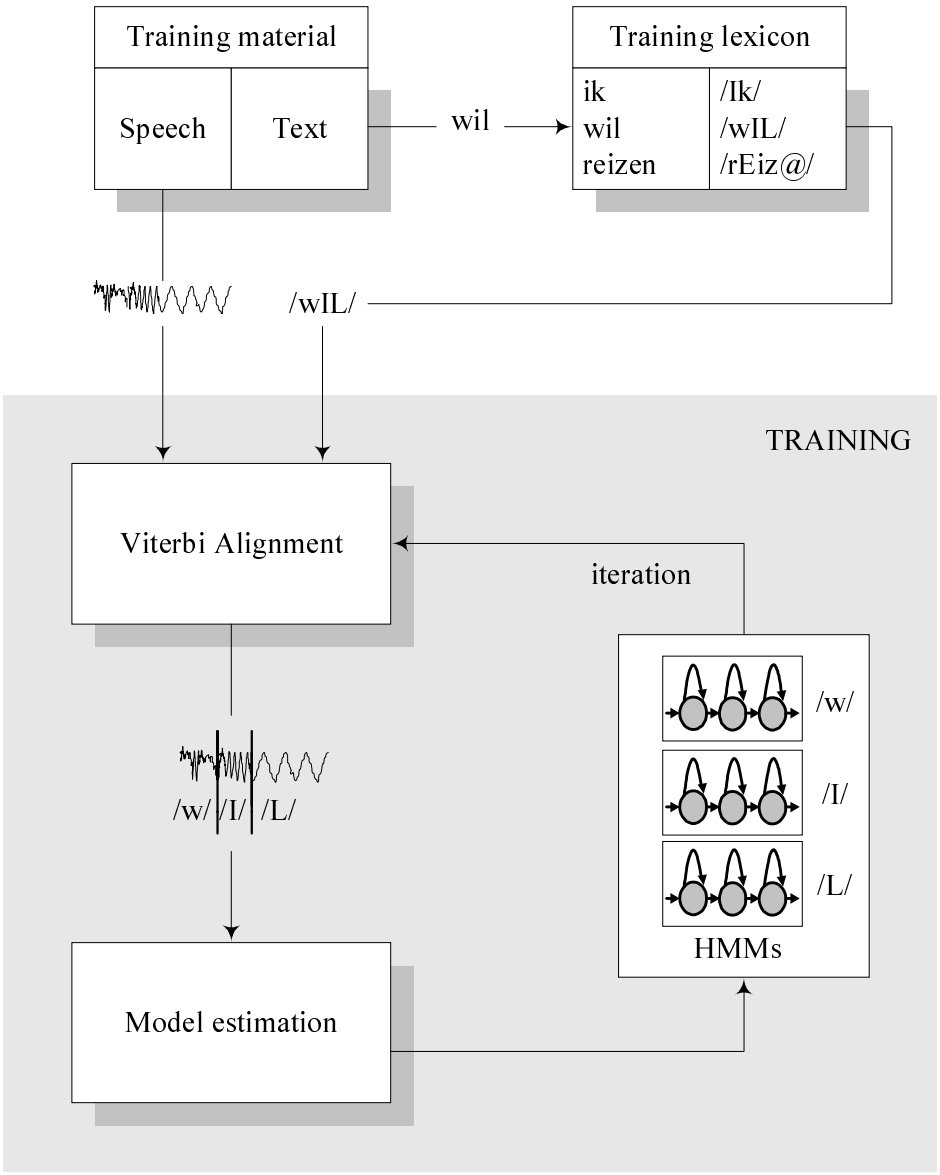

Figure 3: Training the acoustic models

A phone transcription is the sequence of phones that represents the pronunciation of the word. For each utterance in the training material, the phone transcriptions of the words are looked up in the training lexicon. The speech signal together with the concatenated phone transcriptions of the individual words serve as input for training the acoustic models. The training procedure consists of the following two steps:

- **Viterbi alignment.** The goal of alignment is to segment the speech, i.e. given the speech signal and corresponding phone transcription it is determined which parts of the speech signal corresponds to which phone in the phone transcription. An efficient algorithm for finding the optimal alignment is *the Viterbi algorithm*. The Viterbi algorithm finds the optimal alignment based on maximal (acoustic) likelihood.

- **Model estimation.** After alignment, all parts of the speech material that correspond to the same phone are statistically processed. This results in a stochastic model - called a *hidden Markov model* (HMM) - for each basic recognition unit (see Appendix A). Each HMM consists of a sequence of states connected by arcs. Each state consists of an N-dimensional probability density function (pdf), where N is the number of elements in the acoustic vectors. To obtain reliable estimates of the parameters of the pdfs, it is necessary to use a large number of realizations of each phone.

Since no HMMs are available for the calculation of the acoustic likelihoods the first time the Viterbi alignment is made, one usually starts with a linear segmentation, i.e. each phone is assigned an equal duration. Based on this linear segmentation, the initial HMMs are estimated. These HMMs can subsequently be used to make a new Viterbi alignment. Next, the HMMs can be re-estimated based on the new alignment. During each iteration, the likelihood that the models generate the observations increases. The process continues until the likelihood improvements drop below a certain threshold or until a pre-defined number of iterations is reached.

The prior probability $P(W)$ is estimated by the *language model*. A simple but effective way of doing this is to use N-grams, in which it is assumed that the probability of a word is dependent on the previous (N-1) words. To estimate the N-gram probabilities, large amounts of text data are usually used. Our CSR uses a unigram (N=1) and bigram (N=2) language model, which are estimated from the orthographic transcriptions of the training material. The material to train the language model should ideally be recorded with an online version of the application. However, this is a circular problem since a language model is needed in order to be able to use the application. In order to solve this problem, a bootstrap method is often used. This means that an initial language model is constructed (for instance manually), and using this initial language model new material is collected that can subsequently be used to improve the language model. Both the acoustic models and the language model were bootstrapped in OVIS (see section 1.3). In research situations, a speech database is usually divided into two parts: The first part is used to train the acoustic models and the language model (*training material*), whereas the second part is used for recognition experiments (*test material*).

1.2.3 Recognition

Once the acoustic models and language model of the CSR have been trained, the CSR can be used for recognition. An overview of the recognition process is given in Figure 4. Since the CSR can only recognize words that are present in the lexicon, the lexicon needs to contain all the words that one can expect to be used by the people who are addressing OVIS. For instance, all train station names and all days of the week are included in the OVIS recognition lexicon. During the recognition phase, the CSR attempts to recognize an unknown sequence of words. To this end, all possible sequences of words allowed by the lexicon and the language model are generated. If all possible sequences of words had to be evaluated for the full duration of the

utterance, the computational requirements would be prohibitive. Therefore, all hypotheses are scored according to their likelihood. This score is a combination of two scores: the acoustic score determined by the HMMs, and the language model score. The majority of the hypotheses are less likely than the best one, and therefore they can safely be removed from the list of possible solutions. Finally, the output of the recognition process is the most likely sequence of words.

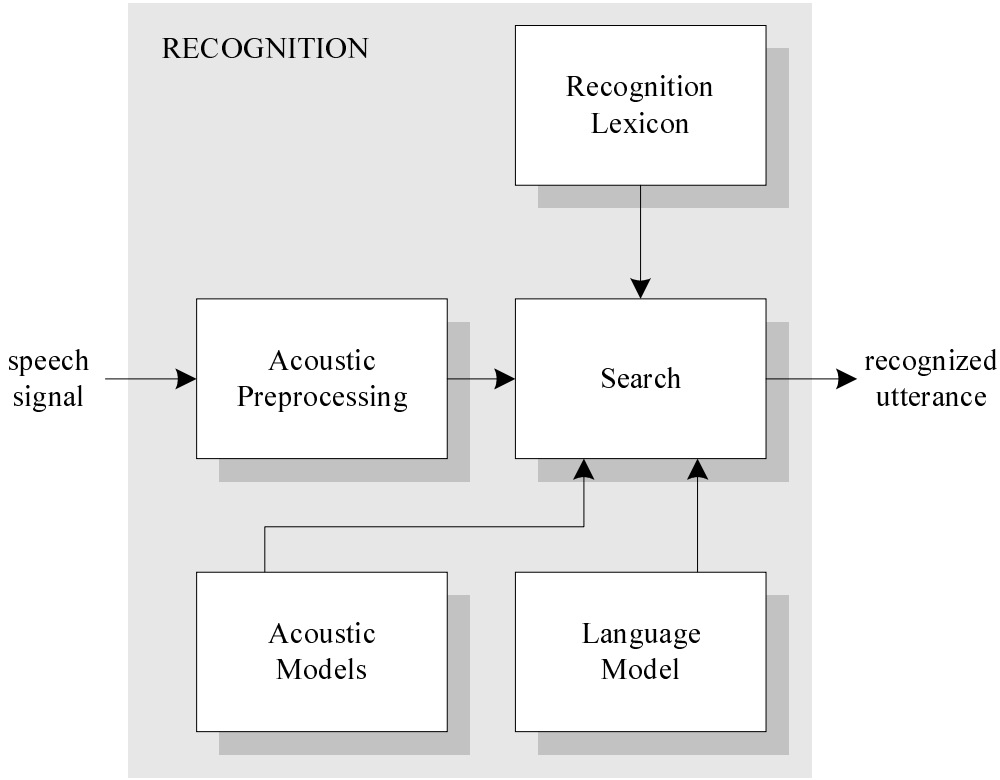

Figure 4: Recognition process

1.3 The VIOS speech material

The first version of OVIS was put into use in December 1995. This version was gradually improved by means of a bootstrap method (Strik et al., 1997). The first version of the phone models was trained using 2,500 Polyphone utterances (den Os et al., 1995). The initial language model was trained on answers of people who addressed a version of OVIS in which - instead of speech - text was used as input for the system. Next, a small group of people received the telephone number of OVIS and were requested to call it regularly. Whenever a sufficient amount of new data was collected, language models and acoustic models were retrained. In this way, the acoustic and language models were gradually improved. From April to June 1997, new speech material was recorded. During this recording period, people from all over the Netherlands were invited to call the system. Compared to the people who called the first version of OVIS, this second group of people is much more heterogeneous and also more representative of the potential users of the OVIS system. The database that

is recorded with OVIS is called ‘VIOS’. The VIOS material was orthographically transcribed by native speakers of Dutch. The output of the CSR (the sequence of words with the highest score) was used as a starting point for transcription by manually correcting it if necessary. The sex of the speakers was determined by the transcribers through auditory impression. Table 1 summarizes the main characteristics of the OVIS speech material.

Table 1: Characteristics of VIOS material

name	VIOS 1	VOIS 2	VIOS 1+2
recording period	Dec. ‘95 – Jun. ‘96	Apr. ‘96 – Jun. ’97	Dec. ‘95 – Jun. ’97
# dialogs	3,531	7,190	10,721
# utterances	33,471	65,929	99,400
# words	108,844	184,513	293,357
# phones [1)]	431,536	718,282	1,149,818
male speakers	57%	75%	68%
female speakers	42%	19%	29%
other speakers [2)]	1%	6%	3%
speech	45%	42%	43%
silence	55%	58%	57%
total duration	25.0 hours	42.6 hours	67.6 hours

[1)] based on canonical transcriptions, [2)] children, mixed speakers or speaker sex unknown

Table 2 shows the selections of the VIOS material used for the various experiments. In the column ‘VIOS’ the recording period is denoted (see Table 1): ‘1’ denotes the first recording period and ‘2’ the second. The column ‘#utts’ shows the number of utterances. For the recognition experiments, the test set perplexity is given in the column “PP”. In article 3, the same material was used for the recognition experiments as for error analysis, whereas in article 4 different sets of material were used. No overlap exists between the material used for training and performing the recognition experiments and the material used for error analysis.

Table2: Selections of VIOS material used in the four articles

	training		recognition experiments			error analysis		phonetic transcriptions	
	# utts	VIOS	# utts	VIOS	PP	# utts	VIOS	# utts	VIOS
article 1	25,104	1	-	-	-	-	-	186	1+2
article 2	25,104	1	482	1+2	33	-	-	482	1+2
article 3	25,104	1	6,276	1	30	6,276	1	-	-
article 4	59,640	1+2	19,880	1+2	28	19,880	1+2	-	-

Figure 5 shows the cumulative proportion of the total VIOS material as a function of word frequency rank. It can be seen that the 10 most frequent words make up about 40% of the total material. They are all short words consisting of one syllable (‘nee’, ‘ja’, ‘naar’, ‘uur’, ‘van’, ‘ik’, ‘wil’, ‘om’, ‘u’, ‘dank’).

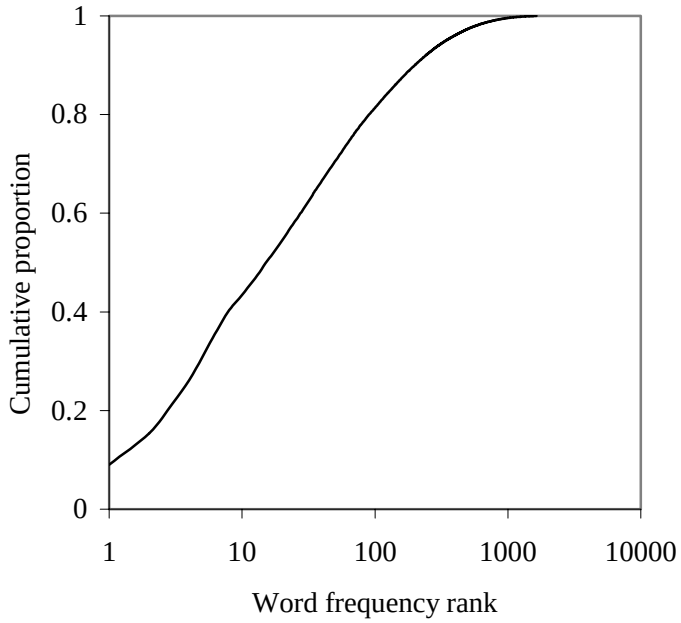

Figure 5: Cumulative distribution of word frequency rank

Figure 6 shows the cumulative proportion of the VIOS utterances as a function of utterance length, i.e. the number of words per utterance. It can be seen that 40% of the utterances consist of a single word.

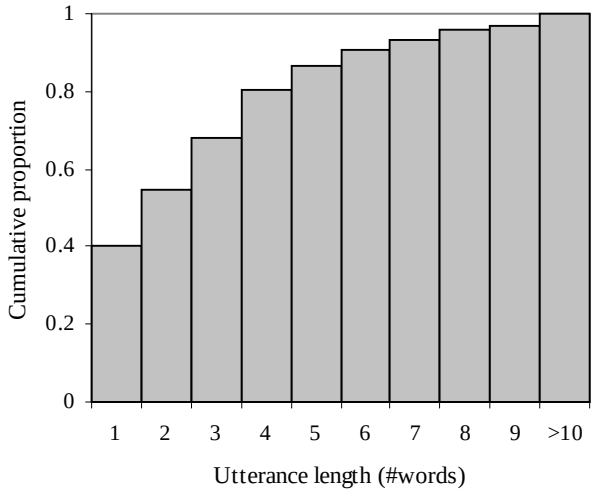

Figure 6: Cumulative distribution of utterance length

2 Pronunciation Variation

When listening to the VIOS speech material, it is immediately clear that words are pronounced in many different ways. The fact that words can be pronounced differently depending on various factors is called *pronunciation variation*. Different sources of pronunciation variation can be distinguished. The distinction given here is adopted from Strik and Cucchiarini (1999).

2.1 Sources of pronunciation variation

A first major distinction can be drawn between *interspeaker* and *intraspeaker* pronunciation variation. *Inter*speaker variation refers to variation in pronunciation of *different* speakers, whereas *intra*speaker variation refers to pronunciation variation of the *same* speaker. To a large degree *inter*speaker variation is caused by anatomical differences between speakers. For example, male and female speakers and children have different speech characteristics. Interspeaker variation also exists due to the fact that speakers of the same language may speak different dialects or speak with a different accent (Laver, 1994). The accent will depend on factors such as region of origin, socioeconomic background, level of education, sex and age. In addition to the factors mentioned so far, another important source of variation is the interlocutor, since it is known that speakers are influenced by the person they are talking to. The interlocutor is a computer in OVIS. However, part of the callers to the system seem to behave as if they were talking to a human being. For instance, people say: "I don't want to go there, madam[3]".

*Intra*speaker pronunciation variation also depends on many different factors. The first factor is the extent to which words are connected to each other. If words are pronounced in isolation, there is almost no interaction between the words. Furthermore, people tend to articulate isolated words more carefully. On the other hand, in connected speech all sorts of interactions may take place such as assimilation, co-articulation, reduction, or deletions and insertions of phones. The degree to which these phenomena occur will vary, depending on the style of speaking. As speech becomes less formal, the syllable structure of words may be reorganized, and there may be changes in pitch and loudness (Laver, 1994, pp. 66-69). The VIOS data show that the manner in which people address the system varies, ranging from very sloppy articulation to hyper-articulation. Another factor that influences the way people speak is their emotional state (Murray and Arnott, 1993). This source of variation is present in the VIOS material. For instance, if the system misunderstands what has been said, people tend to get irritated, which influences the way they speak. A non-linguistic factor that influences the way people speak is background noise. People tend to speak differently in the presence of background noise (Lombard effect). Some environmental noise is present in the VIOS material, e.g. music, other people talking, car noise, or noise due to a low-quality telephone connection.

[3] "Daar wil ik niet naar toe, mevrouw"

The OVIS speech recognizer is an example of a recognizer that handles extemporaneous speech. Furthermore, OVIS can be called nation-wide, and speakers of different sex and age call the system. For all of these reasons it is clear that all the above-mentioned factors that influence the way people speak will vary over a wide range for the VIOS speech material. Therefore, the VIOS material is a good framework for studying the effect of pronunciation variation on the performance of ASR systems. In the next section, we will explain why pronunciation variation is problematic for ASR.

2.2 Why is pronunciation variation problematic for ASR?

In the baseline system, both the lexicons for training and recognition contain a single phone transcription for each word. This phone transcription is the most likely pronunciation according to the linguistic literature and is called the *canonical* phone transcription. Using a lexicon with only one phone transcription per word leads to suboptimal performance when words are not pronounced canonically: Fosler-Lussier (1999, pp.63-64) and McAllaster et al. (1998) showed that the word accuracy on Switchboard data is 11-12% lower for the words that are not pronounced canonically. This degradation in recognition performance is caused by a mismatch between the actual pronunciation of the word and the pronunciation as denoted in the lexicon. This mismatch causes problems both during recognition and training.

To explain why the mismatch is problematic, an example of a non-canonical pronunciation is given in Figure 7. Suppose that the canonical pronunciation of the Dutch city '*Delft*' is /dElft/ (phone transcriptions are given in SAMPA[4] notation). An example of a non-canonical pronunciation is /dEl@f/. In the realized pronunciation, the speech sound /@/ is inserted between the /l/ and the /f/, and the final /t/ is deleted.

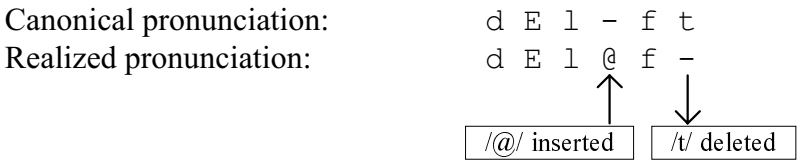

Figure 7: Example of a non-canonical pronunciation (/dEl@f/)

During recognition, the total acoustic score of the realized pronunciation of the word *Delft* (/dEl@f/) is lower than it would have been if the spoken phone sequence had been exactly equal to the canonical phone transcription in the lexicon (/dELft/). The acoustic scores for /l/ and /f/ are likely to be lower, because the part of the acoustic signal that is used to calculate an acoustic score for the phones /l/ and /f/ contains the acoustic signal of the inserted /@/. Furthermore, the acoustic score for the /t/ is also likely to be lower since no /t/ is pronounced, and consequently, parts of the speech

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

signal of the /f/ may be used to calculate the acoustic score for the phone /t/. Figure 8 shows the mismatch due to the non-canonical pronunciation /dEl@f/. The parts of the realized pronunciation that do not match the canonical pronunciation are indicated with grey; these parts of the speech signal will have a low acoustic score.

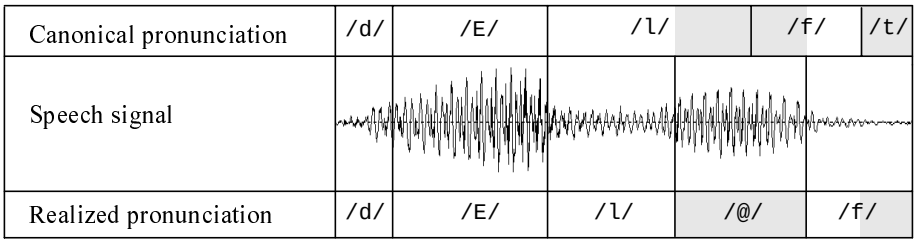

Figure 8: Mismatch due to non-canonical pronunciation (/dEl@f/)

The low acoustic score that is assigned to the mismatching parts might degrade recognition performance. Suppose that there is another word in the lexicon that does not differ very much from the word *Delft*, for example *elf* [5] (/Elf/). This word differs only in two phones from the realized pronunciation (/dEl@f/), since the deletion of the /d/ and /@/ in /dEl@f/ results in the pronunciation /Elf/. If the acoustic score for the word, /Elf/, is higher than the acoustic score of /dELft/, the incorrect word /Elf/ might be recognized.

During training, the mismatch between the actual pronunciation and the canonical pronunciation in the training lexicon will result in contaminated acoustic models. Let us consider the same example and suppose the realized pronunciation is /dEl@f/. During training, the canonical phone transcription is looked up in the training lexicon. Next, the Viterbi algorithm is used to align the canonical transcription with the speech signal. Suppose the alignment between the speech signal and the canonical phone transcription is as given in Figure 8. Since the HMMs are trained on the Viterbi alignments, this means that three HMMs may become contaminated: the /l/-HMM and /f/-HMM are contaminated with parts of the speech signal of the inserted /@/, and the /t/-HMM is contaminated with the acoustic signal of the /f/ or with the acoustic signal of the word that follows 'Delft' in the utterance. The contamination of the HMMs might lead to recognition errors as the contaminated HMMs are less discriminative.

The kind of pronunciation variation described in this section is an example of segmental variation: It can be described as substitutions, insertions and deletions of phones. In this thesis, pronunciation variation is described in this way. In other words; the pronunciation variation is modeled at the level of the phones. Alternatives to phone level pronunciation variation modeling will be discussed in section 5.2.3.

[5] eleven

2.3 Overview of methods to model pronunciation variation in ASR

Strik and Cucchiarini (1999) give an overview of the literature on modeling pronunciation variation for ASR. It is difficult to give a precise definition of pronunciation variation for ASR. Strictly speaking, one could say that almost all ASR research is about modeling pronunciation variation. For example, HMM modeling is a way of accounting for segmental and temporal variation. In this section, the research described in this thesis will be positioned in the categorical framework Strik and Cucchiarini (1999) present.

A first distinction is based on whether the pronunciation variation occurs within words or across word boundaries. In article 3, we modeled both within-word and cross-word variation. We started off by modeling *within-word variation*. The CSR used in this research employs a single-pass search. This type of decoding helps to limit computing time, but one of the limitations of strict single-pass search is that it is difficult to model cross-word processes. To model *cross-word variation* we employed two methods that can be used in our single-pass decoder. To this end, we selected frequent word sequences from the VIOS-material. Next, a number of phonological cross-word phenomena were applied to these word sequences in order to obtain cross-word variants. For the first method, the cross-word variants of individual words in the word sequences were added to the lexicon. For the second method, the word sequences were joined together, thus forming multi-words, and the multi-words and their variants were added to the lexicon.

A second distinction that Strik and Cucchiarini (1999) make concerns the source from which the information on pronunciation variation is retrieved. Two types of information sources are distinguished: In knowledge-based studies, information on possible pronunciation variation is primarily derived from sources that are already available in the literature. In data-driven studies, the information on possible pronunciation variation is obtained from the speech in the training database. In this thesis both approaches are used. Article 3 describes a knowledge-based method of modeling pronunciation, whereas article 4 concerns a data-driven method of modeling pronunciation variation.

A third distinction that is made concerns the information representation. The information about pronunciation variation can be formalized or not. In general, formalization means that a more abstract and compact representation is chosen. Data-driven information on pronunciation variation can be formalized, e.g. by rewrite rules, artificial neural networks, phone confusion matrices, or decision trees. We use rewrite rules in the data-driven approach described in article 4. In knowledge-based studies, the information on possible pronunciation variation obtained from linguistic studies can be formalized in the form of phonological rules. In general, these are optional phonological rules concerning deletions, insertions, and substitutions of phones. We used five optional phonological rules in the knowledge-based approach described in article 3. The obvious alternative to using formalizations is using an approach in which all possible variants are generated without recourse to some form of rules. Generation can be a manual process, or transcriptions observed in a database can be used. The most important difference between using formalizations or not is the way in which

variants can be generated for a specific task. One of the advantages of using formalizations is that variants can be generated for unseen and new words. However, a disadvantage of employing formalizations is possible undergeneration and overgeneration of variants. (Cohen, 1989; Strik and Cucchiarini, 1999)

The last distinction that Strik and Cucchiarini (1999) make concerns the level of modeling. Most CSRs consist of three levels: the lexicon, the phone models and the language model. Pronunciation variation is modeled at all these three levels in this thesis. Section 3.2.2 explains the general method of modeling pronunciation at all three levels of the CSR. This general method is used both in the knowledge-based and in the data-driven methods.

For both the knowledge-based and the data-driven method, information is needed on the frequency and identity of the pronunciation variants that occur in the training data. In order to obtain this information on pronunciation variation, usually phonetic transcriptions of the training material are made. These transcriptions can be obtained manually, but the use of automatically obtained phonetic transcriptions is becoming more common (Strik and Cucchiarini, 1999). An important advantages of making automatic phonetic transcriptions is that it is less time-consuming, and therefore, less costly than making manual transcriptions. Another argument in favor of automatic transcriptions is that they are more in line with the phone strings obtained later during recognition in the system (see Riley et al., 1999). For these reasons, we used automatically obtained phonetic transcriptions in the research reported in this thesis. A detailed analysis of the automatic transcription procedure is presented in article 1 and article 2.

Strik and Cucchiarini (1999) observe that in most studies the emphasis is on reduction of the error rates. In order to find out how and why improvements are obtained, recognition errors should be studied in more detail, i.e. a more detailed error analysis should be carried out. In the research reported in this thesis, we do not limit ourselves to measure performance improvement in terms of Word Error Rate (WER), but an attempt is also made to understand how the recognition process is affected by modeling pronunciation variation. To this end, error-analysis of the results of modeling pronunciation variation is performed; see article 3 and article 4.

3 Goals and methodology

3.1 Goals

The first goal of this thesis is to investigate whether the performance of ASR can be improved by explicit modeling of segmental pronunciation variation. Besides improving recognition performance we also hope to gain more insight into the effect of modeling pronunciation. Furthermore, since automatic phonetic transcription of pronunciation variants forms a vital component of the research methodology, a second goal is to assess the quality of our automatic transcriptions and to investigate how they may best be obtained. In the next two sections, the method for obtaining automatic phonetic transcriptions and the general method for modeling pronunciation variation is explained.

3.2 Methodology

3.2.1 Automatic phonetic transcription

Phonetic transcriptions are needed for two purposes in the research described in this thesis. First of all, for the data-driven method to model pronunciation variation, the pronunciation variants need to be obtained. To this end, phonetic transcriptions of the training material are made. Second, our general method of modeling pronunciation variation requires information on the occurrence of the pronunciation variants to be modeled. In order to obtain this information, phonetic transcriptions are also needed. In this thesis, the phonetic transcriptions are made automatically, i.e. by a speech recognizer. Almost invariably, the automatic phonetic transcriptions are ‘broad phonetic’, or phonemic transcriptions.

Automatic phonetic transcriptions can be made in several ways. One approach that has been used is to perform phone recognition. In this kind of recognition, phones are recognized instead of words. The recognizer is often constrained by a phone N-gram, and by penalties on the generation of sequences comprising many short phones. However, the content of speech (the orthographic transcription) is often available. In this case, the corresponding canonical phonetic transcription can be used as a starting point for automatic transcription. The phonetic transcription is looked up in a lexicon. Based on this phonetic transcription a limited number of possible pronunciation variants are generated by applying some kind of rules, e.g. phonological rules (e.g. Lamel and Adda, 1996), data-derived rules (e.g. Kessens et al., 2000), or by using decision trees (e.g. Riley et al., 1999). The task of the CSR is now to decide for each word which of the possible variants best matches the acoustic signal. This approach to obtaining automatic phonetic transcriptions is called *forced recognition* (or *forced alignment*) and is used in this thesis.

During forced recognition/alignment a Viterbi alignment of the speech material is made for all possible sequences of pronunciation variants and the sequence of variants with the highest likelihood is chosen. If the prior probabilities of the pronunciation variants of the same word are exactly equal during forced

recognition/alignment, the choice for a specific variant is determined solely by the acoustic likelihoods. However, sometimes weighted prior probabilities for the pronunciation variants are used during forced recognition/alignment. Kipp et al. (1997) use manually labeled data in order to obtain the prior probabilities for the pronunciation variants. Riley et al. (1999) and Saraçlar (2000) use the pronunciation probabilities derived from decision trees as weights during alignment. In our automatic transcription procedure, the CSR is forced to choose between the various pronunciation variants by using an utterance specific language model. This language model is trained for each individual utterance on a corpus consisting of 100,000 repetitions of the utterance. In this way, the weight of the language model is largely increased, making it virtually impossible to recognize other words than the ones present in the utterance. During forced recognition, all variants of the same word are assigned equal prior probabilities, thus the choice for a specific pronunciation variant is solely determined by the acoustics.

For the data-driven method to model pronunciation variation, forced recognition is performed twice. The first time, forced recognition is performed in order to obtain transcriptions of the pronunciation variation occurring in the training material. The pronunciation variants that can be chosen during forced recognition are obtained by starting with a canonical phone transcription for each word. Next, a very large number of hypothetical variants are generated. This is done by generating all possible variants in which one or more phones in the canonical phone transcription are deleted. For each utterance, the automatic transcriptions are aligned with the concatenation of the canonical transcriptions of the words in the utterance. On the basis of these alignments, data-driven rules are derived. Next, the data-driven rules are selected and used to generate pronunciation variants. Subsequently, the resulting set of variants is used in a second forced recognition that is carried out to obtain information on the frequency of occurrence of the variants. With the same aim, forced recognition is performed for the knowledge-based method. The pronunciation variants are automatically obtained by applying five phonological rules to the canonical transcriptions of the words in the lexicon. In the next section, it is explained how the information on the occurrence of the variants is used in our general method to model pronunciation variation.

3.2.2 Modeling pronunciation variation

For both the knowledge-based (article 3) and the data-driven approach (article 4), a general method of modeling pronunciation variation was used. This method implies incorporating pronunciation variation at all three levels in the CSR: the lexicon, the phone models, and the language model.

- Modeling pronunciation at the level of the lexicon:
  Pronunciation variants are added to the baseline recognition *lexicon*. In this way, a lexicon is obtained that contains multiple pronunciations for some of the words. By using the multiple pronunciation lexicon, we expect recognition performance to

improve, because the mismatch between the realized pronunciation and the pronunciation in the lexicon is reduced.

For modeling pronunciation variation at the other two levels of the CSR, an extra step is needed. This step consists of obtaining automatic phonetic transcriptions of the training corpus by performing forced recognition (see section 3.2.1).

- Modeling pronunciation variation at the level of the phone models:
  The *phone models* are retrained on the new automatic phonetic transcriptions of the training corpus. Since we expect that there will be less mismatch between the new phone transcriptions and the acoustic signals, the retrained phone models should be less contaminated, and should therefore perform better.

- Modeling pronunciation variation at the level of the language model:
  A new *language model* is calculated from the new automatic transcriptions of the training corpus. In the baseline language model, all pronunciation variants of the same word are assigned equal prior probabilities. However, in the new language model, different variants of the same word are assigned their own specific probabilities. These probabilities are estimated from the automatic transcriptions of the training corpus.

3.2.3 Evaluation

The first objective of this thesis is to improve the recognition performance of our CSR by modeling pronunciation variation. As a measure of recognition performance, we used the WER, which is defined as follows:

- $$\text{WER} = \frac{\text{S} + \text{D} + \text{I}}{\text{N}} \text{x}100\% \qquad (4)$$

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 second objective of this thesis is to assess the quality of our automatic phonetic transcriptions and to investigate how they can best be obtained. As a measure of quality, we used agreement between the automatic phonetic transcriptions and human reference transcriptions; the higher the agreement, the better the quality of the automatic phonetic transcriptions. As a measure of agreement we used Cohen's kappa (κ), which corrects for chance agreement (Cohen, 1968):

- $$\text{Cohen's } \kappa = \frac{P_o - P_c}{100 - P_c} \qquad (5)$$

P_o = percentage of agreeing pairs of judgements (observed agreement)
P_c = percentage of agreeing pairs on the basis of chance (expected agreement)

where, P_c is calculated as follows:

- $$P_c = \sum_{i,j=1}^{v} P_{.j} P_{i.} \text{ x } 100\% \qquad (6)$$

$P_{i.}$ = marginal fraction of row i ($n_{i.}/N$)
$P_{.j}$ = marginal fraction of column j ($n_{.i}/N$)
N = number of judged objects
v = number of categories

When the distribution of scores across the different categories substantially differs from uniformity, P_c is high. The examples given in Table 6 clarify this point. Example A shows a situation in which much more 0-s than 1-s are used. In this case, P_c is 90.5%. Example B shows a situation in which the 0-s and 1-s are more uniformly distributed: P_c is 50%.

Table 6: Two examples of distributions of scores amongst the two judgers (humans and CSR)

A	*humans*		
CSR	*i j*	*0*	*1*	$P_{i.}$
	0	*.9*	*.05*	*.95*
	1	*.05*	*0*	*.05*
	$P_{.j}$	*.95*	*.05*	

B	*humans*		
CSR	*i j*	*0*	*1*	$P_{i.}$
	0	*.45*	*.05*	*.5*
	1	*.05*	*45*	*.5*
	$P_{.j}$	*.5*	*.5*	

In the extreme case that all objects are assigned to the same category, κ cannot be calculated. The values of κ range from –1 (total disagreement) to 1 (perfect agreement).

4 Summaries of the articles

4.1 Summary 1

"Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer", published in Language & Speech 44 (3), pp. 377-403.

In this article, we investigate whether a continuous speech recognizer (CSR) can be used to obtain automatic phonetic transcriptions of speech. The automatic transcriptions were made by using the CSR in forced recognition mode. During forced recognition, the CSR chooses the variant that best matches the acoustic signal from among a number of possible pronunciation variants. The pronunciation variants were automatically generated by applying the following five optional phonological rules to the words in the baseline lexicon: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion (Booij, 1995; Cucchiarini and van den Heuvel, 1999). Two experiments were carried out in which the performance of the CSR was compared to the performance of expert listeners. However, given that human listeners can make mistakes it is not possible to obtain a completely error free human reference transcription with which the automatic transcriptions can be compared (Cucchiarini, 1993). To (partly) circumvent this problem, two strategies were used to obtain a human reference transcription. In the first experiment, a *majority vote* procedure was used, i.e., the reference transcription is based on the judgment of the majority in a group of listeners. In the second experiment, a *consensus* transcription was made, i.e., two (or three) transcribers have to agree on each individual symbol to be transcribed. For evaluation, binary scores were derived: "1" if the rule was applied, or "0" if the rule was not applied. As a measure of agreement we used Cohen's kappa (κ) (see section 3.2.3).

Experiment 1

The manual transcriptions in the first experiment were made by nine expert linguists who all have experience in making phonetic transcriptions for their own investigations. The transcription task was exactly the same for the transcribers and the CSR, namely a forced choice from among a number of possible pronunciation variants. For the 291 words for which the variants had to be chosen, 467 binary scores per subject were obtained. Different reference transcriptions were obtained, depending on the minimum number of listeners that had to agree. The transcriptions for which the minimum number of agreeing listeners is not reached were excluded from analysis. This means that the higher the minimum number of listeners, the stricter the reference transcription, and the more transcriptions are excluded from analysis.

Four types of comparisons were performed in which the CSR's transcriptions were compared to the transcriptions made by the linguists. First, a pairwise comparison was performed for each pair of listeners and for each CSR-listener pair. This comparison showed that the agreement values for six of the listeners do not differ significantly from each other, whereas the agreement values of two listeners are

significantly higher and those of one of the listeners and the CSR are significantly lower than the rest. The average κ-value for the listener-listener pairs is 0.63, whereas the average κ-value for the CSR-listener pairs is 0.55. Second, we compared the listeners' and the CSR's transcriptions to reference transcriptions with varying degrees of strictness. We found that the CSR's agreement values increased if a stricter reference transcription was used. In a third comparison, we investigated the agreement values for the individual rules. This comparison revealed that the results are rule-dependent: the absolute agreement values for both listeners and CSR vary per rule; the differences in agreement values between CSR and listeners vary per rule, and the range in agreement values for the listeners is quite variable per rule. Finally, in a fourth comparison we examined the differences in transcriptions between the listeners and the CSR. We found that the human transcribers scored a phone as present more often than the CSR did. As we hypothesize that this difference might be of durational nature and as the difference is especially large for the /@/-deletion rule, we determined the duration of the /@/s in the context of the /@/-deletion rule based on an automatic segmentation of the transcription material. This analysis showed that half of the /@/s with a very short duration are detected by the humans, but not by the CSR. This result indicates that the human listeners and the CSR may have a different durational threshold for detecting the /@/ in the context of the /@/-deletion rule.

Experiment 2

In order to investigate why the results were quite different for /@/-deletion as opposed to /@/-insertion, we conducted a second experiment. To this end, consensus transcriptions were made for words for which the /@/-deletion and /@/-insertion rule are applicable. Five duos and one trio were asked to reach consensus on the transcription (using IPA[6] symbols) of what was articulated at the indicated spot in the word, i.e., where the conditions for application of the rule were met. The transcribers were students who had all followed the same transcription course. Comparison of the consensus transcriptions with the automatic transcriptions revealed that most of the /@/s that have a short duration according to the listeners were denoted as 'not present' by the CSR. This is further evidence that the listeners and the CSR may have different durational thresholds for detecting the phone /@/. Furthermore, we found that for the /@/s in the context of the /@/-deletion rule often something other than deletion or /@/ was transcribed, indicating that /@/-deletion is a more variable process than /@/-insertion.

The two experiments conducted in this study revealed that overall the machine's transcription performance is significantly different from the listener's performance. However, if we consider the individual rules, not all differences appeared to be significant. Furthermore, it should be kept in mind that significant lower agreement values were also found for one of the listeners. Although there are significant differences between the CSR and the listeners, the difference in performance may be acceptable, depending on what the transcriptions are needed for.

[6] IPA=International Phonetic Alphabet, see http://www2.arts.gla.ac.uk/IPA/fullchart.html

4.2 Summary 2

"On automatic phonetic transcription quality: Lower WERs do not guarantee better transcriptions", submitted to Computer, Speech & Language.

In this study, we investigated a number of issues related to the quality of automatic phonetic transcriptions obtained by using the CSR in forced recognition mode. The pronunciation variants were automatically generated by applying the same five phonological rules as in article 1 to the words in the canonical lexicon. For each phone that could possibly be deleted or inserted, a binary score was obtained: (1) if the rule was applied and (0) if this was not the case. As a quality measure of the automatic transcriptions, we used agreement between the automatic transcriptions and the human reference transcriptions: The higher the agreement with the human reference transcriptions, the better the quality of the automatic transcriptions. As in article 1, Cohen's kappa (κ) was used as a measure of agreement (see section 3.2.3).

Both majority vote and consensus reference transcriptions have been used. The majority vote reference transcriptions were identical to those in article 1; thus, in total, 467 binary scores were obtained. The consensus transcriptions were made by the same students as in article 1. However, a difference is that the transcriptions in this study are made for whole utterances. In total, 770 binary scores were obtained from these utterances, as the context for one of the five rules applying was met 770 times.

Properties of a CSR versus transcription quality

The first goal of this investigation was to determine how various properties of a CSR affect the quality of the resulting automatic transcriptions. The properties of the CSR that were investigated are all related to the acoustic models (HMMs). The first property concerns the HMM topology. In article 1, we found indications that the human listeners and the CSR have a different durational threshold for detecting the phone /@/. Furthermore, Brugnara et al. (1993) found that a better phone accuracy is obtained when HMMs are used with a minimum duration that is shorter than the duration of our baseline HMMs. For these reasons, we investigated whether agreement could be increased by using an HMM topology for the /@/ that has a shorter minimum duration than the baseline /@/-HMM. The results show that the CSR does indeed detect more /@/s when the HMM with a shorter minimum duration is used. However, the increase in the overall agreement values is not very large.

The second property that we investigated concerns the degree of contamination in the HMMs. Since the speech material used for training contains a great deal of variation in pronunciation, but the baseline training lexicon contains only one canonical transcription for each word, the HMMs are contaminated. One of the approaches we used to reduce the contamination due to this mismatch is retraining the phone models using automatic transcriptions of the training material which were obtained through forced recognition in previous research (Wester et al., 1998a). If we use the HMMs from this research to make automatic transcriptions, the overall agreement values improved. Similar results have been reported by Saraçlar (2000). A

second way of reducing the part of the mismatch between the transcription of the training material and the actual pronunciation is to take the most frequently occurring pronunciation to train the HMMs. These HMMs also improved the overall agreement values. A third way of reducing the mismatch is to train the HMMs on read speech instead of on spontaneous speech. As the amount of variation in spontaneous speech tends to be larger than in read speech, it is to be expected that HMMs trained on read speech will also be less contaminated. Our results indeed show that higher agreement values are found for the read speech HMMs.

The third property that we investigated concerns the type of HMMs, namely context-independent (CI) versus context-dependent (CD) HMMs. Since CD-HMMs generally yield lower WERs, one could expect that CD-HMMs would also improve transcription quality. Compared to using CI-HMMs, the agreement values for the CD-HMMs deteriorated for the majority vote material, whereas a small improvement was found for the consensus material. The deterioration in agreement for the majority vote material is mainly caused by the /r/-deletion rule. Using CD-HMMs, the CSR unjustly detects more /r/s. The different /r/-deletion results for the majority vote and the consensus material are probably related to the fact that the words for which the transcriptions of /r/-deletion were made are considerably different in the two types of material.

Finally, we also investigated the effect of combinations of properties. If CD-HMMs are trained on automatic transcriptions (obtained through forced recognition) instead of training them on canonical transcriptions, the contamination within the CD-HMMs is reduced and the quality of the transcriptions is improved. The combination of two other properties, namely pronunciation variation modeling and using a ‘short’ HMM for the phone /@/, also resulted in a further improvement of transcription quality.

In this study, the quality of the automatic transcriptions was evaluated by measuring agreement with human reference transcriptions based on a majority vote principle and with consensus reference transcriptions. For the majority vote transcriptions, the overall κ-value (all rules) varies between 0.46 and 0.63. For the consensus transcriptions, the overall κ-value varies between 0.43 and 0.51. The difference in absolute agreement values might be explained by the transcribers’ differences in level of experience, by the fact that the focus in the two transcription tasks was different, and by differences in the number of transcribers that were used to obtain the reference transcriptions. Although the absolute agreement values varied for the two types of reference transcriptions, the general trends that we observed were very similar. To conclude, we have shown that changing the properties of a CSR can improve the quality of the automatic transcriptions produced. Furthermore, we found that by combining these changes in properties the quality of automatic transcription can be improved even further: The κ-values could be improved by 0.08 for the consensus transcriptions, and by 0.125 for the majority vote transcriptions.

WER versus transcription quality

Intuitively one might expect that the CSR that obtains the lowest WER on some reference recognition task will also yield the best automatic transcriptions. However, on second thoughts, speech recognition may well be quite a different task from automatic transcription. Therefore, our second goal was to investigate whether lower WERs do indeed predict higher quality automatic transcriptions. We observed that there is no clear relation between the WER obtained with a certain CSR and its transcription quality. Therefore, we can conclude that for obtaining automatic transcriptions, taking the CSR with the lowest WER is not always the optimal solution. Rather, one should concentrate on the properties that the CSR should have in order to make optimal transcriptions. The best thing to do is to use a CSR that is optimized for making automatic transcriptions.

4.3 Summary 3

"Improving the performance of a Dut ch CSR by modeling within-word and cross-word pronunciation variation", published in Speech Communication 29, pp. 193-207.

Modeling within-word and cross-word pronunciation variation

This article describes how the performance of our CSR was improved by modeling within- and cross-word pronunciation variation. We propose a general procedure for modeling pronunciation variation (see section 3.2.2). In short, it consists of adding pronunciation variants to the lexicon, retraining the phone models and using variant-specific (prior) probabilities. Within-word pronunciation variants were generated by applying the same five phonological rules as in article 1 to the words in the lexicon. These rules all concern frequent phonological processes. The type of cross-word processes we focused on were contraction, reduction and cliticization (Booij, 1995). It is not straightforward to model variation that occurs across word boundaries in our recognizer, as it uses a single pass search algorithm. Therefore, we employed two methods to model cross-word variation suitable for our single-pass decoder. In the first method a limited number of cross-word processes were modeled by directly adding the cross-word variants to the lexicon. The second method models cross-word variation by using multi-words. We tested the within-word and cross-word methods in isolation, as well as the combinations of the within-word method with each of the cross-word methods.

The recognition experiments that we conducted yielded the following results. We measured a WER of 12.75% for the baseline system in which no pronunciation variants were used. Adding pronunciation variants to the lexicon (without changes elsewhere in the system) did not always result in an improvement of recognition performance. When, on top of adding variants to the lexicon, retrained phone models are used, the WERs for almost all approaches (and combinations of approaches) are improved compared to using the baseline phone models. However, retraining the phone models does not alleviate all the deterioration that is caused by the expansion of the lexicon: Compared to the baseline system, there are still deteriorations. When, in addition to retraining the phone models, variant-specific prior probabilities are

employed, the WERs for all methods improve. Moreover, all WERs are lower than in the baseline system, and the absolute improvements are generally larger than the improvements obtained through using multiple variants in the lexicon or retraining the phone models. These results indicate that employing prior probabilities for the variants is essential when pronunciation variants are added to the lexicon. Comparison of the two cross-word methods revealed that cross-word method 2 performs better than cross-word method 1. The better recognition performance of cross-word method 2 can mainly be attributed to the use of multi-words in the language model, as multi-words increase the span of the language model for the word sequences in the training and test material that occur most frequently. Finally, our results showed that the combination of the within-word method and the cross-word method 2 led to the best results: A total improvement in WER of 1.1% absolute, or 8.8% relative was obtained.

Combination versus isolation

In this article, we also compared the recognition results of the cross-word methods tested in isolation and tested in combination with the within-word method. Furthermore, we tested the five within-word rules in isolation, and we compared the results of these tests to the recognition result of the experiment in which the combination of all five rules is tested. This investigation revealed that the results obtained for testing various sets of pronunciation variants in isolation did not add up to the result of testing the combination of the sets of variants. This is due to a number of factors. First of all, different rules can apply to the same word. Consequently, when the five rules are used in combination, pronunciation variants are generated which are not generated for any of the rules in isolation. Furthermore, the words in the utterances are not recognized independently of each other; thus, interaction between pronunciation variants can occur. The implication of these findings is that it will not suffice to study the effect of modeling pronunciation variants in isolation. Instead, combinations of pronunciation variants have to be studied. However, this poses a practical problem, as there are many possible combinations.

Error analysis

In many studies about modeling pronunciation variation, WER is used as the only measure for performance evaluation. Although this measure gives a global idea of the merits of a method, it certainly does not reveal all details of the effect a method has. This became clear through the error analysis that we conducted, since it revealed that 14.7% of the recognized utterances changed, whereas a net improvement of only 1.3% in the sentence error rate was found. Therefore, it is clear that a more detailed error analysis is necessary to gain real insight into the effects of a certain approach.

4.4 Summary 4

"A data-driven method for modeling pronunciation variation", submitted, reviewed and resubmitted to Speech Communication.

In this article, we describe a data-driven method for modeling pronunciation variation. For two reasons, the kind of pronunciation variation that we modeled was limited to deletion processes. First, we expected deletions (and insertions) to be more important than substitutions, since substitutions can implicitly be modelled in the phone models. Second, deletion processes occur frequently in our speech material (see Wester et al., 1998c). The deletion processes were described by rules that were derived in a data-driven manner. In the first step of the rule-extraction procedure, we generated all possible deletion variants by allowing each phone in the canonical transcription to be deleted. The variants generated in this way are used during forced recognition which was carried out in order to determine which of the possible variants best matches the acoustic signal (see section 3.2.1). The second step was an alignment of the automatic phone transcriptions with a concatenation of canonical phone transcriptions of the words in the utterance. The alignments were used to formulate candidate rules which describe the contexts in which the phones are deleted. Finally, 91 rules were selected by excluding the rules that have a low absolute frequency of rule application (F_{abs}). The main reason for selecting the frequent rules is to filter out rules that might be based on transcription errors. Since it can be expected that transcription errors occur randomly, the rules that are the result of transcription errors are probably not as frequent as the rules that are based on genuine deletion processes.

Reduction in WER through data-driven modeling of pronunciation variation

The first goal of this study is to find out whether the data-driven method used for modeling pronunciation variation leads to improved recognition performance. We tested different subsets of the 91 rules by selecting them based on the relative frequency of rule application (F_{rel}), which is defined as the number of times a rule was applied (F_{abs}) divided by the number of times the rule could have been applied. We started off by testing the rules with the highest F_{rel} and gradually increased the number of rules by lowering the threshold for F_{rel}. We employed the same general method of modeling pronunciation variation as in article 1 (see section 3.2.2). The pronunciation variants were generated by applying the selected set of rules to the words in the lexicon. For the baseline system, in which no pronunciation variants are used, the WER is 16.94%. This WER is higher than the WER for the baseline system in article 3. Two explanations can be given for this difference. First, the test set contains out-of-vocabulary words in this study, whereas this is not the case in article 3. Second, the test set is mainly taken from the second set of VIOS data, in which the variability in the speakers is much larger than in the first set, from which the test set in article 3 is taken (see section 1.3). The recognition experiments revealed results comparable to those presented in article 3. As in article 3 we found that only adding pronunciation variants to the lexicon can deteriorate recognition performance. If the number of added variants is small, the WERs improve compared to the baseline. However, with an

increasing number of added variants, the improvements become smaller until a deterioration in WER is found. This deterioration rapidly increases as a function of the number of added variants. Furthermore, as in article 1, we found that retraining the phone models is only of little benefit. The WERs slightly improve when, in addition to expanding the lexicon, the phone models are retrained. When variant-specific prior probabilities are also used, the WERs improve. For all sets of variants, improvements are found compared to the baseline system using variant-specific prior probabilities. For the best testing condition, a total improvement in WER of 1.2% absolute or 7.3% relative was found. To conclude, our data-driven method of modeling pronunciation variation indeed leads to improvements in recognition performance, provided that prior probabilities for the variants are used.

Error analysis

The second goal of this research is to find out how the changes in WER came about, by performing an error analysis procedure. In this investigation, error analysis was performed at word level, whereas in article 3 error analysis was performed at sentence level. A commonality is that we found that besides improvements, also deteriorations were introduced through the modeling of pronunciation variation. These deteriorations are almost as large as the improvements, so that the total net improvement in SER/WER is small. The current error analysis also gave some new results. Two-thirds of the words that were recognized differently were not recognized as one of the added pronunciation variants. For the other one third of differently recognized words, we could determine which rules caused the change in recognition result. On the basis of this analysis, we determined the number of improvements and deteriorations per rule. A strong correlation between the number of improvements and deteriorations per rule was found, indicating that it is not possible to improve performance by excluding the rules that cause many deteriorations, because these rules also produce a considerable number of improvements. Finally, we found that the contribution to the changes in WER differs per rule. The total improvement in WER could be ascribed to one quarter of the rules. The most important rule was the deletion of word final /n/ preceded by a /@/[7]. To conclude, our error analysis reveals that the gain in recognition performance could be improved by making the balance between introducing and solving errors more positive. However, this cannot be achieved by excluding rules that introduce many errors.

Three criteria for rule selection

The third goal of this study was to examine the adequacy of three criteria for rule selection. In this way, it would be possible to make more sound choices about which rules (or which pronunciation variants) to select. To this end, the following three measures were calculated:

[7] This rule parallels the /n/-deletion rule used in the knowledge-based approach (article 3)

1) F_{abs},
2) F_{rel}, and
3) 'net result of variants'.

The 'net result of variants' was obtained as follows. For the differently recognized words that were recognized as a variant, we determined which rule(s) generated the variant. In this way, it is possible to determine the total number of improvements and deteriorations per rule. The 'net result of variants' is defined as the difference between the number of improvements and the number of deteriorations. The 'net result of variants' is more difficult to obtain, since an error-analysis is necessary to calculate this measure. F_{abs} and F_{rel} are relatively easy to obtain, since they can be calculated from the automatically obtained phone transcriptions of the training material. In order to test the adequacy of the three measures, we selected sets of rules on the basis of the three criteria and measured WER on an independent test set. Next, we calculated the correlation between each of the criteria and the measured WERs. This correlation was highest for F_{abs} (0.92) and 'net result of variants' (0.85). Since F_{abs} is easier to obtain, this measure is to be preferred as a criterion for rule selection. To conclude, our results indicate that rules can best be selected based on the absolute frequency of application (F_{abs}). By selecting the rules on the basis of F_{abs}, in the best testing condition, a total improvement in WER of 1.4% absolute, or 8.2% relative was obtained.

5 Discussion

5.1 Automatic phonetic transcription

The first two articles in this thesis concern automatic phonetic transcription of speech. The differences between automatic versus manually obtained transcriptions will be discussed in section 5.1.1. Next, in section 5.1.2, I will discuss the differences between making automatic transcriptions and performing a normal recognition task. Finally, the last section explains some application areas for automatic transcription.

5.1.1 Automatic versus manual phonetic transcription

The results of the research described in this thesis show that there are differences between the transcriptions made by human transcribers and the transcriptions made the CSR: The average κ-value for the listener-listener pairs was 0.63, whereas the average κ-value for the CSR-listener pairs was 0.55 (article 1). Furthermore, we showed that changing the properties of the CSR can make the CSR's transcriptions more similar to the human transcriptions: The κ-values could be improved by 0.08 for the consensus transcriptions, and by 0.125 for the majority vote transcriptions (article 2). Although we showed that the degree of agreement between phonetic transcriptions made by humans and automatic transcriptions can be diminished, I think it is not possible to eliminate all differences completely.

A reason for believing that there will always be differences between manual and automatic transcriptions is that humans do not even agree on which transcription is 'the correct one'. In the first experiment of article 1, for instance, inter-listener agreement varied between 75% and 82% (Wester et al., 1998b). Kipp et al. (1997) found an inter-labeler agreement ranging from 79% to 83% (Verbmobil corpus). For the Switchboard corpus, inter-labeler agreement (at the phonetic segment level) between highly experienced transcribers varied between 72% and 80% (Greenberg, 1999). If humans do not agree, it cannot be expected that the CSR is able to produce a transcription that can be expected to be the 'correct one'.

One of the reasons why making phonetic transcriptions of speech is so difficult (for both humans and the CSR), is that the continuously changing signal has to be divided into discrete non-overlapping segments. In non-linear, auto-segmental phonology, a representation has been proposed in which speech is represented by many *parallel* tiers, representing the parallel activities of the articulators in speech that do not necessarily begin and end simultaneously (Goldsmith, 1990). Some other authors state that speech cannot be described fully in terms of sequential units (see e.g. Greenberg, 1997).

Besides the non-sequential and continuous character of speech that poses problems for both humans and the CSR in making phonetic transcriptions, there are also differences in the way transcriptions are made by human listeners and CSRs. First of all, humans and CSRs analyze the speech signal differently. For example, Strik

(2001) states that several important assumptions for signal analysis are made in standard CSRs, which do not correspond to the way humans perceive speech. For instance, the analysis window for which feature values are calculated is usually very short. Although some dynamic information can be obtained from the derivatives of these features, humans may very well rely on information from a wider time span for speech recognition. A second factor that causes differences between automatic and manual phonetic transcriptions is that human listeners are influenced by various factors, for instance, spelling, phonotactics, semantics, fatigue and level of experience (for an overview see Cucchiarini, 1993). These factors do not influence the transcriptions made by the CSR, or when they do (as is usually the case for phonotactics) they are likely to have different effects.

5.1.2 Automatic transcription quality versus recognition performance

One of the main conclusions of article 2 is that there is no clear relation between the WER obtained with a certain CSR and its transcription quality. Saraçlar (2000) reported similar results showing that better quality transcriptions do not always lead to improved WERs. In my view, these results are not surprising, since automatic transcription and automatic recognition are completely different tasks. For automatic transcription, we performed forced recognition. The phone-level constraints applied during forced recognition are different than the word-level constraints applied during a normal recognition task. The sequences of words that can be recognized during a normal recognition task are constrained by the lexicon and the language model. During forced recognition, the phone sequences for each word are restricted to the pronunciation variants contained in the lexicon. Consequently, other causes of errors play a role during a normal recognition task than during forced recognition: For instance, lexical confusability is a great source of errors in a normal recognition task (words are recognized incorrectly), whereas lexical confusability cannot cause errors during forced recognition. However, I do not think automatic transcription quality and recognition performance are completely uncorrelated. In order to make good quality transcriptions a certain level of recognition performance is necessary, and the other way around, a CSR that performs very badly is useless for making high quality transcriptions.

5.1.3 Application areas of automatic phonetic transcription

Although there are significant differences between the CSR and the listeners, the difference in performance may be acceptable, depending on what the transcriptions are needed for. The question that arises then is for what applications our automatic transcription tool can be used. It is obvious that it cannot be used to obtain phonetic transcriptions from scratch, but it is clearly limited to hypothesis verification. A first application of our automatic transcription tool, of course, is our research on modeling pronunciation variation (articles 3+4). Second, it could be used in various fields of linguistics, like phonetics, phonology, sociolinguistics, and dialectology. In practice,

this tool could be a useful aid in all research situations in which phonetic transcriptions have to be made by one person, since this tool could resolve possible doubts about what was actually realized. Given that a CSR does not suffer from fatigue and loss of concentration, it could assist the transcriber who is likely to make mistakes owing to concentration loss. By comparing his/her own transcriptions with those produced by the CSR a transcriber could spot possible errors that are due to absent-mindedness. Furthermore, a transcriber may be biased by his/her own hypotheses and expectations with obvious consequences for the transcriptions, while the biases for the automatic tool may be controlled. Checking the automatic transcriptions may help discover possible biases in the listener's data. Finally, an important contribution of automatic transcription to linguistics would be that it makes it easier to use existing (very large) speech databases for the purpose of linguistic research. With this tool, large amounts of material can be analyzed in a relatively short time (about 2x real time), and at relatively low costs. Although the CSR is not infallible, the advantages of a very large dataset might very well outweigh the errors introduced by the occasional mistakes of the CSR.

5.2 Modeling pronunciation variation

The last two articles in this thesis concern pronunciation variation modeling. In section 5.2.1, the general method that we employed for modeling of pronunciation variation will be discussed. Next, in section 5.2.2, it will be discussed why the improvements that we found were small. Finally, alternatives to phone level modeling of pronunciation variation will be discussed in section 5.2.3.

5.2.1 General method of modeling pronunciation variation

The general method of modeling pronunciation variation consisted of incorporating pronunciation variation at all three levels of the CSR (i.e. the lexicon, the phone models and the language model). In this section, the results of modeling pronunciation variation at each of the levels will be discussed. An essential part of our general method is to make automatic transcriptions of the training material. The new transcriptions are used to re-estimate the phone models and the language model. This process can be repeated iteratively. To this end, the retrained phone models are used to make new transcriptions. The new transcriptions are used in turn to train new phone models and to re-estimate the prior probabilities of the variants. The results of iteration will be discussed in the last paragraph of this section.

Adding variants to the lexicon

Adding pronunciation variants to the recognition lexicon without changes elsewhere in the system was not always beneficial to recognition performance (articles 3+4). In article 4 we found that, if a small number of variants are added to the lexicon, recognition performance improves, but with an increasing number of added variants the gain in recognition performance becomes smaller. Above a certain number of added variants (an average of 2.5 variants per word), a deterioration in recognition

performance is found. This deterioration rapidly increases as a function of the number of added variants. These results are comparable with the results of Yang and Martens (2000) and Fukada (1999).

Retraining the phone models

In the studies reported in this thesis, recognition performance generally improved if in addition to expanding the lexicon, the phone models were retrained. Several other authors also found improvements in recognition performance by retraining the phone models (e.g. Aubert and Dugast, 1995; Lamel and Adda, 1996; Riley et al., 1999). However, the improvements in recognition performance are generally not very large. Other authors even found deteriorations in recognition performance when retrained phone models were used (Beulen et al., 1998; Wester, 2001).

Our research (Kessens et al., 1997; Wester et al., 1998a) and other research (e.g. Lamel and Adda, 1996; Schiel et al., 1998) revealed that retraining the phone models is only beneficial if the pronunciation variants which are used during training are also used during recognition (by including variants in the lexicon). This result can be explained as follows. By retraining the phone models, part of the contamination within the phone models disappears. Consequently, the phone models can better discriminate between various pronunciation variants. However, during recognition this greater discriminative ability cannot be used, since no alternative pronunciation variants are present in the lexicon. Moreover, if a word is not pronounced canonically, the acoustic likelihood scores for the mismatching parts of the speech are probably lower than the acoustic likelihood scores obtained with the 'contaminated' baseline phone models. Therefore, the risk of the recognition of an incorrect word is increased.

Incorporating pronunciation variants in the language model

The difference between incorporating pronunciation variants in the language model or not is that in the first case the variants are assigned their own specific prior probabilities, whereas in the second case each variant is assumed to be equally likely. The level of the CSR in which prior probabilities are used is system dependent. For instance, in our system the prior probabilities are defined in the language model, whereas prior probabilities can also be defined in the lexicon (see e.g. Fosler-Lussier, 1999; Wester and Fosler-Lussier, 2000). The results reported in this thesis show that using prior probabilities for pronunciation variants is crucial when modeling pronunciation variation. We found that the positive effect of adding variants to the lexicon is much larger when prior probabilities are assigned to the variants. A possible explanation for the importance of employing variant-specific probabilities is as follows. By adding variants to the lexicon, a number of recognition errors are solved, as the variants match the actual pronunciation for some of the words better. On the other hand, new errors are introduced because lexical confusability increases. By treating each pronunciation variant as being equally likely, the damage done by the increase in lexical confusability is probably large, since the probabilistic framework of the speech recognizer is violated: Pronunciation variants of frequently occurring words are assigned high prior probabilities, despite the fact that they may be highly unlikely.

Consequently, these variants might introduce more errors than they correct. Many other authors have reported on the importance of prior probabilities for pronunciation variants (e.g. Fukada et al., 1999; Peskin et al., 2000; Saraçlar, 2000, pp. 118; Yang and Martens, 2000; Jurafsky et al., 2001).

Iteration

The results of our research (Kessens and Wester, 1997; Kessens et al., 1999) show that iteration only has small effects on recognition performance: Most of the changes in the transcriptions and WERs occur the first time an improved transcription is made. After the first iteration, the transcriptions and WERs do not change very much. Beringer and Schiel (1999) calculated phone error rates of automatic transcriptions (compared to manual transcriptions). As no further improvements in phone error rates were observed in later iterations, Beringer and Schiel conclude that the process of iterative transcription converges after the second iteration. To conclude, the process of iterative transcription seems to converge very fast (after one or two iterations).

5.2.2 Why are the improvements so small?

In this thesis we have shown that recognition performance can be improved by modeling pronunciation variation at the level of the phones. However, the improvements obtained were, in general, not very large (relative reductions of 8-9% in WER). This observation is not restricted to the research reported in this thesis, but seems to be a general finding among the researchers in the field of pronunciation variation modeling. In 1998, an ESCA workshop “Modeling Pronunciation Variation for ASR” was held at Rolduc, Kerkrade, in the Netherlands. As a result of this workshop a special issue of Speech Communication was published. The relative reductions in WER reported in that journal issue ranged between 0 and 20% (Strik and Cucchiarini, 1999). Since then,only Yang and Martens (2000) reported larger improvements (30-45% relative WER reduction). However, the results of Yang and Martens are found for read speech material (TIMIT); such large reductions in WER have not yet been obtained for spontaneous speech. There are a number of factors that could explain why the improvements due to modeling pronunciation variation are generally not very large. These factors will be discussed below.

One of the factors that play a role is that not all variants that occur in the test set are included in the lexicon (*undercoverage*), and the other way around: variants that do not occur in the test set are included in the lexicon (*overcoverage*). Saraçlar (2000) performed ‘cheating’ experiments that revealed that if one were able to construct a lexicon that has no undercoverage and overcoverage, a relative reduction in WER of 19% can be obtained. Similar results have been reported by McAllaster et al. (1998). These authors performed recognition experiments on simulated speech data fabricated from the acoustic models. Using the simulated data, a relative reduction of 24-42% in WER can be obtained if a lexicon is used that contains all and only the variants in the test set.

In the two approaches for modeling pronunciation variation used in this thesis, the degree of mismatch between the lexicon and the test sets is different. One of the

drawbacks of knowledge-based modeling of pronunciation variation is that the knowledge on pronunciation variation that can be found in the literature is incomplete (see e.g. Strik and Cucchiarini, 1999). In our data-driven method, the information on the pronunciation variation is derived from exactly the same kind of speech as the material that is used for the recognition experiments. Consequently, the coverage is expected to be better. However, since we only concentrated on deletion processes, not all variation in the data is covered. Moreover, both in the knowledge-based and the data-driven approach we used rules to generate possible variants. One of the advantages of using rules is that they generalize to unseen contexts and that they are not corpus/task dependent. A disadvantage of employing rules is possible undergeneration and overgeneration of variants due to incorrect specifications of the rules applied (Cohen, 1989; Strik and Cucchiarini, 1999). To conclude, I hypothesize that the coverage for both approaches could be improved; for the knowledge-based method by modeling more pronunciation variation, and for the data-driven method by extending the method to substitutions and insertions of phones and by refining the way the rules are defined (e.g. by using more context information).

Coverage is not the only (and maybe not even the most important) factor that plays a role. This is suggested by the results of the error analysis that we performed in this thesis (articles 3+4). Despite the fact that we used prior probabilities for the pronunciation variants (thus reducing the negative effects of overcoverage), new errors are also introduced due to the addition of pronunciation variants: These deteriorations counterbalance part of the improvements, so that only a small total net improvement in SER/WER is obtained. A possible explanation for the introduction of new errors is *lexical confusability*: (sequences of) pronunciation variants of incorrect words are confused with (sequences of) correct words. Some researches have tried to estimate the amount of lexical confusability of pronunciation variants (Sloboda, 1995; Torre et al., 1997; Wester and Fosler-Lussier, 2000). There are various reasons that could explain why attempts to reduce confusability do not always translate to large reductions in WER. First of all, lexical confusability will always exist, since homophony (and near homophony) is part of the language. By excluding confusable variants, the benefits that these variants could have for recognition performance also disappear. Second, it is difficult to find a measure that takes completely into account all factors that explain lexical confusability. For instance, the confusability measure of Wester and Fosler-Lussier (2000) only takes into account confusions between words that exactly match (parts of) other words, whereas most of the recognition errors concerns confusions of words that do not exactly match.

In addition to coverage and lexical confusability, there is a third factor that partly explains why the improvements in recognition performance are not very large for our method of pronunciation variation modeling. An implicit assumption in our method is that pronunciation variation can be modeled at the level of the phones. This means that a phone can be deleted, substituted or inserted; no intermediate models are used. This way of modeling pronunciation is obviously a simplification of what actually happens, since changes in pronunciation are not discrete, but rather gradual in nature (see e.g. Saraçlar, 2000).

5.2.3 Alternatives to phone level modeling of pronunciation variation

A way of partially circumventing the problems connected with phone level modeling of pronunciation variation is by modeling the variation implicitly in the acoustic models. In this way, the level of modeling the pronunciation variation has shifted from the phone level to the level of states or densities. One way of implicitly modeling pronunciation variation in the phone models is by using context-dependent (CD) phone models. The recognition results presented in article 2 of this thesis (Figure 16) show that the amount of improvement obtained by using context-dependent phone models is about equal to the improvement in recognition result obtained with a combination of context-independent phone modeling and pronunciation variation modeling. Furthermore, in article 2, it was also shown that pronunciation variation in combination with context-dependent phone models does not improve recognition performance. These results are in line with the results of Ma (1998), since Ma showed that the gain in recognition performance from pronunciation variation modeling reduces if CD models are used and if the complexity of the models is increased. However, not all pronunciation variation is well captured by CD models: Jurafsky et al. (2001) showed that phone substitutions and vowel reduction can be adequately captured in CD models, but syllable deletions are poorly modeled.

Another way of modeling pronunciation variation implicitly in the phone models is to use a State-Level Pronunciation Model (SLPM) (Saraçlar, 2000). Saraçlar showed that the improvements in recognition performance are larger for state-level modeling of pronunciation variation than for phone-level modeling, but the differences were not very large. Lee and Wellekens (2001b) also used a SLPM, but the improvements were not very large.

Pronunciation variation can also be modeled implicitly in the acoustic models by using larger basic units than phones, like (demi-)syllables (see e.g. Heine et al., 1998; Wu, 1998; Greenberg, 1999; Ganapathiraju et al., 2001) or even whole word models. In this way, the pronunciation variation contained in the syllable/word is captured within the acoustic model. However, a problem with using larger basic units is that for large vocabulary tasks, the number of syllables/words is much larger than the number of phones. As a consequence, the number of model parameters is also larger, and the danger of under-training increases. Furthermore, these larger basic units do not provide a solution for cross-word or cross-syllable pronunciation variation. These two limitations are probably the reasons why using larger basic units does not often result in large improvements in recognition performance.

An approach that can perhaps be used to describe pronunciation variation in a more appropriate manner is to use articulatory features. Compared to phones, articulatory features provide a more adequate description of pronunciation variation, as the variation can be described in terms of feature spreading and assimilation, instead of categorical phone substitutions, deletions and insertions. Articulatory features are often used as sub-phonemic units, as an intermediate level between the level of the acoustically-based features and the phone level, which makes it necessary to transform the articulatory-based features into phonetic segments. Although high frame-level feature classification accuracies are found, it appears to be difficult to transform the

frame-level transcriptions into phone/word-level transcriptions with higher word accuracy (King et al. 1998; Kirchhoff 1999; Koreman et al. 1999). However, Kirchhoff (1999) has shown that articulatory features provide complementary information to acoustically-based features. This suggests that *combining* articulatory features with other acoustic input could possibly improve pronunciation variation modeling.

6 Conclusions and future work

6.1 Conclusions

Several conclusions can be drawn from the results presented in this thesis. One of the goals of this thesis was to assess the quality of our automatic transcription procedure. The first conclusion is that it is possible to use the CSR for automatic transcription. Whether the differences in performance between the machine and the human transcribers are acceptable, depends on the purpose for which the transcriptions are needed. Furthermore, we conclude that using the CSR with the lowest WER measured on an independent test set does not guarantee that optimal automatic transcriptions are obtained. In order to obtain optimal automatic transcriptions, one should rather concentrate on those properties of the CSR that are important for automatic transcription. The quality of the automatic transcriptions can be improved by using 'short' HMMs and by reducing the amount of contamination in the HMMs. Furthermore, it appeared that CD-HMMs should not be trained on canonical transcriptions, since the transcriptions obtained with these HMMs are too much biased towards the canonical transcriptions. We also found that by combining these changes in properties of the CSR the quality of automatic transcription can be further improved.

Another goal of this thesis was to investigate whether the recognition performance of our CSR could be improved by modeling pronunciation variation at the level of the phones. We conclude that with our general approach to model pronunciation variation it is indeed possible to improve recognition performance. Knowledge-based and data-driven modeling of pronunciation variation led to the same degree of improvement in recognition performance. However, the degree of improvement was generally not very large.

Our general method of modeling pronunciation variation involves all three levels of the CSR. More specific conclusions can be drawn concerning the results of modeling pronunciation variation at each level. First of all, expanding the lexicon by adding pronunciation variants is no guarantee for improved recognition performance. Second, retraining the phone models on (iterative) automatic transcription of the training material has only very small effects on recognition performance. A third important conclusion is that it is crucial to use prior probabilities for the pronunciation variants in order to ensure improvements in recognition performance.

6.2 Future work

6.2.1 Automatic phonetic transcription

In the literature, only few examples of optimising automatic phonetic transcriptions can be found. In my view, more work should be done in that direction. The goal of this kind of research should not be to minimize the differences between automatic and manual phonetic transcriptions, but rather to find out in what respect manual and

automatic transcriptions are different. A question that is worth investigating is whether the differences that we found between the manual and automatic transcriptions are caused by the CSR, the human transcribers, or both, or whether it is not possible to say what caused the difference. In articles 1 and 2, for instance, we showed that part of the difference between the transcriptions made by the CSR and the human transcribers is due to a bias of the CSR towards the deletion of segments. Furthermore, we found indications that part of this bias of the CSR is of durational nature. Another example of research that provided more insight into what respect manual and automatic transcriptions are different is the work of Saraçlar (2000). His work shows that the phone error rate between human and automatic transcriptions dramatically increases (>60%) for the proportion of transcriptions where the human transcribers disagree.

Furthermore, I think it is worthwhile to investigate whether measures can be developed to assess the quality of automatic transcriptions beforehand, i.e. without comparing them to manual transcriptions. In ASR, confidence measures are often used in order to estimate the reliability of correctness of the recognition output. Confidence measures might also appear to be useful in order to estimate the reliability of automatic transcriptions. Using such a kind of measure makes it easier to quantify the differences between automatic transcriptions and manual transcriptions. As a consequence, it will be less difficult to decide in what research situations automatic transcriptions can be used.

6.2.2 Improving pronunciation variation modeling

There are many differences in the way that people and machines perceive and process speech. One of the main differences between human and machine speech decoding is that humans use multiple sources of information in parallel. Linguistic theory assumes that language is represented on many organizational tiers. If information from one of the tiers is damaged or completely missing, human beings tend to use cues from other tiers. In contrast, current ASR-systems focus on just a few of the linguistically relevant tiers. For this reason, many authors have suggested that speech recognition could be improved by performing many parallel analyses at the various linguistic levels, for instance analyses at the articulatory-acoustic, phonological, grammatical, and semantic levels (e.g. Greenberg, 1997; Pols, 1999). Some researchers have already investigated whether using information from other linguistic tiers can help to rule out some of the errors that are introduced by modeling pronunciation variation. For instance, Fosler-Lussier (1999) investigated the dependence of pronunciation variation on word-predictability and speaking rate. Jurafsky et al. (1998) investigated how filled pauses, disfluencies, segmental context, speaking rate and word predictability relate to the realization of the ten most common function words in the Switchboard corpus. Finke and Waibel (1997) have introduced speaking mode as means to reduce confusability by probabilistically weighting alternative variants depending on the speaking style. These studies found correlations between each of the investigated factors and pronunciation variability, but the interactions seem to be interdependent. For instance, Fosler-Lussier (1999) found that a combination of word predictability and speaking

rate can best explain pronunciation phenomena. For this reason, I think it is important to investigate how the various factors that can predict pronunciation variability interact.

Another important difference between human speech perception and the way speech recognizers process speech, is that human speech recognition is much more flexible. Humans continually adapt the prior probabilities in their lexicon depending on various factors like the person(s) they are talking to, the situations they are in, and the state of the conversation. In many speech recognition systems, however, the words that can be recognized (and their corresponding prior probabilities) are fixed. By dynamically adapting the language model (and/or the lexicon), recognition performance can be improved. For instance, in applications like spoken dialogue systems, the language model can be adapted depending on the dialogue state, which results in a decrease in task perplexity and error rates (Popovici and Baggia, 1997; Baggia et al., 1999; Wessel and Baader, 1999).

Besides the positive effects that dynamic modeling has on speech recognition in general, I think it can be especially beneficial to pronunciation variation modeling. A major problem connected with adding pronunciation variants to the lexicon is that lexical confusability is increased. In my view, the best way to combat this lexical confusability is by dynamic modeling of pronunciation variation. An approach to dynamic modeling of pronunciation variation is to perform two-pass decoding. In the second pass, the lattice (or list) of N-best hypotheses from the first pass is expanded with pronunciation variants. Next, the expanded lattice is re-scored and the best hypothesis is selected. Saraçlar (2000) showed that if the lattice is only expanded with pronunciation variants that actually occur in the utterance, recognition performance is considerably improved compared to using a static lexicon; the WER reduced from 38% to 27%. This result shows that a large gain can be expected by dynamic pronunciation variation modeling. Some authors reported small improvements for dynamic modeling of pronunciation variation compared to static modeling (see e.g. Weintraub et al., 1996; Fosler-Lussier, 1999; Lee and Wellekens, 2001b), but recently Lee and Wellekens (2001a) found a much larger relative improvement of 16.7% WER for dynamic versus static modeling of pronunciation variation. In my view, dynamic modeling of pronunciation variation is a promising research direction, especially if it is combined with information from other linguistic tiers (e.g. phone context, speaking rate, word predictability, stress and the presence of disfluencies).

6.2.3 Comparison of methods

In the literature almost no research can be found in which various techniques for modeling pronunciation variation are compared. Strik and Cucchiarini (1999) mention in their overview article that several factors make it difficult to compare methods, namely differences between corpora and ASR systems, differences in the measures used for evaluation, and differences in the baseline system. I agree with Strik and Cucchiarini (1999) that it is advisable to strive towards an objective evaluation of methods. In my opinion, just reporting WERs is not sufficient, as WERs only reveal

the net changes. In order to make it easier to compare the effects of different methods, it is important to separate the effects of the different factors that can influence the amount of improvement that can be obtained with a certain method of modeling pronunciation variation. Questions that could help to compare various methods are the following:

- What is the amount of undercoverage and overcoverage?
- How many changes occur due to pronunciation variation modeling? How many improvements and how many deteriorations? Which part of the errors in the baseline system is affected by pronunciation variation modeling?
- How dependent are the results on the average number of variants per word in the lexicon? Is there an optimum?
- How system and language dependent are the results?
- How corpus dependent are the results? Does the type of speech play a role?

Error analysis as done in this thesis and by others (e.g. Weintraub et al., 1996; Fosler-Lussier, 1999) and ‘cheating’ experiments like McAllaster et al. (1998) and Saraçlar, (2000) may shed more light on the possibilities of different methods of modeling pronunciation variation.

References

Aubert, X. and Dugast, C. (1995). Improved acoustic-phonetic modeling in Philips'dictation system by handling liaisons and multiple pronunciations. *Proceedings of Eurospeech*, Madrid, Spain, 767-770.

Baggia, P., Kellner, A., Pérennou, G., Popovici, C., Sturm, J. and Wessel, F. (1999). Language modelling and spoken dialogue systems - the ARISE experience. *Proceedings of Eurospeech*, Budapest, Hungary, 1767-1770.

Beringer, N. and Schiel, F. (1999). Independent automatic segmentation of speech by pronunciaton modeling. *Proceedings of ICPhS*, San Fransisco, USA, 1653-1656.

Beulen, K., Ortmanns, S., Eiden, A., Martin, L., Welling, L., Overmann, J. and Ney, H. (1998). Pronunciation modeling in the RWTH large vocabulary speech recognizer. *Proceedings of the ESCA workshop "Modeling Pronunciation Variation for Automatic Speech Recognition"*, Kerkrade, The Netherlands, 13-16.

Bonnema, R., Bod, R. and Scha, R. (1997). A DOP model for semantic interpretation. *Proceedings of the ACL/EACL workshop on spoken dialogue systems*, Madrid, Spain, 159-167.

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

Brugnara, F., Falavigna, D. and Omologo, M. (1993). Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication* **12**: 357-370.

Cohen, J. A. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**: 213-220.

Cohen, M. (1989). *Phonological Structures for Speech Recognition*, Ph.D. thesis, University of California. Berkeley, CA, USA.

Cucchiarini, C. (1993). *Phonetic Transcription: a Methodological and Empirical Study*, Ph.D. thesis, University of Nijmegen. Nijmegen, The Netherlands.

Cucchiarini, C. and van den Heuvel, H. (1999). /r/-deletion in Dutch: more experimental evidence. *Proceedings of ICPhS*, San Francisco, USA, 1673-1676.

Davis, K. H., Biddulph, R. and Balashek, S. (1952). Automatic recognition of spoken digits. *Journal of the Acoustic Society of America* **24 (6)**: 637-642.

den Os, E. A., Boogaart, T. I., Boves, L. and Klabbers, E. (1995). The Dutch Polyphone Corpus. *Proceedings of Eurospeech*, Madrid, Spain, 825-828.

Finke, M. and Waibel, A. (1997). Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. *Proceedings of Eurospeech*, Rhodes, Greece, 2379-2382.

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

Fukada, T., Yoshimura, T. and Sagisaka, Y. (1999). Automatic generation of multiple pronunciations based on neural networks. *Speech Communication* **27**: 63-73.

Ganapathiraju, A., Hamaker, J., Ordowski, M., Doddington, G. and Picone, J. (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing* **9 (4)**: 358-366.

Greenberg, S. (1997). On the origins of speech intelligibility in the real world. *Proceedings of the Workshop on "Robust Speech Recognition for Unknown Communication Channels"*, Pont-a-Mousson, France, 23-32.

Greenberg, S. (1999). Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* **29**: 159-176.

Heine, H., Evermann, G. and Jost, U. (1998). An HMM-based probabilistic lexicon. *Proceedings of the ESCA workshop "Modeling Pronunciation Variation for Automatic Speech Recognition"*, Kerkrade, The Netherlands, 57-62.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C. and Raymond, W. (1998). Reduction of English function words in Switchboard. *Proceedings of ICSLP*, Sydney, Australia, 3111-3114.

Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y. and Zhang, S. (2001). What kind of pronunciation variation is hard for triphones to model? *Proceedings of ICASSP*, Salt Lake City, USA, 577-580.

Kessens, J. M., Strik, H. and Cucchiarini, C. (2000). A bottom-up method for obtaining information about pronunciation variation. *Proceedings of ICSLP*, Beijing, China, 274-277.

Kessens, J. M. and Wester, M. (1997). Improving recognition performance by modelling pronunciation variation. *Proceedings of the CLS opening Academic Year '97/'98*, Nijmegen, The Netherlands, 1-20.

Kessens, J. M., Wester, M., Cucchiarini, C. and Strik, H. (1997). Testing a method for modelling pronunciation variation. *Proceedings of the COST workshop "Speech Technology in the Telephone Network: Where are we today?"* Rhodes, Greece, 37-40.

Kessens, J. M., Wester, M. and Strik, H. (1999). Modeling within-word and cross-word pronunciation variation to improve the performance of a Dutch CSR. *Proceedings of ICPhS*, San Francisco, USA, 1665-1668.

Kipp, A., Wesenick, B. and Schiel, F. (1997). Pronunciation modeling applied to automatic segmentation of spontaneous speech. *Proceedings of Eurospeech*, Rhodes, Greece, 1023-1026.

Kirchhoff, K. (1999). *Robust Speech Recognition Using Articulatory Information*, Ph.D.thesis, University of Bielefeld, Germany.

Klabbers, E. (2000). *Segmental and Prosodic Improvements to Speech Generation*, Ph.D. thesis, Technical University of Eindhoven. Eindhoven, The Netherlands.

Lamel, L. and Adda, G. (1996). On designing pronunciation lexicons for large vocabulary, continuous speech recognition. *Proceedings of ICSLP*, Philadelphia, USA, 6-9.

Laver, J. (1994). *Principles of Phonetics*. Cambridge, Cambridge University Press.

Lee, K.-T. and Wellekens, C. J. (2001a). Dynamic lexicon using phonetic features. *Proceedings of Eurospeech*, Aalborg, Denmark, 1413-1416.

Lee, K.-T. and Wellekens, C. J. (2001b). Dynamic sharings of Gaussian densities using phonetic features. *Proceedings of ASRU*, Madonna di Campiglio, Italy.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication* **22**: 1-15.

Ma, K., Zavaliagkos, G. and Iyer, R. (1998). Pronunciation modeling for large vocabulary conversational speech recognition. *Proceedings of ICSLP*, Sydney, Australia, 2455-2458.

McAllaster, D., Gillick, L., Scattone, F. and Newman, M. (1998). Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. *Proceedings of ICSLP*, Sydney, Australia, 1847-1850.

Murray, I. R. and Arnott, J. L. (1993). Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustic Society of America* **93 (2)**: 1097-1108.

Peskin, B., Newman, M., McAllaster, D., Venkatesh, N., Hywel, R., Wegmann, S., Hunt, M. and Gillick, L. (2000). Improvements in recognition of conversational telephone speech. *Proceedings of ICASSP*, Phoenix, USA, 53-56.

Pols, L. C. W. (1999). Flexible, robust, and efficient human speech processing versus present-day speech technology. *Proceedings of ICPhS*, San Francisco, USA, 9-15.

Popovici, C. and Baggia, P. (1997). Specialized language models using dialogue predictions. *Proceedings of ICASSP*, München, Germany, 815-818.

Riley, M., Byrne, B., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraçlar, M., Wooters, C. and Zavaliagkos, G. (1999). Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication* **29**: 209-224.

Saraçlar, M. (2000). *Pronunciation Modeling for Conversational Speech Recognition*, Ph.D. thesis, John Hopkins University. Baltimore, Maryland.

Schiel, F., Kipp, A. and Tillmann, H. G. (1998). Statistical modelling of pronunciation: it's not the model, it's the data. *Proceedings of the ESCA workshop "Modeling Pronunciation Variation for Automatic Speech Recognition"*, Kerkrade, The Netherlands, 131-136.

Sloboda, T. (1995). Dictionary Learning: Performance through consistency. *Proceedings of ICASSP*, 453-456.

Strik, H. (2001). Pronunciation adaptation at the lexical level. *Proceedings of the ITRW Adaptation Methods for Speech Recognition*, Sophia-Antopolis, France, 123-130.

Strik, H. and Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication* **29**: 225-246.

Strik, H., Russel, A. J. M., van den Heuvel, H., Cucchiarini, C. and Boves, L. (1997). A spoken dialog system for the Dutch public transport information service. *International Journal of Speech Technology* **2**: 119-129.

Theune, M. (2000). *From Data to Speech: Language Generation in Context*, Ph.D.thesis, Technical Univerity of Eindhoven. Eindhoven, The Netherlands.

Torre, D., Villarrubia, L., Hernandez, L. and Elvira, J. M. (1997). Automatic alternative transcripition generation and vocabulary selection for flexible word recognizers. *Proceedings of ICASSP*, Munich, Germany, 1463-1466.

Van Noord, G., Bouma, G., Koeling, R. and Nederhof, M. (1999). Robust grammatical analysis for spoken dialogue systems. *Journal of Natural Language Engineering* **5 (1)**: 45-93.

Veldhuijzen van Zanten, G. (1998). Adaptive mixed-initiative dialogue modelling. *Proceedings of IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications*, Turin, Italy, 65-70.

Weintraub, M., Fosler-Lussier, E., Galles, C., Kao, Y.-H., Khudanpur, S., Saraçlar, M. and Wegmann, S. (1996). Automatic learning of word pronunciation from data. *1996 LVCSR Summer Research Workshop Technical Reports, chapter 3*, Center for Language and Speech Processing, John Hopkins University, USA.

Wessel, F. and Baader, A. (1999). Robust dialogue-state dependent language modeling using leaving-one-out. *Proceedings of ICASSP*, Phoenix, USA, 741-744.

Wester, M. (2001). Pronunciation modeling for ASR - knowledge-based and data derived methods. *submitted to Computer, Speech and Language*.

Wester, M. and Fosler-Lussier, E. (2000). A comparison of data-derived and knowlegde-based modeling of pronunciation variation. *Proceedings of ICSLP*, Beijing, China, 270-273.

Wester, M., Kessens, J. M. and Strik, H. (1998a). Improving the Performance of a Dutch CSR by Modeling Pronunciation Variation. *Proceedings of the ESCA workshop "Modeling Pronunciation Variation for Automatic Speech Recognition"*, Kerkrade, The Netherlands, 145-150.

Wester, M., Kessens, J. M. and Strik, H. (1998b). Selection of pronunciation variants in spontaneous speech: Comparing the performance of man and machine. *Proceedings of the ESCA workshop on the Sound Patterns of Spontaneous Speech: Production and Perception*, Aix-en-Provence, France, 157-160.

Wester, M., Kessens, J. M. and Strik, H. (1998c). Two automatic approaches for analyzing the frequency of connected speech processes in Dutch. *Proceedings of ICSLP*, Sydney, Australia, 3351-3356.

Wu, S.-L. (1998). *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*, Ph.D.thesis, University of California. Berkeley, USA.

Yang, Q. and Martens, J.-P. (2000). On the importance of exception and cross-word rules for the data-driven creation of lexica for ASR. *Proceedings of 11th ProRisc, Workshop*, Veldhoven, The Netherlands, 589-593.

Appendix A

Set of Dutch phones and other speech sounds for which HMMs are trained

Dutch phones			Dutch phones		
#	SAMPA[1]	Example	#	SAMPA[1]	Example
Vowels			Fricatives		
1	I	p**i**t	22	f	**f**el
2	E	p**e**t	23	v	**v**el
3	A	p**a**t	24	s	**s**ein
4	O	p**o**t	25	z	**z**ijn
5	Y	p**u**t	26	x	to**ch**
6	@	g**e**mak	27	h	**h**and
7	i	v**ie**r	28	S	**sh**ow
8	y	v**uu**r	Nasals, liquids, glides		
9	u	v**oe**r	29	m	**m**et
10	a:	n**aa**m	30	n	**n**et
11	e:	v**ee**r	31	N	ba**ng**
12	2:	d**eu**r	32[2]	l	**l**and
13	o:	v**oo**r	33[2]	L	ha**l**
14	Ei	f**ij**n	34[2]	r	**r**and
15	9y	h**ui**s	35[2]	R	to**r**
16	Au	g**ou**d	36	w	**w**it
Plosives			37	j	**j**a
17	p	**p**ak	**Other speech sounds**		
18	b	**b**ak	#	**symbol**	**description**
19	t	**t**ak	38	<n>	noise
20	d	**d**ak	39	<sil>	silence
21	k	**k**ap	40[3]	@=	filled pause

[1] See http://www.phon.ucl.ac.uk/home/sampa/dutch.htm
[2] In article 2, no distinction is made between post- and prevocalic /l/ and /r/
[3] Only used in article 2

The articles

Article 1

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

*Obtaining Phonetic Transcriptions: A Comparison between Expert Listeners and a Continuous Speech Recognizer**

Mirjam Wester, Judith M. Kessens, Catia Cucchiarini, and Helmer Strik

University of Nijmegen

Key words

automatic transcription

continuous speech recognition

pronunciation variation

Abstract

In this article, we address the issue of using a continuous speech recognition tool to obtain phonetic or phonological representations of speech. Two experiments were carried out in which the performance of a continuous speech recognizer (CSR) was compared to the performance of expert listeners in a task of judging whether a number of prespecified phones had been realized in an utterance. In the first experiment, nine expert listeners and the CSR carried out exactly the same task: deciding whether a segment was present or not in 467 cases. In the second experiment, we expanded on the first experiment by focusing on two phonological processes: schwa-deletion and schwa-insertion. The results of these experiments show that significant differences in performance were found between the CSR and the listeners, but also between individual listeners. Although some of these differences appeared to be statistically significant, their magnitude is such that they may very well be acceptable depending on what the transcriptions are needed for. In other words, although the CSR is not infallible, it makes it possible to explore large datasets, which might outweigh the errors introduced by the mistakes the CSR makes. For these reasons, we can conclude that the CSR can be used instead of a listener to carry out this type of task: deciding whether a phone is present or not.

* *Acknowledgments:* We kindly thank Prof. Dr. W.H. Vieregge for integrating our transcription material in his course curriculum. We are grateful to the various members of *A²RT* who gave their comments on previous versions of this article. We would like to thank Stephen Isard, Julia McGory, and Ann Syrdal for their useful comments on an earlier version of this article. The research by J.M. Kessens was carried out within the framework of the Priority Program Language and Speech Technology, sponsored by NWO (Dutch Organization for Scientific Research). The research by Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

Address for correspondence: Mirjam Wester, *A²RT*, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands;
e-mail: <M.Wester@let.kun.nl>

1 Introduction

In the last decade, an increasing number of databases have been recorded for the purpose of speech technology research (see for instance: <http://www.ldc.upenn.edu> and <http://www.icp.inpg.fr/ELRA/>). What started out as recordings of isolated words in restricted domains has now evolved to recordings of spontaneous speech in numerous domains. Since these databases contain a wealth of information concerning human language and speech, it seems that they should somehow be made available for linguistic research in addition to the speech technology research for which they were originally constructed and are currently being employed.

The use of such databases for linguistic research has at least two important advantages. First, many of them contain spontaneous speech. Most of the knowledge on speech production and perception is based on so-called "laboratory speech," while spontaneous speech is still under-researched (Cutler, 1998; Duez, 1998; Mehta & Cutler, 1988; Rischel, 1992; Swerts & Collier, 1992). Since it is questionable whether the findings concerning laboratory speech generalize to spontaneous speech, it seems that more emphasis should be placed on studying spontaneous speech. Second, these databases contain large amounts of speech material, which bodes well for the generalizability of the results of research that uses these databases as input.

Recent studies that have made use of such large databases of spontaneous speech reveal that this line of research is worth pursuing (Greenberg, 1999; Keating, 1997). On the basis of these observations one could get the impression that analysis of the speech data contained in such databases is within the reach of any linguist. Unfortunately, this is not true. The information stored in these databases is not always represented in a way that is most suitable for linguistic research. In general, before the speech material contained in the databases can be used for linguistic research it has to be phonetically transcribed (see, for instance, Greenberg, 1999). Phonetic transcriptions are obtained by analyzing an utterance auditorily into a sequence of speech units represented by phonetic symbols and making them is therefore extremely time-consuming. For this reason, linguists often decide not to have whole utterances transcribed, but only those parts of the utterance where the phenomenon under study is expected to take place (e.g., Kuijpers & van Donselaar, 1997). In this way, the amount of material to be transcribed can be limited in a way that is least detrimental for the investigation being carried out. Nevertheless, even with this restriction, making phonetic transcriptions remains a time-consuming, costly and often tedious task.

Another problem with manual phonetic transcriptions is that they tend to contain an element of subjectivity (Amorosa, von Benda, Wagner, & Keck, 1985; Laver, 1965; Oller & Eilers, 1975; Pye, Wilcox, & Siren, 1988; Shriberg & Lof, 1991; Ting, 1970; Witting, 1962). These studies reveal that transcriptions of the same utterance may show considerable differences, either when they are made by different transcribers (between-subjects variation) or when they are made by the same transcriber, but at different times or under different conditions (within-subjects variation). Since the presence of such discrepancies throws doubt on the reliability of phonetic transcription, it has become customary among researchers who use transcription data for their studies to have more than one person transcribe the speech material (e.g., Kuijpers & van Donselaar, 1997). This of course makes the task of transcribing speech even more time-consuming and costly.

To summarize, the problems connected with obtaining good manual phonetic transcriptions impose limitations on the amount of material that can be analyzed in linguistic research, with obvious consequences for the generalizability of the results. This suggests that if it were possible to obtain good phonetic transcriptions automatically, linguistic research would be made easier. Furthermore, in this way linguistic research could make profitable use of the large speech databases.

In speech technology, various tools have been developed that go some way toward obtaining phonetic representations of speech in an automatic manner. It is possible to obtain complete unrestricted phone-level transcriptions from scratch. However, phone accuracy turns out to vary between approximately 50% and 70%. For our continuous speech recognizer, we measured a phone accuracy level of 63% (Wester, Kessens, & Strik, 1998). In general, such levels of phone accuracy are too low for many applications. Therefore, to achieve acceptable recognition results, top-down constraints are usually applied.

The top-down constraints generally used in standard CSRs are a lexicon and a language model. With these constraints, word accuracy levels are obtained which are higher than the phone accuracy levels just mentioned. However, the transcriptions obtained with standard CSRs are not suitable for linguistic research because complete words are recognized, leading to transcriptions that are not detailed enough. The transcriptions thus obtained are simply the canonical transcriptions that are present in the lexicon. More often than not, the lexicon contains only one entry for each word thus always leading to the same transcription for a word regardless of pronunciation variation, whereas for linguistic research it is precisely this detail, a phone-level transcription, which is needed.

A way of obtaining a representation that approaches phonetic transcription is by using forced recognition, also known as forced (Viterbi) alignment. In forced recognition, the CSR is constrained by only allowing it to recognize the words present in the utterance being recognized. Therefore, in order to perform forced recognition, the orthographic transcription of the utterance is needed. The forced choice entails choosing between several pronunciation variants for each of the words present in the utterance. In this way, the variants that most closely resemble what was said in an utterance can be chosen. In other words, by choosing alternative variants that differ from each other in the representation of one specific segment, the CSR can be forced, as it were, to choose between different transcriptions of that specific segment thus leading to a transcription which is more detailed than a simple word-level transcription.

A problem of automatic transcription is the evaluation of the results. Given that there is no absolute truth of the matter as to what phones a person has produced, there is also no reference transcription that can be considered correct and with which the automatic transcription can be compared (Cucchiarini, 1993, pp. 11 – 13). To try and circumvent this problem as much as possible, different procedures have been devised to obtain reference transcriptions. One possibility consists in using a consensus transcription, which is a transcription made by several transcribers after they have agreed on each individual symbol (Shriberg, Kwiatkowski, & Hoffman, 1984). Another option is to have more than one transcriber transcribe the material and to use only that part of the material for which all transcribers agree or at least the majority of them (Kuijpers & van Donselaar, 1997).

The issues of automatic transcription and its evaluation have been addressed for example, by Kipp, Wesenick, and Schiel (1997) within the framework of the Munich

Automatic Segmentation System. The performance of MAUS has been evaluated by comparing the automatically obtained transcriptions with transcriptions made by three experts. The three manual transcriptions were not used to compose a reference transcription, but were compared pairwise with each other and with the automatic transcriptions to determine the degree of agreement. The results showed that the percentage agreement ranged from 78.8% to 82.6% for the three human transcribers, while agreement between MAUS and any of the human transcriptions ranged from 74.9% to 80.3% using data-driven rules, and from 72.5% to 77.2% using rules compiled by an experienced phonetician. These results indicate how the degree of agreement differs between expert transcribers and an automatic system, and, in a sense, this is a way of showing that the machine is just one of the transcribers. However, this is not sufficient because it does not say much about the quality of the transcriptions of the individual transcribers. Therefore, we propose the use of a reference transcription.

The aim of our research is to determine whether the automatic techniques that have been developed to obtain some sort of phonetic transcriptions for CSR can also be used meaningfully, in spite of their limitations, to obtain phonetic transcriptions for linguistic research. To answer this question, we started from an analysis of the common practice in many (socio/psycho) linguistic studies in which, as mentioned above, only specific parts of the speech material have to be transcribed. In addition, we further restricted the scope of our study by limiting it to insertion and deletion phenomena, which is to say that we did not investigate substitutions. The rationale behind this choice is that it should be easier for a CSR to determine whether a segment is present or not than to determine which one of several variants of a given segment has been realized. If the technique presented here turns out to work for deletions and insertions it could then be extended to other processes. In other words, our starting point was a clear awareness of the limitations of current CSR systems, and an appreciation of the potentials that CSR techniques, despite their present limitations, could have for linguistic research.

In this study, we describe two experiments in which different comparisons are carried out between the automatically obtained transcriptions and the transcriptions made by human transcribers. In these experiments the two most common approaches to obtaining a reference transcription are used: the majority vote procedure and the consensus transcription.

In the first experiment, four kinds of comparisons are carried out to study how the machine's performance relates to that of nine listeners. First of all the degree of agreement in machine-listener pairs is compared to the degree of agreement in listener-listener pairs, as in the Kipp et al. (1997) study. Second, in order to be able to say more about the quality of the machine's transcriptions and the transcriptions by the nine listeners, they are all compared to a reference transcription (majority vote procedure). Third, because it can be expected that not all processes give the same results, the comparisons with the reference transcription are carried out for each individual process of deletion and insertion. Fourth, a more detailed comparison of the choices made by the machine and by the listeners is carried out to get a better understanding of the differences between the machine's performance and that of the listeners.

The results of this last comparison show that the CSR systematically tends to choose for deletion (non-insertion) of phones more often than listeners do. To analyze this to a further

extent, we carried out a second experiment in order to find out why and in what way the detection of a phone is different for the CSR and for the listeners. In order to study this, a more detailed reference transcription was needed. Therefore, we used a consensus transcription instead of a majority vote procedure to obtain a reference transcription.

The organization of this article is as follows: First, the methodology of the first experiment is explained followed by the presentation of the results. Before going on to the second experiment a discussion of the results of Experiment 1 is given. Following on from this, the methodology of the second experiment is explained, subsequently the results are shown and also discussed. Finally, conclusions are drawn as to the merits and usability of our automatic transcription tool.

2 Experiment 1

2.1 *Method and Material*

2.1.1 *Phonological variation*

The processes we chose to study concern insertions and deletions of phones within words (i.e., alterations in the number of segments). Five phonological processes were selected for investigation: /n/-deletion, /r/-deletion, /t/-deletion, schwa-deletion and schwa-insertion. The main reasons for selecting these five phonological processes are that they occur frequently in Dutch and are well described in the linguistic literature. Furthermore, these phonological processes typically occur in fast or extemporaneous speech, but to a lesser extent in careful speech; therefore it is to be expected that they will occur in our speech material (for more details on the speech material, see the following section).

The following description of the four processes: /n/-deletion, /t/-deletion, schwa-deletion and schwa-insertion is according to Booij (1995), and the description of the /r/-deletion process is according to Cucchiarini and van den Heuvel (1999). The descriptions given here are not exhaustive, but describe the conditions of rule application which we formulated to generate the variants of the phonological processes.

1. /n/-deletion:

In standard Dutch, syllable-final /n/ can be dropped after a schwa, except if that syllable is a verbal stem or if it is the indefinite article *een* [ən] ‘a’. For many speakers, in particular in the western part of the Netherlands, the deletion of /n/ is obligatory.

Example: *reizen* [rɛizən] → [rɛizə] ‘to travel’

2. /r/-deletion:

According to Cucchiarini and van den Heuvel (1999), /r/-deletion can take place in Dutch when /r/ is preceded by a vowel and followed by a consonant in a word. Although this phenomenon is attested in various contexts, it appears to be significantly more frequent when the vowel preceding the /r/ is a schwa.

Example: *Amsterdam* [ɑmstərdɑm] → [ɑmstədɑm] ‘Amsterdam’

3. /t/-deletion:

If a /t/ in a coda is preceded by an obstruent, and followed by another consonant, the /t/ may be deleted.

Example: *rechtstreeks* [rɛxtstreks] → [rɛxstreks] 'directly'

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

Example: *'s avonds* [savɔnts] → [savɔns] 'in the evening'

Finally, we also included /t/-deletion in word-final position following an obstruent.

Example: *Utrecht* [ytrɛxt] → [ytrɛx] 'Utrecht'

4. schwa-deletion:

When a Dutch word has two consecutive syllables headed by a schwa, the first schwa may be deleted, provided that the resulting onset consonant cluster consists of an obstruent followed by a liquid.

Example: *latere* [latərə] → [latrə] 'later'

5. schwa-insertion:

In nonhomorganic consonant clusters in coda position schwa may be inserted. Schwa-insertion is not possible 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.

Example: *Delft* [dɛlft] → [dɛləft] 'Delft'

2.1.2
Selection of speech material

The speech material used in the experiments was selected from a Dutch database called VIOS, which contains a large number of telephone calls recorded with the on-line version of a spoken dialog system called OVIS (Strik, Russel, Van Den Heuvel, Cucchiarini, & Boves, 1997). OVIS is employed to automate part of an existing Dutch public transport information service. The speech material consists of interactions between man and machine, and can be described as extemporaneous speech.

The phonological rules described in the previous section were used to automatically generate pronunciation variants for the words being studied. In some cases, it was possible to apply more than one rule to the same word. However, in order to keep the task relatively easy for the listeners we decided to limit to two the number of rules which could apply to a single word.

From the VIOS corpus, 186 utterances were selected. These utterances contain 379 words with relevant contexts for one or two rules to apply. For 88 words, the conditions for rule application were met for two rules simultaneously and thus four pronunciation variants were generated. For the other 291 words, only one condition of rule application was relevant and two variants were generated. Consequently, the total number of instances in which a rule could be applied is 467. Table 1 shows the number of items for each of the different rules and the percentages of the total number of items. This distribution (columns 2 and 3) is not uniform, because the distribution in the VIOS corpus (columns 4 and 5) is

TABLE 1

Number of items selected per process for Experiment 1, and the percentage of the total number of items in Experiment 1. Number of items and their corresponding percentages in the VIOS corpus, for each process

phonological process	*# Exp. 1*	*% Exp. 1*	*# VIOS corpus*	*% VIOS corpus*
/n/-deletion	155	33.2	10,694	45.2
/r/-deletion	127	27.2	7,145	30.2
/t/-deletion	84	18.0	3,665	15.5
schwa-deletion	53	11.3	275	1.2
schwa-insertion	48	10.3	1,871	7.9

not uniform. However, we tried to ensure a more even distribution by having at least a 10% representation for each phonological process in the material which was selected for Experiment 1.

2.1.3 *Experimental procedure*

Nine expert listeners and the continuous speech recognizer (CSR) carried out the same task, that is, deciding for the 379 words which pronunciation variant best matched the word that had been realized in the spoken utterances (forced choice).

Listeners. The nine expert listeners are all linguists who were selected to participate in this experiment because they have all carried out similar tasks for their own investigations. For this reason, they are representative of the kind of people that make phonetic transcriptions and who may benefit from automatic ways of obtaining such transcriptions. The 186 utterances were presented to them over headphones, in three sessions, with the possibility of a short break between successive sessions. The orthographic representation of the whole utterance was shown on screen, see Figure 1. The words which had to be judged were indicated by an asterisk. Beneath the utterance, the phonemic transcriptions of the pronunciation variants were shown. The listeners' task was to indicate for each word which of the phonemic transcriptions presented best corresponded to the spoken word. The listener could listen to an utterance as often as he/she felt was necessary in order to judge which pronunciation variant had been realized.

CSR. The utterances presented to the listeners were also used as input to the CSR which is part of the spoken dialog system OVIS (Strik et al., 1997). The orthography of the utterances was available to the CSR. The main components of the CSR are a lexicon, a language model, and acoustic models.

For the automatic transcription task, the CSR was used in forced recognition mode. In this type of recognition, the CSR is "forced" to choose between different pronunciations of a word instead of between different words. Hence, a lexicon with more than one possible pronunciation per word was needed. This lexicon was made by generating pronunciation

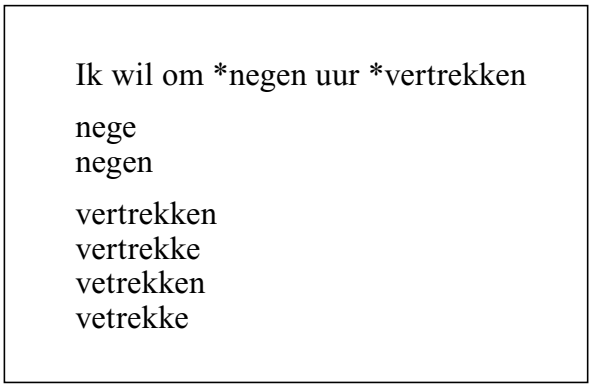

Figure 1

Pronunciation variant selection by the nine expert listeners. The left-hand panel shows an example of the manner in which the utterances were visually presented to the listeners. The right-hand panel shows the translation

variants for the words in the lexicon using the five phonological rules described earlier. Pronunciation variants were only generated for the 379 words under investigation, for the other words present in the 186 utterances the canonical transcription was sufficient. The canonical phone transcription is the phone transcription generated with the Text-to-Speech system developed at the University of Nijmegen (Kerkhoff & Rietveld, 1994). The language model (unigram and bigram) was restricted in that it only contained the words present in the utterance which was being recognized.

Feature extraction was done every 10 ms for frames with a width of 16 ms. The first step in feature analysis was an FFT analysis to calculate the spectrum. Next, the energy in 14 mel-scaled filter bands between 350 and 3400 Hz was calculated. The next processing stage was the application of a discrete cosine transformation on the log filterband coefficients. Besides 14 cepstral coefficients (c_0–c_{13}), 14 delta coefficients were also used. Thus, a total of 28 feature coefficients were used.

The acoustic models which we used are monophone hidden Markov models (HMM). The topology of the HMMs is as follows: Each HMM is made up of six states, and consists of three parts. Each of the parts has two identical states, one of which can be skipped (Steinbiss et al., 1993). In total, 40 HMMs were trained. For 33 of the phonemes, one context-independent HMM was used. For the /l/ and the /r/, separate models were trained depending on their position in the syllable, that is, different models were trained for prevocalic and postvocalic position. In addition to these 37 acoustic models, three other models were trained: an HMM for filled pauses, one for nonspeech sounds and a one-state HMM to model silence. Furthermore, the acoustic models which were used for the automatic transcription task were "retrained" models. Retrained acoustic models, in our case, are HMMs which are trained on a training corpus in which pronunciation variation has been transcribed. This is accomplished by performing forced recognition of the training corpus using a lexicon which contains pronunciation variants, thus adding variants to the training corpus at the appropriate places. Subsequently, the resulting corpus is then used to retrain the HMMs. The main reason for using retrained acoustic models is that we expect these

models to be more precise and therefore better suited to the task. For more details on this procedure see Kessens, Wester, and Strik (1999).

Note that we use monophone models rather than diphone or triphone models although in state-of-the-art recognition systems diphone and triphone models have proven to outperform monophone models. This is the case in a recognition task, but not necessarily in forced recognition.

2.1.4 *Evaluation*

Binary scores. On the basis of the judgments made by the listeners and the CSR, scores were assigned to each item. For each of the rules two categories were defined: (1) "rule applied" and (0) "rule not applied." For 88 words four variants were present, as mentioned earlier. For each of these words two binary scores were obtained, that is, for each of the two underlying rules it was determined whether the rule was applied (1) or not (0). For each of the remaining 291 words one binary score was obtained. Thus, 467 binary scores were obtained for each of the listeners and for the CSR.

Agreement. We used Cohen's kappa (Cohen, 1968) to calculate the degree of agreement between listeners and the CSR. The reason we chose to use Cohen's κ instead of for instance percentage agreement is that the distributions of the binary scores may differ for the various phonological processes, and in that case, it is necessary to correct for chance agreement in order to be able to compare the processes to each other. Cohen's κ is a measure which corrects for chance:

$$\kappa = \frac{(P_o - P_c)}{(1 - P_c)} \qquad -1 \leq \kappa \leq 1$$

where: P_o = observed proportion of agreement; P_c = proportion of agreement on the basis of chance

Table 2 shows the qualifications for κ-values greater than zero, to indicate how the κ-values should be interpreted (taken from Landis & Koch, 1977).

TABLE 2

Qualifications for κ-values >0

k-value	*qualification*
0.00 – 0.20	slight
0.21 – 0.40	fair
0.41 – 0.60	moderate
0.61 – 0.80	substantial
0.81 – 1.00	almost perfect

Reference transcriptions. In the introduction, we mentioned various strategies that can be used to obtain a reference transcription. In this first experiment, we used the majority vote procedure. Two types of reference transcriptions were composed using the majority vote

procedure: 1) reference transcriptions based on eight listeners, and 2) a reference transcription based on all nine listeners.

The reference transcriptions based on eight listeners were used to compare the performance of each individual listener to the performance of the CSR. For each listener, the reference transcription was based on the other eight listeners. By using a reference transcription based on eight listeners, it is possible to compare the CSR and an individual listener to exactly the same reference transcription, thus ensuring a fair and correct comparison. If, instead, one were to use a reference transcription based on all nine listeners, the comparison would not be as fair because, in effect, the listener would be compared to herself/himself due to the fact that the results of that individual listener would be included in the reference transcription.

Consequently, nine sets of reference transcriptions were compiled each with four different degrees of strictness. The different degrees of strictness which we used were A: a majority of at least five out of eight listeners agreeing, B: six out of eight, C: seven out of eight, and finally D: only those cases in which all eight listeners agree. Subsequently, the degree of agreement for an individual listener with the reference transcription was calculated and the same was done for the CSR with the various sets of reference transcriptions.

The reference transcription based on nine listeners was used to analyze the differences between the listeners and the CSR. In this case, it is also possible to use different degrees of strictness. However, for the sake of brevity, we only show the results for a majority of five out of nine listeners agreeing. The reason for choosing five out of nine is that as the reference becomes stricter, the number of items in it reduces, whereas, for this degree of strictness all items (467) are present.

2.2 *Results*

Analysis of the results was done by carrying out four comparisons. First, pairwise agreement was calculated for the various listeners and for the listeners and the CSR. Pairwise agreement gives an indication of how well the results of the listeners compare to each other and to the results of the CSR. However, as we explained in the introduction, pairwise agreement is not the most optimal type of comparison, as the transcriptions of individual transcribers may be incorrect. To circumvent this problem as much as possible, we used the majority vote procedure to obtain reference transcriptions. Thus, we also calculated the degree of agreement between the individual listeners and a reference transcription based on the other eight listeners and between the CSR and the same sets of reference transcriptions. These results give a further indication of how well the listeners and the CSR compare to each other, but we were also curious whether the same pattern exists for the various phonological processes. Therefore, for the third comparison, the data were split up for the separate processes and the degree of agreement between the CSR and the reference transcriptions was calculated for each of the phonological processes. These data showed that there are indeed differences between the various phonological processes. In an attempt to understand the differences, we analyzed the discrepancies between the CSR and the listeners. In this final analysis, the reference transcription based on a majority of five out of nine listeners agreeing was employed.

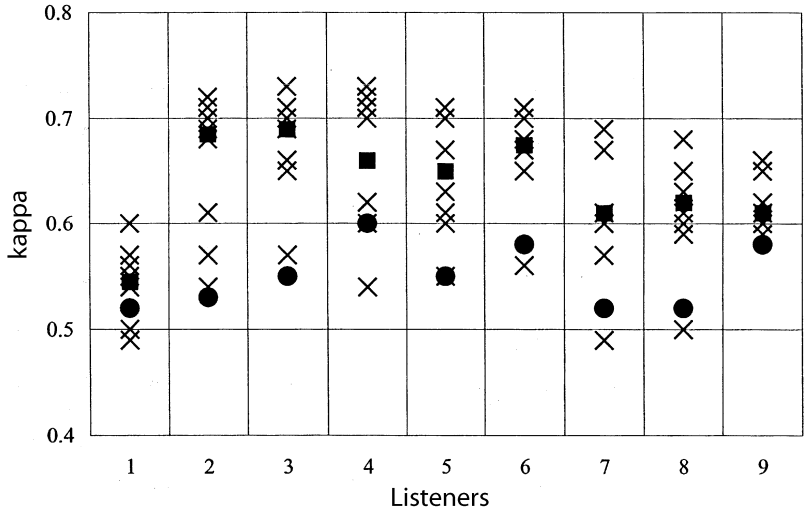

Figure 2
Cohen's κ for the agreement between the CSR and each listener (●), for listener pairs (×) and the median of the listeners (■)

2.2.1
Pairwise agreement between CSR and listeners

For each listener, pairwise agreement was calculated for each pair of listeners and for each CSR-listener pair. In this analysis, no reference transcription was used. Figure 2 shows the results of the pairwise comparisons. For instance, in the first "column" in Figure 2, the crosses (×) indicate the comparison between listener 1 and each of the other listeners, the square (■) shows the median for all listener pairs, and the circle (●) indicates the degree of agreement between the CSR and listener 1.

The results for pairwise agreement in Figure 2 show that there is quite some variation among the different listener pairs. The κ-values vary between 0.49 and 0.73, and the median for all listener pairs is 0.63. The median κ-value for all nine listener-CSR pairs is 0.55. In Figure 2, it can also be seen that the degree of agreement between each of the listeners and the CSR is lower than the median κ-value for the listeners. Statistical tests (Mann-Whitney test, $p < .05$) show that the CSR and listeners 1, 3, and 6 behave significantly different from the other listeners. For both the CSR and listener 1, agreement is significantly lower than for the rest of the listeners whereas for listeners 3 and 6 agreement is significantly higher.

2.2.2
Agreement with reference transcriptions with varying degrees of strictness

In order to further compare the CSR's performance to the listeners', nine sets of reference transcriptions were compiled, each based on eight listeners and with four different degrees of strictness. With an increasingly stricter reference transcription, the differences between listeners are gradually eliminated from the set of judgments under investigation. It is to be expected that if we compare the performance of the CSR with the reference transcriptions of type A, B, C, and D, the degree of agreement between the CSR and the reference transcription will increase when going from A to D. The rationale behind this is that those cases for which a greater number of listeners agree should be easier to judge for the listeners. Therefore, it can be expected that those cases should be easier for the CSR too. In going from A to D the number of cases involved is reduced (see Appendix 1 for details on numbers).

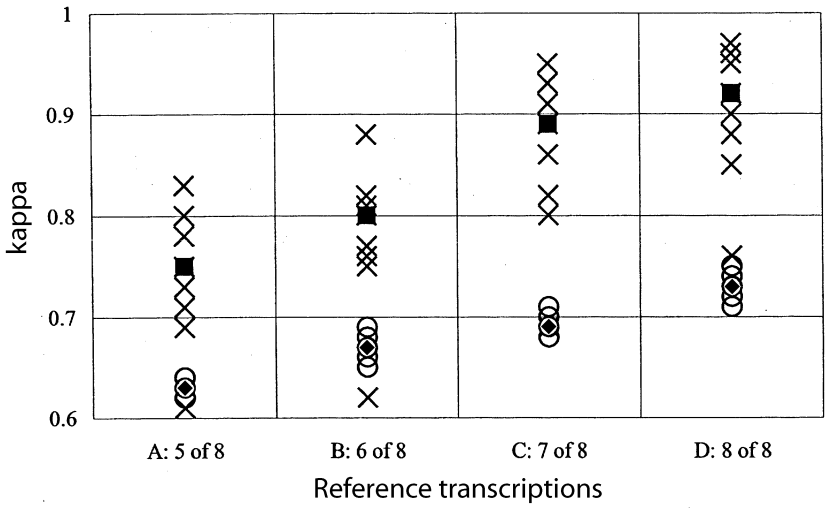

Figure 3

Cohen's κ for CSR (○) and listeners (×) compared to various se of reference transcripti based on responses of eight listeners, and median κ for the sets o reference transcriptions for the CSR (◆) and th listeners (■)

Figure 3 shows the κ-values obtained by comparing each of the listener's transcriptions to the relevant set of reference transcriptions (×) and the median for all listeners (■). In addition, the κ-values obtained by comparing the CSR's transcriptions to each of the sets of reference transcriptions (○), and the median for all the CSR's κ-values (◆) are shown. It can be seen that in most cases the degree of agreement between the different sets of reference transcriptions and the listeners is higher than the degree of agreement between the reference transcriptions and the CSR. These differences between the CSR and the listeners are significant. (Wilcoxon signed ranks test, $p < .05$.) However, as we expected, the degree of agreement between the reference transcription and both the listeners and the CSR gradually increases, as the reference transcription becomes stricter.

2.2.3
Agreement with reference transcription for the separate phonological processes

In the previous section, we compared results in which items of the various phonological processes were pooled. However, it is possible that the CSR and the nine listeners perform differently on different phonological processes. Therefore, we also calculated the results for the five phonological processes separately, once again using a majority vote based on eight listeners (see Appendix 2 for the number of items in each set of reference transcriptions). The results are shown in Figure 4. For each process, the degree of agreement between each of the sets of reference transcriptions and the nine listeners (×) and the CSR (○) is shown, first for all of the processes together and then for the individual processes. The median for the nine listeners (■) and the median for the results of the CSR (◆) are also shown. Furthermore, for three of the listeners, the data points have been joined to give an indication of how an individual listener performs on the different processes in relation to the other listeners.

For instance, if we look at the data points for listener A (dotted line) we see that this listener reaches the highest κ-values for all processes except for /n/-deletion in which case the listener is bottom of the group of listeners. The data points for listener B (solid line) fall in the middle of the group of listeners, except for the processes of /r/-deletion and /t/-deletion, where this listener is bottom of the group. The data points for listener C (dashed line) show a poor performance on schwa-insertion and schwa-deletion compared to the

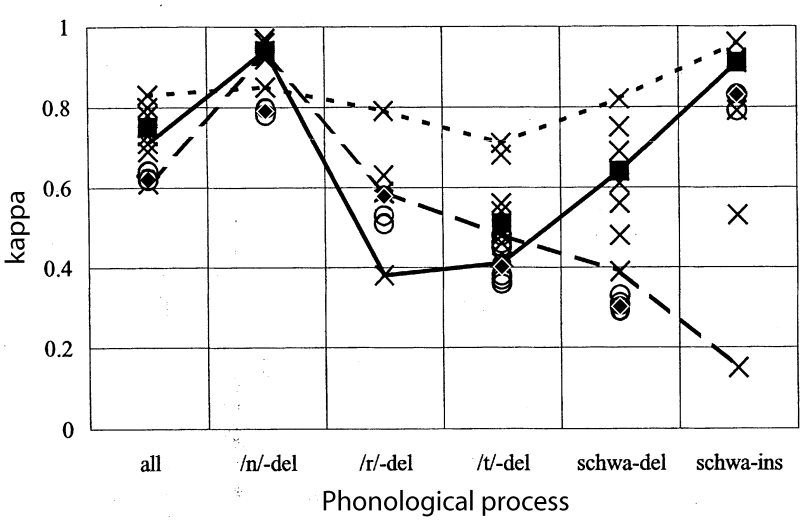

Figure 4

Cohen's κ for the listeners and the CSR compared to the sets of reference transcriptions (5 of 8) for the various phonological processes (○ = CSR, × = listener, ■ = median listeners, ◆ = median CSR, dotted line = listener A, solid line = listener B, and dashed line = listener C)

rest of the listeners, but a more or less average performance on the other processes. These three examples indicate that none of the listeners is consistently better or worse than the others in judging the various phonological processes. Furthermore, on the basis of the medians for the listeners, we can conclude that /n/-deletion and schwa-insertion are the easiest processes to judge, whereas the processes of /r/-deletion, /t/-deletion and schwa-deletion are more difficult processes for listeners to judge. This is also the case for the CSR.

As far as the difference between the CSR and the listeners is concerned, statistical analysis (Wilcoxon signed ranks test, $p < .05$) shows that for the phonological processes of /r/-deletion and schwa-insertion there is no significant difference between the CSR and the listeners. For the other three processes the difference is significant, and this is also the case for all of the phonological processes grouped together. This is also reflected in Figure 4, as there is almost no difference in the median for the CSR and the listeners for /r/-deletion (0.01) and for schwa-insertion (0.08). For /n/-deletion (0.15) and /t/-deletion (0.11), the difference is larger, and comparable to the results found for all rules pooled together (0.12), leaving the main difference in the performance of the listeners and the CSR to be found for schwa-deletion (0.34).

2.2.4
Differences between CSR and listeners

The results in the previous section give rise to the question of why the results are different for various phonological processes and what causes the differences in results between the listeners and the CSR. In this section, we try to answer the question of what causes the discrepancy, by looking more carefully at the differences in transcriptions found for the listeners and the CSR. In these analyses, we used the reference transcription based on a majority of five out of nine listeners agreeing. The reason we use five of nine instead of five of eight is because we wanted to include all of the material used in the experiment in this analysis. Furthermore, instead of using the categorization "rule applied" and "rule not applied" the categories "phone present" and "phone not present" are used to facilitate presentation and interpretation of the data. Each item was categorized according to whether agreement was found between the CSR and the reference transcription or not.

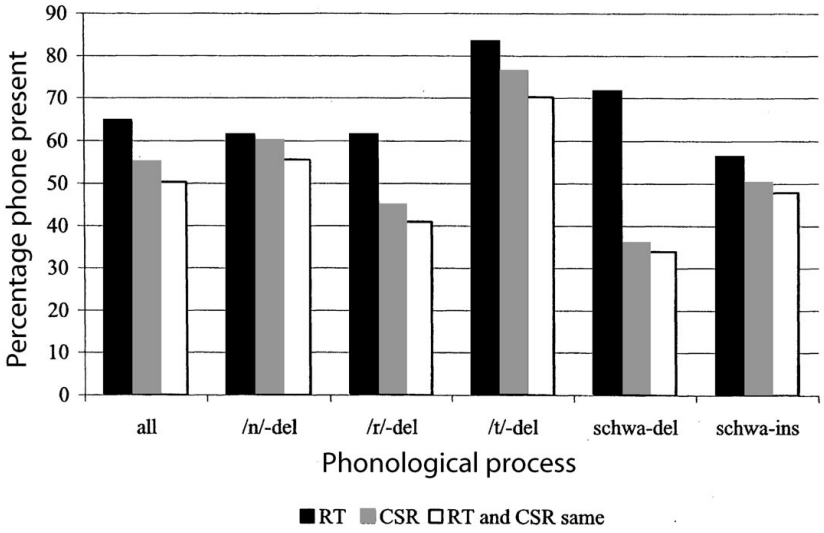

Figure 5

Percentages of phone present for the reference transcription (RT), the CSR, and the CSR and RT together, for the various phonological processes

Figure 5 shows the percentages of phone present according to the reference transcription (RT, dark gray bar) and the CSR (gray bar). It also shows the percentages of phone present for which the RT and CSR agree (white bar). For exact counts and further details, see Appendix 3. It can be seen in Figure 5 that, for all phonological processes pooled, the phones in question are realized in 65% of all cases according to the reference transcription and in 55% of the cases according to the CSR. In fact for every process the same trend can be seen: The RT bar is always higher than the CSR bar. Furthermore, the CSR bar is never much higher than the RT-CSR bar, which indicates that the CSR rarely chooses phone present when the RT chooses phone not present. The differences between the CSR and the listeners are significant for /r/-deletion, for schwa-deletion and for all rules pooled (Wilcoxon signed ranks test, $p < .05$).

An explanation for the differences between the CSR and the listeners may be that they have different durational thresholds for detecting a phone, in the sense that phones with a duration that falls under a certain threshold are less likely to be detected. This sounds plausible if we consider the topology of the HMMs. The HMMs we use have at least three states, thus phones which last less than 30 ms are less likely to be detected. (Feature extraction is done every 10 ms.)

To investigate whether this explanation is correct, we analyzed the data for schwa-deletion and /r/-deletion in terms of the duration of the phones. The speech material was automatically segmented to obtain the durations of the phones. The segmentation was carried out using a transcription that did not contain deletions to ensure that durations could be measured for each phone. Due to the typology of the HMMs durations shorter than 30 ms are also classified as 30 ms As a result, the 30 ms category may contain phones that are shorter in length.

Figures 6 and 7 show the results for schwa-deletion and /r/-deletion, respectively. These figures show that the longer the phone is the less likely that the CSR and the listeners consider it deleted, and the higher the degree of agreement between the CSR and the listeners is. Furthermore, the results for schwa-deletion seem to indicate that the listeners and the CSR do indeed have a different threshold for detecting a phone. Figure 6 shows that the listeners perceive more than 50% of the schwas that are 30 ms or less long, whereas

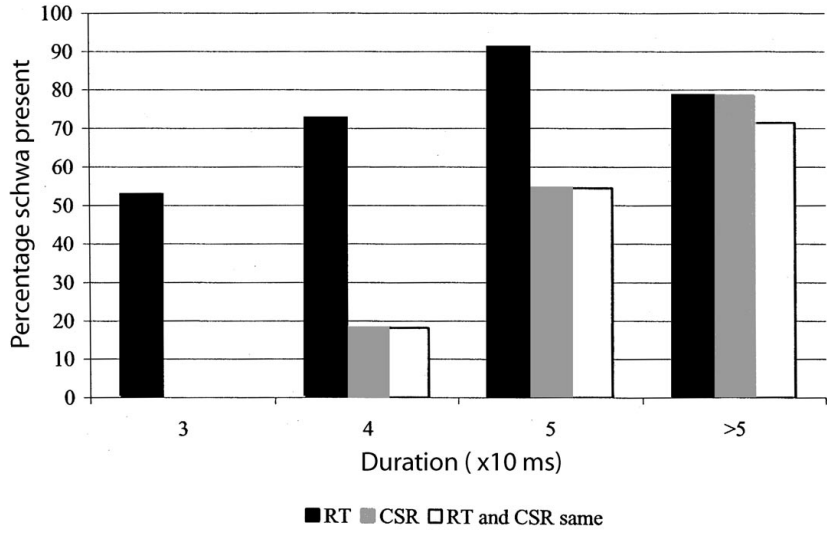

Figure 6

Percentage schwas present, as a function of the duration of the phones, according to the reference transcription (RT), the CSR, and the CSR and RT together

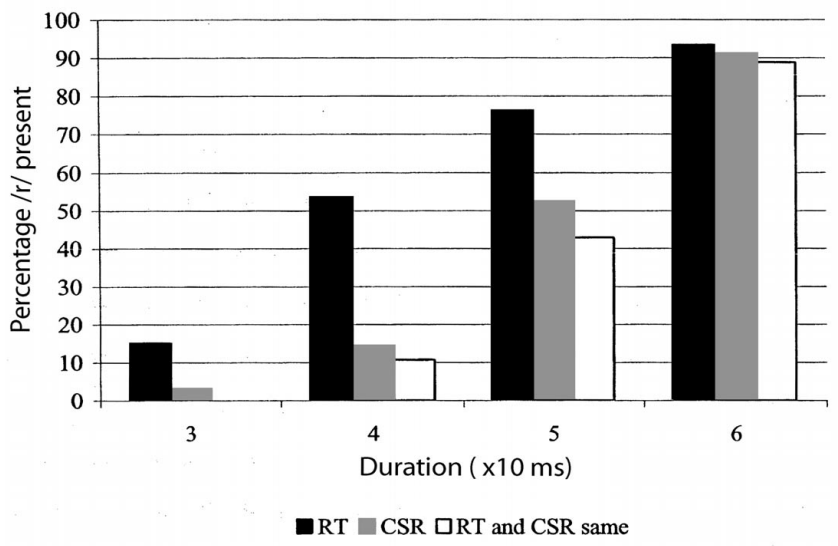

Figure 7

Percentage/r/s present, as a function of the duration of the phones, according to the reference transcription (RT), the CSR, and the CSR and RT together

the CSR does not detect any of them. However, for /r/-deletion this is not quite the case as neither the CSR nor the listeners detect most of the /r/s with a duration of 30 ms or less.

2.3 *Discussion*

The results concerning pairwise agreement between the listeners and the CSR show that the agreement values obtained for the machine differ significantly from the agreement values obtained for the listeners. However, the results of three of the listeners also differ significantly from the rest. Thus, leaving a middle group of six listeners that do not significantly differ from each other. On the basis of these pairwise agreement results, we must conclude that the CSR does not perform the same as the listeners, and what is more that not all of the listeners perform the same either.

A significant difference between the machine's performance and the listeners' performance also appeared when both the CSR transcription and those of the nine listeners were

compared with reference transcriptions of various degrees of strictness. However, the cases that were apparently easier to judge for the listeners, that is, a greater number of them agreed, also presented fewer difficulties for the CSR.

The degrees of agreement observed in this experiment, both between listeners and between listeners and machine, are relatively high. This is all the more so if we consider that the degree of agreement was not calculated over all speech material, as in the Kipp et al. (1997) study, but only for specific cases which are considered to be among the most difficult ones. As a matter of fact, all processes investigated in these experiments are typical connected speech processes that in general have a gradual nature and are therefore difficult to describe in categorical terms (Booij, 1995; Kerswill & Wright, 1990).

In addition, more detailed analyses of the degree of agreement between humans and machine for the various processes revealed that among the phenomena investigated in these experiments there are differences in degree of difficulty. Also in this case the machine's performance turned out to be similar to the listeners', in the sense that the processes that presented more difficulties for the listeners also appeared to be more difficult for the machine. Statistical analyses were carried out for the various phonological processes. The results of these tests are shown in Table 3.

TABLE 3

Results of the statistical analyses for the individual phonological processes from Figure 4 and Figure 5. S = significant; N = not significant difference

Figure	*/n/-deletion*	*/r/-deletion*	*/t/-deletion*	*schwa-deletion*	*schwa-insertion*
4	S	N	S	S	N
5	N	S	N	S	N

Table 3 shows that the comparisons carried out for the individual processes do not present a very clear picture. For schwa-deletion the differences are always significant and for schwa-insertion they are always not significant. For the remaining three processes, the results of the statistical analyses seem to contradict each other. This is maybe less puzzling than it seems if we consider that the comparisons that were made are of a totally different nature. In Figure 4, nine pairs of kappas were compared to each other and in Figure 5, many pairs of "rule applied" and "rule not applied" were compared (the number varies per rule). Still the question remains how we are to interpret these results. The objective was to find out whether the CSR differs significantly from the listeners or not. If we look at the global picture of all rules pooled together then we must conclude that this is indeed the case; the CSR differs significantly from the listeners. However, if we consider the individual processes, we find that the differences for schwa-deletion are significant, for schwa-insertion they are not and that for the other three processes no definite conclusion can be drawn, as it depends on the type of analysis. In other words, only in the case of schwa-deletion are the results of the CSR significantly different from the results of the listeners.

The fact that the degree of agreement between the various listeners and the reference transcriptions turned out to be so variable depending on the process investigated deserves attention, because, in general, the capabilities of transcribers are evaluated in terms of

global measures of performance calculated across all kinds of speech processes, and not as a function of the process under investigation (Shriberg, Kwiatowski, & Hoffman, 1984). However, this experiment has shown that the differences in degree of agreement between the various processes can be substantial.

These results could be related to those presented by Eisen, Tillman, and Draxler (1992) about the variability of interrater and intrarater agreement as a function of the sounds transcribed, although there are some differences in methodology between our experiment and theirs. First, Eisen et al. (1992) did not analyze whether a given segment had been deleted/inserted or not, but whether the same phonetic symbol had been used by different subjects or by the same subject at different times. The degree of agreement in this latter case is directly influenced by the number of possible alternatives, which may be different for the various sounds. In our experiment, on the other hand, this number is constant over all cases. Furthermore, the relative difficulty in determining which particular type of nasal consonant has been realized may be different from the difficulty in determining whether a given nasal consonant is present or not. Second, these authors expressed the degree of agreement using percentage agreement, which, as explained above, does not take chance agreement into account, and therefore makes comparisons rather spurious. In general, however, Eisen et al. (1992) found that consonants were more consistently transcribed than vowels. In our experiment, there is no clear indication that this is the case. Within the class of consonants, Eisen et al. (1992) found that laterals and nasals were more consistently transcribed than fricatives and plosives, which is in line with our findings that higher degrees of agreement were found for /n/-deletion than for /t/-deletion. For liquids no comparison can be made because these were not included in the Eisen et al. (1992) study. As to the vowels, Eisen et al. (1992) found that central vowels were more difficult to transcribe. In our study we cannot make comparisons between different vowel types because only central vowels were involved. In any case, this provides further evidence for the fact that the processes studied in our experiments are among those considered to be more difficult to analyze.

Another important observation to be made on the basis of the results of this experiment is that apparently it is not only the sound in question that counts, be it an /n/ or a schwa, but rather the process being investigated. This is borne out by the fact that the results are so different for schwa-deletion as opposed to schwa-insertion. This point deserves further investigation.

The fourth comparison carried out in Experiment 1 was aimed at obtaining more insight into the differences between the machine's choices and the listeners' choices. These analyses revealed that these differences were systematic and not randomly distributed over presence or absence of the phone in question. Across-the-board the listeners registered more instances of insertion and fewer instances of deletion than the machine did, thus showing a stronger tendency to perceive the presence of a phone than the machine. Although this finding was consistent over the various processes, it was most pronounced for schwa-deletion.

In view of these results, we investigated whether the CSR and the listeners possibly have different durational thresholds in detecting the presence of a phone. This analysis showed that it is clear that duration does certainly play a role, but there is no unambiguous threshold which holds for all phones.

Another possible explanation for these results could be the very nature of the HMMs. These models do not take much account of neighboring sounds. This is certainly true in our case as we used context independent phones, but even when context dependent phone models are used this is still the case. With respect to human perception, on the other hand, we know that the way one sound is perceived very much depends on the identity of the adjacent sounds and the transitions between the sounds. If the presence of a given phone is signaled by cues that are contained in adjacent sounds, the phone in question is perceived as being present by human listeners, but would probably be absent for the machine that does not make use of such cues. A third possible explanation for the discrepancies between the machine response and the listeners' responses lies in the fact that listeners can be influenced by a variety of factors (Cucchiarini, 1993, p. 55), among which spelling and phonotactics are particularly relevant to our study. Since in our experiments the subjects listened to whole utterances, they knew which words the speaker was uttering and this might have induced them to actually "hear" an /r/, a /t/, an /n/ or a schwa when in fact they were not there. In other words, the choice for a nondeletion could indeed be motivated by the fact that the listener knew which phones were supposed to be present rather than by what was actually realized by the speaker. This kind of influence is known to be present even in experienced listeners like those in our experiments. A problem with this argument is that while it can explain the lower percentages of deletion by the humans, it does not explain the higher percentages of insertions. A further complicating factor in our case is that the listeners are linguists and may therefore be influenced by their knowledge and expectations about the processes under investigation. Finally, schwa-insertion happens to be a phenomenon that is more common than schwa-deletion (Kuijpers & Van Donselaar, 1997) which could explain part of the discrepancy found for the two processes.

3 Experiment 2

In Experiment 1, analysis of the separate processes showed that both for listeners and the CSR some processes are more easily agreed on than others. Closer inspection of the differences showed that the CSR systematically tends to choose for deletion (non-insertion) of phones more often than listeners do. This finding was consistent over the various processes and most pronounced for schwa-deletion. Furthermore, we found that the results were quite different for schwa-deletion as opposed to schwa-insertion. To investigate the processes concerning schwa to a further extent, a second experiment was carried out in which we focused on schwa-deletion and schwa-insertion. The first question we would like to see answered pertains to the detectability of schwa: is the difference between listeners and machine truly of a durational nature? In order to try to answer this question, it was necessary to make use of a more detailed transcription in which it was possible for transcribers to indicate durational aspects and other characteristics of schwa more precisely. To achieve this, we used the method of consensus transcriptions to obtain reference transcriptions of the speech material.

The second question is why the processes of schwa-deletion and schwa-insertion lead to such different results. In Experiment 1, the machine achieved almost perfect agreement with listeners on judging the presence of schwa in the case of schwa-insertion, whereas only fair agreement was achieved in the case of schwa-deletion. This difference is quite

large and it is not clear why it exists. Looking at these two processes in more detail could shed light on the matter.

3.1 *Method and Material*

3.1.1 *Phonological variation and selection of speech material*

As was mentioned above, in this second experiment, we concentrated on the phonological processes of schwa-deletion and schwa-insertion. For both processes the material from Experiment 1 was used and both sets were enlarged to include 75 items.

3.1.2 *Experimental procedure*

Listeners. The main difference in the experimental procedure, compared to the previous experiment, is that the consensus transcription method was used instead of the majority vote procedure to obtain a reference transcription. The listeners that participated in this experiment were all Language and Speech Pathology students at the University of Nijmegen. All had attended the same transcription course. The transcriptions used in this experiment were made as a part of the course examination. Six groups of listeners (5 duos and 1 trio, i.e., 13 listeners) were each asked to judge a portion of the 75 schwa-deletion cases and the 75 schwa-insertion cases. The words were presented to the groups in the context of the full utterance. They were instructed to judge each word by reaching consensus of transcription for what was said at the indicated spot in the word (where the conditions for application of the rule were met). The groups were free to transcribe what they heard using a narrow phonetic transcription.

CSR. The CSR was employed in the same fashion as it was in the first experiment; the task was to choose whether a phone was present or not. Because of this, the tasks for the listeners and the machine were not exactly the same. The listeners were not restricted to choosing whether a phone was present or not as the CSR was, but were free to transcribe whatever they heard.

Evaluation. By allowing the listeners to use a narrow phonetic transcription instead of a forced choice, the consensus transcriptions resulted in more categories than the binary categories used previously: "rule applied" and "rule not applied." This is what we anticipated and an advantage in the sense that the transcription is bound to be more precise. However, in order to be compared with the CSR transcriptions, the multivalued transcriptions of the transcribers have to be reduced to dichotomous variables of the kind "rule applied" and "rule not applied." In doing this different options can be taken which lead to different mappings between the listeners' transcriptions and the CSR's and possibly to different results. Below, two different mappings are presented. Furthermore, for the analysis of these data, we once again chose to use the categories "phone present" and "phone not present" to facilitate the comparison of the processes of deletion and insertion.

The transcriptions pertaining to schwa-deletion obtained with the consensus method were: deletion: ø, different realizations of schwa: ə, ŏ, ə̲, ə̣, əˑ, and other vowels: ĕ, ӟ. There were fewer transcriptions pertaining to schwa-insertion, viz.: not present: ø, different realizations of schwa: ə, ŏ and other vowels: ɐ, ɪ. The mappings chosen in this case were based on the idea that duration may be the cause of the difference between man and machine. Thus, for both processes, we used the following two mappings:

I. deletions (ø) are classified as "phone not present" and the rest is classified as "phone present" [ə, ŏ, ə̲, ə̣, əˑ, ĕ, ӟ, ɐ, ɪ]

II. deletions (ø) and short schwas (ŏ) are classified as "phone not present" and the rest is classified as "phone present": [ə, ə̲, ə̣, əˑ, ĕ, ӟ, ɐ, ɪ]

3.2 *Results*

Tables 4 and 5 show the different transcriptions given by the transcribers for schwa-deletion and schwa-insertion, respectively. The first row shows which transcriptions were used, the second row shows the number of times they were used by the transcribers, the third row indicates the number of times the CSR judged the item as phone present and the last row shows the number of times the CSR judged the item as phone not present. These tables show that deletion, schwa and short schwa were used most frequently, thus the choice of the two mappings is justified as the number of times other transcriptions occurred is too small to have any significant impact on further types of possible mappings.

TABLE 4

Reference transcriptions obtained for the process of schwa-deletion, and the classification of these items by the CSR as present or not present

	ø	ə	ŏ	ə̲	ə̣	əˑ	ĕ	ӟ	**total**
RT	18	37	15	1	1	1	1	1	75
phone present	1	21	5	–	1	1	–	1	30
phone not present	17	16	10	1	–	–	1	–	45

TABLE 5

Reference transcriptions obtained for the process of schwa-insertion and the classification of these items by the CSR as present or not present

	ø	ə	ŏ	ɪ	ɐ	**total**
RT	32	32	8	2	1	75
phone present	6	28	3	2	–	39
phone not present	26	4	5	–	1	36

Figure 8 shows the percentage of schwas present in the CSR's transcriptions and in the reference transcriptions for the processes of schwa-deletion and schwa-insertion, for both mappings. Comparing the CSR's transcriptions to the reference transcriptions once

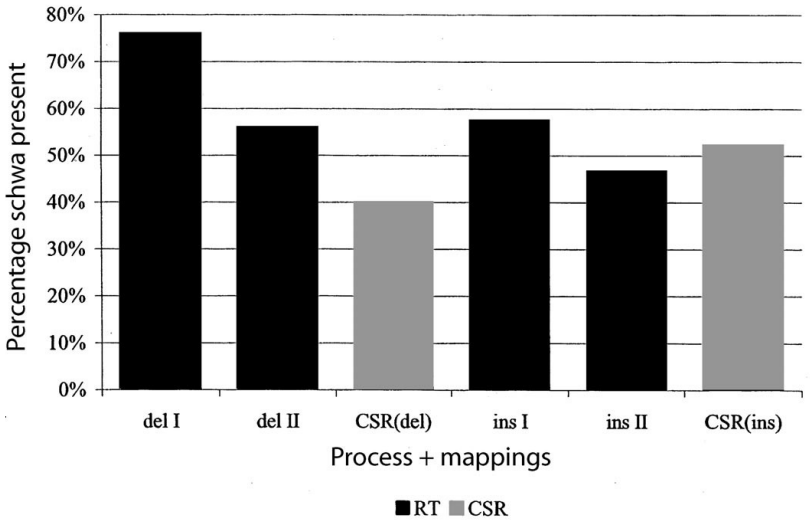

Figure 8

Percentage schwas present for the reference transcription (RT) and for the CSR, for different mappings for the processes of deletion and insertion

again shows that the CSR's threshold for recognizing a schwa is different from the listeners'. In the case of schwa-deletion, this difference becomes smaller when mapping I is replaced by mapping II. For schwa-insertion, replacing mapping I with mapping II leads to a situation where the CSR goes from having a lower percentage of schwa present to having a higher percentage of schwa present than the reference transcription. The difference between the CSR and the reference transcription is significant for schwa-deletion and not significant for schwa-insertion (Wilcoxon, $p < .05$).

Tables 6 and 7 illustrate more precisely what actually occurs. The difference in phone detection between the CSR and the listeners becomes smaller for schwa-deletion (Table 6) if mapping II is used. For this mapping, ə is classified as "phone not present" which causes the degree of agreement between the CSR and the reference transcription to increase. However, it is not the case that all short schwas were classified as "phone not present" by the CSR.

For schwa-insertion (Table 7), the differences in classification by the CSR and by the listeners are not as large. In this case, when the ə is classified as "phone not present" the CSR shows fewer instances of schwa present than the listeners do.

3.3 *Discussion*

The results of this experiment underpin our earlier statement that the CSR and the listeners have different durational thresholds for detecting a phone. A different mapping between the machine and the listeners' results can bring the degree of agreement between the two sets of data closer to each other. It should be noted that the CSR used in this experiment was not optimized for the task, we simply employed the CSR which performed best on a task of pronunciation variation modeling (Kessens, Wester, & Strik, 1999). Although this has not been tested in the present experiment, it seems that changing the machine in such a way that it is able to detect shorter phones more easily should lead to automatic transcriptions that are more similar to those of humans. In other words, in addition to showing how machine and human transcriptions differ from each other, these results also indicate

TABLE 6

Counts of agreement/disagreement CSR and reference transcription (RT) for different mappings of RT categories, for schwa-deletion. **Y**(es) phone present, and **N**(o) phone not present

		RT I			***RT II***		
Mappings		*Y*	*N*	*SUM*	*Y*	*N*	*SUM*
CSR	Y	29	1	30	24	6	30
	N	28	17	45	18	27	45
	SUM	57	18	75	42	33	75

TABLE 7

Counts of agreement/disagreement CSR and reference transcription (RT) for different mappings of RT categories, for schwa-insertion. **Y**(es) phone present, and **N** (o) phone not present

		RT I			***RT II***		
		Y	*N*	*SUM*	*Y*	*N*	*SUM*
CSR	Y	33	6	39	30	9	39
	N	10	26	36	5	31	36
	SUM	43	32	75	35	40	75

how the former could be brought closer to the latter. For instance, the topology of the HMM could be changed by defining fewer states, or by allowing states to be skipped, thus facilitating the recognition of shorter segments.

Although schwa is involved in both cases in this experiment, not much light is shed on the issue of why the processes of insertion and deletion lead to such different results. A possible explanation as far as the listeners are concerned could be the following: For 20 of the schwa-deletion cases, something other than deletion or schwa was transcribed by the listeners compared to nine such cases for schwa-insertion. This indicates that schwa-deletion may be a less straightforward and more variable process. Furthermore, as was mentioned earlier, schwa-deletion is less common than schwa-insertion, which might also influence the judgments of the listeners. So there are two issues playing a role here; the process of deletion might be more gradual and variable than the process of insertion and the listeners may have more difficulties because schwa-deletion is a less frequently occurring process.

Another explanation for the difference is that there is an extra cue for judging the process of schwa-insertion. When schwa-insertion takes place, the /l/ and /r/, which are the left context for schwa-insertion, change from postvocalic to prevocalic position (see Table 8). This change in position within the syllable also entails a change in the phonetic properties of these phones. In general postvocalic /l/s tend to be velarized while postvocalic /r/s tend to be vocalized or to disappear. This is not the case for schwa-deletion, whether or not the schwa is deleted does not influence the type of /l/ or /r/ concerned. These extra cues regarding the specific properties of /l/ and /r/ can be utilized quite easily by listeners, and

TABLE 8

Examples of application of schwa-deletion and schwa-insertion. Syllable markers indicate pre- and postvocalic position of /l/ and /r/

	base form	*rule applied*
schwa-deletion	[la-tə-rə]	[la-trə]
schwa-insertion	[dɛlft]	[dɛ-ləft]

most probably are. They can also be utilized by our CSR because different monophone models were trained for /l/ and /r/ in pre- and post-vocalic position. Thus, whether a schwa is inserted may be easier to judge than whether a schwa is deleted due to these extra cues.

4 General discussion

In this paper, we explored the potential that a technique developed for CSR could have for linguistic research. In particular, we investigated whether and to what extent a tool developed for selecting the pronunciation variant that best matches an input signal could be employed to automatically obtain phonetic transcriptions for the purpose of linguistic research.

To this end, two experiments were carried out in which the performance of a machine in selecting pronunciation variants was compared to that of various listeners who carried out the same task or a similar one. The results of these experiments show that overall the machine's performance is significantly different from the listeners' performance. However, when we consider the individual processes, not all the differences between the machine and the listeners appear to be significant. Furthermore, although there are significant differences between the CSR and the listeners, the differences in performance may well be acceptable depending on what the transcriptions are needed for. Once again it should be kept in mind that the differences that we found between the CSR and the listeners were also in part found between the listeners.

In order to try and understand the differences in degree of agreement between listeners and machine, we carried out further analyses. The important outcome of these analyses is that the differences between the listeners' performance and the machine's did not have a random character, but were of a systematic nature. In particular, the machine was found to have a stronger tendency to choose for absence of a phone than the listeners: the machine signaled more instances of deletion and fewer instances of insertion. Furthermore, in the second experiment, we found that the majority of instances where there was a discrepancy between the CSR's judgments and listeners', it was due to the listeners choosing a short schwa and the CSR choosing a deletion. This underpins the idea that durational effects are playing a role.

In a sense these findings are encouraging because they indicate that the difference between humans and machine is a question of using different thresholds and that by adjusting these thresholds some sort of tuning could be achieved so that the machine's performance becomes more similar to the listeners'. The question is of course whether

this is desirable or not. On the one hand, the answer should be affirmative, because this is also in line with the approach adopted in our research. In order to determine whether the machine's performance is acceptable we compare it with the listeners' performance, which, in the absence of a better alternative, constitutes the point of reference. The corollary of this view is that we should try to bring the machine's performance closer to the listeners' performance. On the other hand, we have pointed out above that human performance does not guarantee hundred percent accuracy. Since we are perfectly aware of the shortcomings of human performance in this respect, we should seriously consider the various cases before unconditionally accepting human performance as the authoritative source.

To summarize, the results of the more detailed analyses of human and machine performance do not immediately suggest that by using an optimization procedure that brings the machine's performance closer to the listeners', better machine transcriptions would be obtained. This brings us back to the point where we started, namely taking human performance as the reference. If it is true that there are systematic differences between human and machine, as appeared from our analyses, then it is not surprising that all agreement measures between listeners were higher than those between listeners and machine. Furthermore, if we have reasons to question the validity of the human responses, at least for some of the cases investigated, it follows that the machine's performance may indeed be better than we have assumed so far.

Going back to the central question in this study, namely whether the techniques that have been developed in CSR to obtain some sort of phonetic transcriptions can be meaningfully used to obtain phonetic transcriptions for linguistic research, we can conclude that the results of our experiments indicate that the automatic tool proposed in this paper can be used effectively to obtain phonetic transcriptions of deletion and insertion processes. It remains to be seen whether these techniques can be extended to other processes.

Another question that arises at this point is how this automatic tool can be used in linguistic studies. It is obvious that it cannot be used to obtain phonetic transcriptions of complete utterances from scratch, but is clearly limited to hypothesis verification, which is probably the most common way of using phonetic transcriptions in various fields of linguistics, like phonetics, phonology, sociolinguistics, and dialectology. In practice, this tool could be used in all research situations in which the phonetic transcriptions have to be made by one person. Given that a CSR does not suffer from tiredness and loss of concentration, it could assist the transcriber who is likely to make mistakes owing to concentration loss. By comparing his/her own transcriptions with those produced by the CSR a transcriber could spot possible errors that are due to absent-mindedness.

Furthermore, this kind of comparison could be useful for other reasons. For instance, a transcriber may be biased by his/her own hypotheses and expectations with obvious consequences for the transcriptions, while the biases which an automatic tool may have can be controlled. Checking the automatic transcriptions may help discover possible biases in the listener's data. In addition, an automatic transcription tool could be employed in those situations in which more than one transcriber is involved; in order to solve possible doubts about what was actually realized. It should be noted that using an automatic transcription tool will be less expensive than having an extra transcriber carry out the same task.

Finally, an important contribution of automatic transcription to linguistics would be that it makes it possible to use existing speech databases for the purpose of linguistic research. The fact that these large amounts of material can be analyzed in a relatively short

time, and with relatively low costs makes automatic transcription even more important (see for instance Cucchiarini & van den Heuvel, 1999). The importance of this aspect for the generalizability of the results cannot be overestimated. And although the CSR is not infallible, the advantages of a very large dataset might very well outweigh the errors introduced by the mistakes the CSR makes.

Received: December 21, 1999; revised manuscript received: October 5, 2000; accepted: December 21, 2000

References

AMOROSA, H., BENDA, U. von, WAGNER, E., & KECK, A. (1985). Transcribing phonetic detail in the speech of unintelligible children: A comparison of procedures. *British Journal of Disorders of Communication*, **20**, 281–287.

BOOIJ, G. (1995). *The phonology of Dutch.* Oxford, U.K.: Clarendon Press.

COHEN, J. A. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213–220.

CUCCHIARINI, C. (1993). *Phonetic transcription: A methodological and empirical study*. Ph.D. thesis, University of Nijmegen.

CUCCHIARINI, C., & HEUVEL, H. van den (1999). Postvocalic /r/-deletion in Dutch: More experimental evidence. *Proceedings of the, 14th International Congress of Phonetic Sciences*, San Francisco, **3**, 1673–1676.

CUTLER, A. (1998). The recognition of spoken words with variable representations. *Proceedings of the ESCA Workshop on the Sound Patterns of Spontaneous Speech: Production and Perception*, Aix-en-Provence, France, 83–92.

DUEZ, D. (1998). The aims of SPoSS. *Proceedings of the ESCA Workshop on the Sound Patterns of Spontaneous Speech: Production and Perception*, Aix-en-Provence, France, VII–IX.

EISEN, B., TILLMANN, H. G., & DRAXLER, C. (1992). Consistency of judgments in manual labeling of phonetic segments: The distinction between clear and unclear cases. *Proceedings of the International Conference on Spoken Language Processing '92*, Banff, Canada, 871–874.

GREENBERG, S. (1999). Speaking in shorthand — a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, **29**(2–4), 159–176.

KEATING, P. (1997). Word-level phonetic variation in large speech corpora. To appear in an issue of *ZAS Working Papers in Linguistics*, Ed. Berndt Pompino-Marschal. Available as <http://www.humnet.ucla.edu/humnet/linguistics/people/keating/berlin1.pdf>.

KERKHOFF, J., & RIETVELD, T. (1994). Prosody in NIROS with FONPARS and ALFEIOS. In P. de Haan & N. Oostdijk (Eds.), *Proceedings of the Department of Language and Speech. University of Nijmegen*, **18**, 107–119.

KERSWILL, P., & WRIGHT, S. (1990). The validity of phonetic transcription: Limitations of a sociolinguistic research tool. *Language Variation and Change*, **2**, 255–275.

KESSENS, J. M., WESTER, M., & STRIK, H. (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication*, **29**(2–4), 193–207.

KUIJPERS, C., & DONSELAAR, W. van (1997). The influence of rhythmic context on schwa epenthesis and schwa deletion in Dutch. *Language and Speech*, **41**(1), 87–108.

KIPP, A., WESENICK, B., & SCHIEL, F. (1997). Pronunciation modeling applied to automatic segmentation of spontaneous speech. *Proceedings of EUROSPEECH '97*, Rhodes, Greece, 1023–1026.

LANDIS, J. R., & KOCH, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

LAVER, J. D. M. (1965). Variability in vowel perception. *Language and Speech*, **8**, 95–121.
MEHTA, G., & CUTLER, A. (1998). Detection of target phonemes in spontaneous and read speech. *Language and Speech*, **31**, 135–156.
OLLER, D. K., & EILERS, R. E. (1975). Phonetic expectation and transcription validity. *Phonetica*, **31**, 288–304.
PYE, C., WILCOX, K. A., & SIREN, K. A. (1988). Refining transcriptions: The significance of transcriber "errors." *Journal of Child Language*, **15**, 17–37.
RISCHEL, J. (1992). Formal linguistics and real speech. *Speech Communication*, **11**, 379–392.
SHRIBERG, L. D., & LOF, L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics*, **5**, 225–279.
SHRIBERG, L. D., KWIATKOWSKI, J., & HOFFMAN, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research*, **27**, 456–465.
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. *Proceedings of EUROSPEECH '93*, Berlin. Germany, 2125–2128.
STRIK, H., RUSSEL, A., HEUVEL, H. van den, CUCCHIARINI, C., & BOVES, L. (1997). A spoken dialog system for the Dutch public transport information service. *International Journal of Speech Technology*, **2**(2), 119–129.
SWERTS, M., & COLLIER, R. (1992). On the controlled elicitation of spontaneous speech. *Speech Communication*, **11**, 463–468.
TING, A. (1970). Phonetic transcription: A study of transcriber variation. *Report from the Project on Language Concepts and Cognitive Skills Related to the Acquisition of Literacy* (Madison: Wisconsin University).
WESTER, M., KESSENS, J. M., & STRIK, H. (1998). Two automatic approaches for analyzing the frequency of connected speech processes in Dutch. *Proceedings of the International Conference on Spoken Language Processing*, Sydney, **7**, 3351–3356.
WITTING, C. (1962). On the auditory phonetics of connected speech: Errors and attitudes in listening. *Word*, **18**, 221–248.

Appendix 1

Number of items in each reference transcription set per excluded listener

	Set of reference transcriptions								
RT Strictness	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5 of 8	445	448	449	443	449	454	453	454	448
6 of 8	407	399	395	403	407	399	403	404	398
7 of 8	353	349	340	341	345	338	347	348	354
8 of 8	273	249	251	256	250	250	262	254	258

Appendix 2

Number of items in each reference transcription set per excluded listener for each of the phonological processes. (Strictness: 5 out of 8 listeners agreeing)

	Set of reference transcriptions								
Phonological processes	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
/n/-del	152	151	155	151	153	152	154	153	154
/r/-del	116	120	115	114	117	120	117	121	118
/t/-del	79	80	81	79	80	82	82	80	78
schwa-del	51	50	51	51	51	52	53	52	51
schwa-ins	47	47	47	48	48	48	47	48	47

Appendix 3

Counts (percentages between brackets) of agreement/disagreement CSR and reference transcription (RT) based on a majority of 5 of 9 listeners agreeing, for all items together and split up for each of the processes. Phone present = Y, and phone not present = N

	phonological processes					
	all	***/n/-del***	***/r/-del***	***/t/-del***	***schwa-del***	***schwa-ins***
RT=Y, CSR=Y	235 (50)	86 (55)	52 (41)	59 (70)	18 (34)	23 (48)
RT=N, CSR=N	143 (31)	53 (34)	44 (35)	9 (11)	14 (26)	20 (42)
RT=Y, CSR=N	67 (14)	9 (6)	26 (20)	11 (13)	20 (38)	4 (8)
RT=N, CSR=Y	22 (5)	7 (5)	5 (4)	5 (6)	1 (2)	1 (2)
Total RT=Y	302 (65)	95 (61)	78 (61)	70 (83)	38 (72)	27 (56)
Total CSR=Y	257 (55)	93 (60)	57 (45)	64 (76)	19 (36)	24 (50)
Total items	467 (100)	155 (100)	127 (100)	84 (100)	53 (100)	48 (100)

Article 2

J. M. Kessens and H. Strik. On automatic phonetic transcription quality: Lower WERs do not guarantee better transcriptions, submitted to Computer, Speech & Language.

Abstract

The first goal of this study is to investigate the effect of several properties of a continuous speech recognizer (CSR) on automatic phonetic transcription. Our results show that changing certain properties of the CSR affects the resulting automatic transcriptions. The quality of the automatic transcriptions can be improved by using 'short' HMMs and by reducing the amount of contamination in the HMMs. The amount of contamination can be reduced by training the HMMs on the basis of a transcription that better matches the actual pronunciation, e.g. by modeling pronunciation variation or by training HMMs on read speech. Furthermore, it appeared that context-dependent HMMs should not be trained on canonical transcriptions since the transcriptions obtained with these HMMs are too much biased towards the canonical transcriptions. Finally, we found that by combining these changes in properties of the CSR the quality of automatic transcription can be further improved.

The second goal of this study is to find out whether there exists a relation between the word error rate (WER) and transcription quality. As no clear relation was found, we conclude that in order to obtain automatic transcriptions taking the CSR with the lowest WER does not always provide the optimal solution.

1. Introduction

Phonetic Transcriptions (PTs) of speech are needed in many disciplines. In linguistic research, for instance phonetics, phonology, sociolinguistics, and dialectology, PTs form a vital component of the research methodology. In speech pathology, PTs are needed in research and in clinical practice. In clinical applications, PTs are used for diagnostic purposes in order to measure the severity of the handicap or disability (Shriberg & Lof, 1991), and during treatment programmes, to monitor and document progress (or lack thereof). Furthermore, PTs are used in speech technology, both in speech synthesis and in automatic speech recognition (ASR). For the development of speech synthesis systems, a phonetically transcribed database is needed from which diphones and/or larger concatenation segments can be extracted, and of which the segmentation can be used for duration modeling of the concatenation units (Ljolje, Hirschberg, & van Santen, 1997). During the last decades, one of the approaches that has been used to improve ASR is by modeling pronunciation variation (for an overview see Strik and Cucchiarini, 1999). Reliable and accurate PTs of speech form an essential resource for this type of research.

PTs can be obtained in two ways. Manual Phonetic Transcriptions (MPTs) are made by experts who listen to an utterance and transcribe it into a sequence of speech units represented by phonetic symbols. These experts may use the full set of IPA[8] symbols, including diacritics, to produce what is known as 'narrow phonetic transcriptions'. However, making MPTs is extremely time-consuming and therefore costly. Moreover, MPTs tend to contain an element of subjectivity (Shriberg & Lof, 1991). The time needed to make MPTs can be reduced by limiting the transcription

[8] http://www2.arts.gla.ac.uk/IPA/ipa.html

process to a few phenomena that are of special interest for the study at hand, such as the presence or absence of the vowel schwa (Kuijpers & van Donselaar, 1997). Time investment can also be diminished – and accuracy improved (see Shriberg & Lof, 1991) - by using broad phonetic transcriptions, i.e. transcriptions in which only the subset of the symbols is used that correspond to the phonemes of the language.

PTs can also be made automatically, i.e. by a speech recognizer: This results in what we will call Automatic Phonetic Transcriptions (APTs) in this paper. Almost invariably, APTs are 'broad phonetic', or phonemic transcriptions. This is a direct consequence of the fact that virtually all operational ASR systems are trained to handle only the 'phonemes' of the target language. APTs are much faster to make, and therefore much cheaper, than MPTs. However, before APTs of large corpora can be used as the raw data for research in speech science or technology, many questions about accuracy and also reliability must be answered. APTs are certainly reliable in the sense that the same material transcribed by the same ASR will result in identical output. However, it is much less self-evident that transcriptions of the same material by different ASR systems will show a high degree of agreement. Differences between the transcriptions of ASR systems parallel the subjectivity that is inherent in MPTs.

APTs can be made in various ways. One approach is to perform phone recognition. In this kind of recognition, instead of words - as is the case during a normal recognition task - phones are recognized. Often, the recognizer is constrained by a phone N-gram, and by penalties on the generation of many short sequences of phones. In the cases when the content of an utterance (the orthographic transcription) is available, a second kind of APTs can be made. The phonetic transcriptions of the words in the utterance are then used as a starting point for automatic transcription. This phonetic transcription can be looked up in a lexicon or can be obtained by means of a grapheme-phoneme converter. Next, a number of possible pronunciation variants are generated on the basis of the phonetic transcription, e.g. by applying phonological rules (e.g. Adda-Decker & Lamel, 1998), data-derived rules (e.g. Kessens & Strik, 2001) or by means of D-trees (e.g. Riley et al., 1998). The task of the recognizer is then to decide for each word, which of the variants best matches the acoustic signal. This study is an example of this second kind of APT. The number of transcription variants is restricted by allowing only pronunciation variants generated by applying five phonological rules to the canonical transcriptions. Other research (Wester, Kessens, Cucchiarini, & Strik, 2001; Saraclar, 2000) showed that for such a transcription task, APTs can be made that form acceptable substitutes for MPTs.

In order to evaluate the quality of our APTs, each APT is compared to a human Reference Transcription (RT). However, given that humans can make mistakes there is no completely error free RT with which the automatic transcriptions can be compared (Cucchiarini, 1993: 11-13). To circumvent this problem (at least partly), the following two strategies have been devised for obtaining a human RT:

1) A consensus transcription is used, which is a transcription made by several transcribers after they have agreed on each individual symbol (Shriberg, Kwiatkowski & Hoffman, 1984).

2) A majority vote principle is used, which means that the material is transcribed by more than one transcriber and that only the part of the material is used for which all transcribers agree (Kuijpers & van Donselaar, 1997), or at least the majority of them (Wester et al., 2001).

In this paper, both strategies to obtain RTs are used. We use agreement between the APTs and the human RTs as a measure of quality for the various APTs; the higher the agreement with the human RTs, the better the quality of the APT.

In our previous study (Wester et al., 2001), we simply employed the CSR used in other research without trying to optimize it to make the CSR's transcriptions more similar to the human transcriptions. It is likely that properties of the CSR, such as for instance the speech material used for training and the procedure to estimate the acoustic models, all influence the APTs. This holds true for phone recognition and for selection of pre-defined pronunciation variants. Some research on this issue has already been carried out. In the study reported in Saraçlar (2000a) and Saraçlar, Nock & Khudanpur (2000b) different techniques to improve APTs are investigated. For evaluation, phone accuracy is calculated with MPTs as the reference. These experiments reveal that the following techniques hardly influence the accuracy of the APTs: speaker and channel adaptation, acoustic models with lower resolution (less Gaussian mixtures) and jack-knifing, i.e. one half of the training data is used to transcribe the other half. They conclude that it is quite difficult to further improve automatic phonetic transcription using acoustic models trained on canonical transcriptions. For this reason, acoustic models are trained on hand-labeled data or on data for which automatic transcriptions are made using a pronunciation model based on the same hand-labeled data. These acoustic models appeared to substantially improve automatic transcription compared to the baseline models that are trained on canonical transcription of the training material. Another study is conducted by Cox, Brady & Jackson (1998). These authors compared various automatic transcription systems by calculating phone accuracy between APTs and MPTs made by a professional phonetician. They found that speaker adaptation improves the quality of APTs. Besides adaptation, they used confidence measures to label phones, and trained acoustic models using the phones for which the confidence value exceeded a certain threshold. These acoustic models further improved the quality of the APTs. Finally, the work of Brugnara, Falavigna & Omologo (1993) is mainly concerned with segmentation of speech. As part of this research, these authors investigated the effects of the topology of the HMMs with phone accuracy as evaluation criterion. They found an optimal accuracy for HMMs that have a minimum duration of 20 ms.

The first goal of this paper is to investigate and compare a number of properties of ASR systems for their effects on the quality of APTs. In addition to some of the properties described above, we will also investigate the impact of the type of acoustic models (e.g., context dependent versus context independent models) and the type of speech material used to train these models.

In previous research on APT (Wester et al., 2001), we simply took the CSR with the lowest Word Error Rate (WER) that was available from our research on pronunciation variation modeling. In other research on APT, the choice of the CSR usually is not clearly motivated. Intuitively one might expect that the ASR system that obtains the lowest WER on some reference recognition task will also yield the best APTs. However, on second thoughts (automatic) speech recognition may well appear to be quite a different task than (automatic) phonetic transcription. Therefore, it is worthwhile to investigate whether lower WERs do indeed indicate higher quality APTs. This is the second goal of the research reported here.

This paper is organized as follows: In section 2, the method that we employed is illustrated. Subsequently, in section 3, we present the results for each of the properties of the ASR system that are investigated. The relation between degree of agreement and WER is examined in section 4. Finally, in section 5, we discuss the results, while in section 6 we present our general conclusions.

2. Method

As explained in the introduction, the focus of this study is a restricted form of automatic transcription. Only the pronunciation variants that are automatically generated by the application of five phonological rules to the canonical transcriptions can be chosen by the recognizer. In section 2.1, we will first explain which pronunciation variants are selected for transcription. Next, in section 2.2, we will describe the speech material and our CSR. Section 2.3 describes how the APTs and the two kinds of RTs are obtained and what the differences are between the two kinds of transcription tasks. For evaluation of the various APTs, we calculate agreement between the various APTs and human RTs, as will be explained in section 2.4.

2.1 Pronunciation variants

The pronunciation variants were automatically generated by applying a set of phonological rules to the canonical transcriptions of the words that occur in the transcription material. For variant generation, we used five phonological rules concerning deletions and insertions of phones: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion (SAMPA[9] notation is used throughout this paper). The main reasons for selecting these five phonological processes are that they occur frequently in Dutch and are well described in the linguistic literature. Furthermore, these phonological processes typically occur in fast or extemporaneous speech; therefore, it is to be expected that they will occur in the speech material that we use (see section 2.2). Table 1 provides an example of each rule. The deleted phones are shown between '(..)', and the inserted phone is indicated by '[..]'. For more details and a description of the five phonological rules, see Wester et al. (2001).

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

Table 1: Examples of the five phonological rules

Rule	Example			Orthography	Translation
/n/-deletion	rEiz@(n)	→	rEiz@	reizen	to travel
/r/-deletion	Amst@(r)dAm	→	Amst@dAm	Amsterdam	Dutch city: ‘Amsterdam’
/t/-deletion	sa:vOn(t)s	→	sa:vOns	's avonds	in the evening
/@/-deletion	la:t(@)r@	→	la:tr@	latere	later
/@/-insertion	dElft	→	dEl[@]ft	Delft	Dutch city: ‘Delft’

The transcription task can be considered to be a binary decision task, since the CSR (and the humans in one of the two approaches used to produce RTs) must decide whether a rule was applied or not. For analysis purposes, we treated the transcription task as a binary decision task: For each phone that can possibly be deleted or inserted since the condition for one of the five rules is met, a binary score is obtained: (1) if the rule is applied and (0) if this is not the case. To clarify this, let us consider the following example: For the word /dELft/ (‘Delft’) the rule conditions for the /t/-deletion and the /@/-insertion are met; thus, four pronunciation variants are generated. Table 2 shows the four variants (column 1), the rules that are applied (column 2), and the corresponding binary scores (column 3).

Table 2: Example of pronunciation variants and corresponding binary scores

pronunciation variant	rules that are applied	binary scores
/dELft/	none	/t/-deletion=0, /@/-insertion=0
/dElf/	/t/-deletion	/t/-deletion=1, /@/-insertion=0
/dEl@ft/	/@/-insertion	/t/-deletion=0, /@/-insertion=1
/dEl@f/	/@/-insertion + /t/-deletion	/t/-deletion=1, /@/-insertion=1

2.2 Speech material and CSR

The speech material used in the experiments is taken from a Dutch database, which contains a large number of telephone calls recorded with the on-line version of a spoken dialogue system called OVIS (Strik, Russel, van den Heuvel, Cucchiarini & Boves, 1997). OVIS is employed to automate part of an operational Dutch public transport information service. The speech material consists of interactions between man and machine, and can be described as extemporaneous or spontaneous. From the VIOS material, two sets of data are selected and for each data set a different kind of human RT is obtained. For the first set, a reference transcription is employed based on a *majority vote* procedure. This set is equal to the one that was used in Wester et al. (2001). For the second set, a *consensus transcription* is made. The statistics of the two sets of transcription material are given in Table 3. In the column ‘#utts’ and ‘#words’, the number of utterances and words in the set is given. The remaining columns display the number of times a condition for rule application is met, and thus the number of binary scores that are obtained. The two sets of material are selected in such a way that the relative frequencies of potential and actual application of the rules correspond more or less to the relative rule frequencies in the training material. For the /@/-

deletion and /@/-insertion rules, the relative frequencies of potential application are higher so as to obtain a sufficiently high number of observations.

Table 3: Statistics of transcription material

set	reference	# utts	# words	/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins	all
1	majority vote	186	1208	155	127	84	53	48	467
2	consensus	296	2035	287	230	109	41	103	770
	TOTAL:	482	3243	442	357	193	94	151	1237

We used a standard CSR that is part of the spoken dialogue system OVIS (Strik et al., 1997). The baseline phone models are continuous density HMMs with 32 Gaussians per state. Every 10 ms, 14 cepstral coefficients (including c_0) and their deltas are calculated for frames with a width of 16 ms. The HMMs are trained on 25,104 VIOS utterances (81,090 words), which do not overlap with the material that was manually transcribed. The baseline HMMs consist of a tripartite structure; each of the three parts consists of two identical states, one of which can be skipped (Steinbiss et al., 1993). In total, 38 HMMs are trained. For 35 of the phonemes, context-independent HMMs are trained. In addition, one model is trained for non-speech sounds, one model is used for filled pauses, and a model consisting of one state is employed to model silence. The baseline lexicon contains one transcription for each word. These canonical transcriptions are obtained using the grapheme-phoneme-converter which is part of a Text-to-Speech system for Dutch (Kerkhoff & Rietveld, 1994), followed by a manual correction. The only rule that is applied in the canonical transcriptions is the /n/-deletion rule, since the pronunciation without the /n/ is considered to be the most likely pronunciation according to the linguistic literature (van de Velde, 1996). The CSR uses a unigram and bigram language model, which is trained on the same 25,104 VIOS utterances used to train the acoustic models.

2.3 Automatic transcriptions and human reference transcriptions

2.3.1 Automatic transcriptions

The CSR is used to make the APTs. To this end, pronunciation variants are automatically generated by applying the five phonological rules (see section 2.1) to the canonical transcriptions of the words. The task of the CSR is to determine which of the generated variants best matches the acoustic signal. We refer to this type of recognition as *forced recognition*, since the CSR is forced to choose among a number of pronunciation variants. During forced recognition, all variants of the same words are assigned the same language model probability; thus, variant selection is completely determined by the acoustics. For more details on our approach to forced recognition, see Wester et al. (2001). The details on each investigated property of the CSR are given together with the results in section 3.

2.3.2 Majority vote reference transcriptions

The majority vote reference transcriptions are identical to those made in Wester et al. (2001). We briefly summarize the relevant points of this transcription task; for more details, see Wester et al. (2001). The transcriptions were made by nine expert listeners who listened to the speech signal and decided which pronunciation variant best matched the realization that they had just heard for each of the 379 words in Table 3. In this sense, their task was exactly the same as the CSR's, i.e. deciding which pronunciation variant best matched the speech signal. The listeners were selected to participate in this experiment because they all had carried out similar tasks for their own investigations. For this reason, they are representative of the kind of people who may benefit from automatic ways of obtaining such transcriptions. The RTs were determined by a majority vote procedure, which implies that the transcription that is produced by the majority of the listeners (5 or more out of 9) is taken to be the human RT.

2.3.3 Consensus reference transcriptions

The transcribers who made the consensus reference transcriptions are Language and Speech Pathology students at the University of Nijmegen. They had all attended the same transcription course including 32 hours contact time. The transcriptions used in this experiment were made as part of the final examination. The IPA transcription alphabet is used in this course. The transcribers all worked in one of 12 groups of two or three people (eleven duos and one trio) and based their transcriptions on auditory analysis of the full utterances without any kind of visual support. The groups of listeners made consensus transcriptions for whole utterances, which implies that two (or three) listeners had to agree on each symbol in the utterance. The utterances of the transcription material were distributed over the groups in such a way that the number of words that each group had to transcribe was about equal. No overlap existed between the transcription material of the different groups.

The consensus transcriptions cannot directly be used for analysis, as they are produced using the whole range of IPA symbols and diacritics, whereas the CSR uses a limited set of SAMPA symbols. For this reason, the diacritics are discarded and the IPA-symbols are mapped to SAMPA symbols, as is shown in Table 4.

Table 4: mapping of IPA to SAMPA symbols

IPA	n, m*	r, ʀ, ʁ, ɾ, ɻ , ɽ , ɺ	t	ə, ʒ
SAMPA	n	r	t	@

* *the /*m*/ is only allowed in case of nasal assimilation*

The different IPA symbols shown in Table 4 are all allophonic variants of the phone that is represented by the corresponding SAMPA symbol. In case the consensus transcription is not an allophonic variant but a different phoneme, then this transcription is excluded from further analysis. In total 22 consensus transcriptions were excluded: 1 /n/-deletion, 16 /r/-deletion, 2 /t/-deletion, 2 /@/-deletion and 1 /@/-

insertion transcriptions. This results in the number of transcribed phones as presented in Table 3.

2.3.4 Differences between majority vote and consensus reference transcriptions

As mentioned above, the two transcription procedures described in the previous sections are two attempts of obtaining human transcriptions that approach the actual speech realisations as much as possible. However, there are differences between these two procedures, which might have effects on the results obtained. First, the majority vote transcription is based on transcriptions that are made independently by various transcribers, whereas in the consensus transcription task the transcribers work together to produce one single transcription. In other words, in the first case the transcribers do not influence each other, while in the second case they do. This form of influence between transcribers may work either positively or negatively. It has been reported that if one of the transcribers is clearly more experienced and competent than the others, "the consensus transcription may be biased to reflect the judgements of the more competent, higher ranked, or 'forceful' transcriber" (Shriberg et al., 1984: 458). However, in many cases this influence helps resolve cases of disagreement between transcribers that are caused by the fact that one of the transcribers was "inattentive to a particular phonetic behaviour, which was immediately obvious upon replay" (Shriberg et al. 1984: 464).

Another difference between the two procedures as they were applied in our experiments is that the transcribers that made the majority vote transcriptions (linguists) were much more experienced than those who made the consensus transcriptions (Language and Speech Pathology students). In view of the possibilities of having bias in the data when transcribers of different status make the consensus transcription, this choice appears to be a plausible one, as status differences seem more likely among linguists than among students. However, the differences in degree of experience may affect the results in another way. For example, it seems reasonable to assume that linguists will be much more aware of the various phonological processes that can occur in Dutch than students are. As a consequence they may be more attentive to details that are otherwise ignored by students. However, these types of expectations may also bias their transcriptions.

Furthermore, an important difference between these two procedures concern the number of subjects involved. The majority vote transcription was based on input from nine subjects, whereas the consensus transcription was produced by two and, in one case, by three subjects. Given the differences in procedure, this seems logical, as it would be very time-consuming to obtain a consensus transcription from nine people. However, we have to realise that this has methodological consequences in terms of transcription reliability. The notion of reliability in relation to phonetic transcription is described in Cucchiarini (1993: 10): "The reliability of a measuring instrument represents the degree of consistency observed between repeated measurements of the same object made with that instrument. It is an indication of the degree of accuracy of a measuring device [...] The notion of reliability is related to the idea that each

measurement is subject to some degree of error and, therefore, each score can be seen as a combination of error and true value [...] Mathematically, the true value is defined as the limit of the average as the number of observations approaches infinity. ” It follows that a measurement based on larger number of observations is bound to be more accurate than one based on a smaller number of observations. Therefore, in our case the majority vote transcription can be assumed to be more accurate than the consensus transcription.

Finally, the last difference is that the two transcription tasks were quite different. The majority vote transcribers were specifically instructed to decide whether one of the five optional phonological rules under investigation was (or was not) applied in specific words in the utterance. The consensus transcribers, on the other hand, were not aware of the purpose of the investigation. Their task consisted of transcribing all sounds in the utterances, which means that they had to pay attention to all phonetic phenomena in the utterances. Through this difference in focus, the majority vote transcribers probably base their decisions on more subtle differences than the listeners who make the consensus transcriptions. Furthermore, by focusing on a few phenomena, the reliability of the transcriptions might also be improved.

2.4 Evaluation of the APTs

The APTs are evaluated by comparing them to the human RTs. To this end, the binary scores of the APTs are compared to the binary scores that are derived from the RTs. As a measure of agreement between the APTs and the RTs we use Cohen’s κ, which corrects percentage agreement for chance agreement (Cohen, 1968):

- $\text{Cohen's } \kappa = \frac{P_o - P_c}{100 - P_c}$ (1)

  $-1 \leq \kappa \leq 1$

  P_c = percentage agreement on the basis of chance

  $P_o = 100\% \times \frac{\#\text{agreements}}{\#\text{agreements} + \#\text{disagreements}}$ (2)

Table 5 shows the qualifications for κ-values greater than zero, to indicate how the κ-values should be interpreted (taken from Landis & Koch, 1977).

Table 5: qualifications for κ-values > 0

κ-value	qualification
0.00 - 0.20	slight
0.21 - 0.40	fair
0.41 - 0.60	moderate
0.61 - 0.80	substantial
0.81 - 1.00	almost perfect

3 Results

The first aim of this investigation is to determine how various properties of ASR systems affect the quality of APTs. The properties of the CSR that are investigated are all related to the HMMs. The general procedure is to take our baseline CSR and substitute it with a different set of HMMs for which each of the following properties is changed:

1) HMM topology (section 3.1)
2) Degree of contamination of the HMMs (section 3.2)
3) Context-independent versus context-dependent HMMs (section 3.3)
4) Combinations of 1) to 3) (section 3.4)

Each section in this chapter starts with a description of the investigated property of the CSR. As there are differences between the majority vote en consensus transcription procedures (see section 2.3.4), we present κ-values for the two sets of material separately. Both the total agreement values (=agreement for all rules) and the agreement values per rule are presented. Finally, each section ends with a discussion of the results and some concluding remarks.

3.1 Topology of the HMMs

In Wester et al. (2001), we found that, in general, our CSR detects fewer phones as present than the humans do. Figure 1 shows the percentages 'phone present' in the human RTs and in the APTs made with the baseline HMMs. In Figure 1, these percentages are given for: a) the majority vote material, and b) the consensus material. Figure 1 shows that for all rules and in each of the data sets the humans tend to detect more phones than the CSR. For the majority vote material, the difference is largest for the /@/-deletion and /@/-insertion rules, whereas for the consensus material the differences in percentages 'phone present' are comparable across rules.

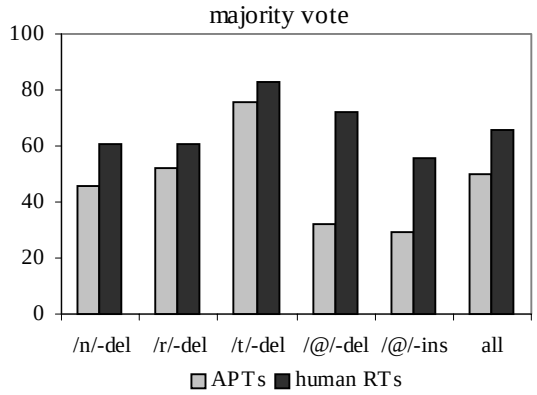

Figure 1a: Percentages 'phone present' for the majority vote material

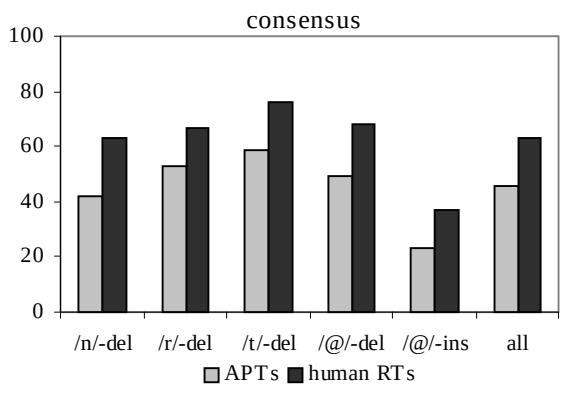

Figure 1b: Percentages 'phone present' for the consensus material

The results in Wester et al. (2001) showed that agreement between the APTs and the human RTs (consensus transcriptions) increased if the /@/s which were judged to be short in duration by the humans were denoted as 'not present'. This could be an indication that the minimum duration associated with the HMM topology is too long, with the consequence that it may be difficult for the CSR to detect short duration /@/s. In this paper, we define *topology length* as the duration corresponding to the minimum number of states to visit from the beginning to the end of the HMM model. Since the baseline HMMs consists of 6 states of which 3 can be skipped, the topology length of the baseline /@/ HMM is 3 states, or 30 ms.

Brugnara et al. (1993) pointed out that topology length is a critical point for automatic segmentation of speech. The topology length of an HMM should be shorter than the minimum phone duration in order to avoid skipping of models. However, using a too short topology length (without any duration model) can cause a high insertion rate. In order to investigate the optimal topology length, Brugnara et al. (1993) compared various HMM topologies, with phone recognition rate as an evaluation criterion. They found an optimal accuracy for HMMs that have a minimum duration of 20 ms. This result might be an indication that our HMM topology length of 30 ms is suboptimal for the task of automatic transcription.

We decided to investigate the effect of using HMMs with topology lengths shorter than 30 ms on the task of automatic transcription. For two reasons, we started off by only changing the HMM topology for the phone /@/. First, the majority vote transcriptions showed very large differences in the numbers of /@/s that are denoted as present by the CSR and by the humans. Second, the results reported in Wester et al. (2001) indicate that duration might be a factor that plays a role in the difference in the number of /@/s transcribed by humans and CSR. For training of the short /@/ HMMs, we first made a segmentation of the training material using the baseline HMMs. In order to determine the duration of the phone /@/ in the training material, the /@/ must be present in the transcriptions used for segmentation. Therefore, the canonical transcriptions were used for all words, except for those to which the /@/-insertion rule is applicable. For these words, we inserted a /@/ at all places where the rule condition for /@/-insertion was met. Subsequently, we determined the number of frames that were assigned to each /@/. Next, we divided the /@/-s into two categories:

1. *short /@/*: the duration in the segmentation is exactly 3 frames (30 ms); 1796 /@/s
2. *long /@/*: the duration in the segmentation is > 3 frames (>30 ms);18,640 /@/s

All short /@/s were then used to train an HMM consisting of 1 segment (2 identical states of which one can be skipped), with a topology length of 10 ms. The long /@/s were used for training the long-/@/ HMM, consisting of 3 segments. In addition to this HMM set, another set of HMMs was trained. For this model set the short /@/ HMM has a 2 segment topology, and thus a topology length of 20 ms.

In order to find out whether using a short /@/ HMM indeed results in higher frequencies of /@/ in the APTs, which in turn increases agreement, the results of the /@/-deletion and /@/-insertion rule are investigated in more detail. First of all, we expect that by using the short /@/ HMM, more /@/s will be transcribed by the CSR. Table 6 shows the percentage of /@/s that are denoted as 'present' by the CSR. The

following abbreviations are used: ‘3seg’ denotes the baseline HMMs with a 3-segment topology for the phone /@/, ‘2seg’ denotes the 2-segment topology, and ‘1seg’ denotes the 1 segment topology. In Table 6, it can be seen that the percentages ‘/@/ present’ indeed increase when using the short-@/ HMM. Especially for the /@/-deletions of the majority vote material the discrepancy between the percentages ‘/@/ present’ is decreased (the percentages ‘phone present’ are doubled).

Table 6: Percentages ‘/@/ present’ for HMMs with various topology lengths and human RTs

rule	majority vote				consensus			
	APTs			human RTs	APTs			human RTs
	3 seg	2 seg	1 seg		3 seg	2 seg	1 seg	
/@/-deletion	32%	57%	68%	72%	49%	51%	61%	68%
/@/-insertion	29%	33%	38%	56%	23%	32%	33%	37%

Second, we expect that agreement will increase for the /@/-deletion and /@/-insertion rule. Figure 2 shows the agreement values per rule. These data reveal that the increase in the number of /@/s that are detected by the CSR (as shown in Table 6) does not necessarily mean that agreement is also increased: There is a decrease in agreement for the majority vote transcriptions of the /@/-insertion rule. Another observation that can be made from Figure 2 is that the use of a shorter topology length for the phone /@/ also influences the agreement values for the other rules.

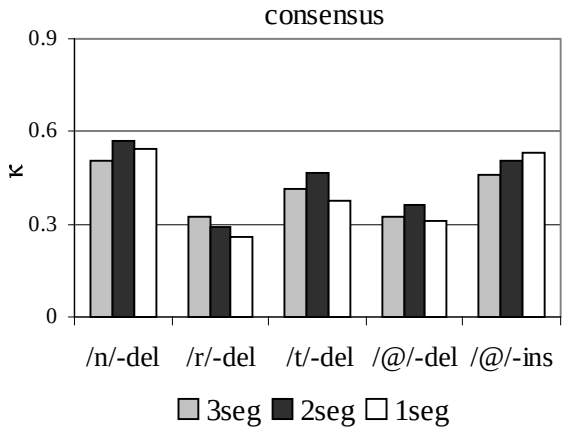

Figure 2a: Agreement values per rule majority vote transcriptions

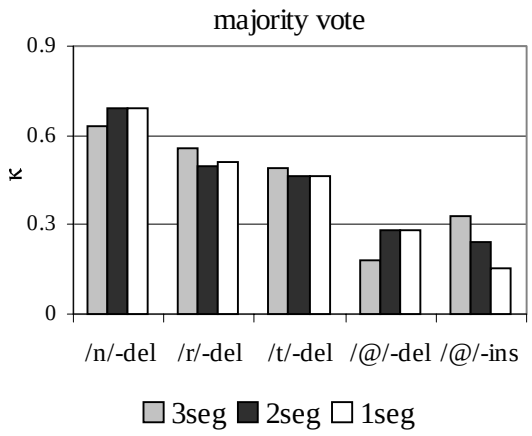

Figure 2b: Agreement values per rule for consensus transcriptions

In Figure 3, the total agreement values are given for HMMs with various topology lengths for the phone /@/. As agreement deteriorates for the /@/-insertion rule of the majority vote material and also for some of the other rules (see Figure 2), it is not surprising that the improvement in the total agreement values is not very large when a short /@/ HMM is used. Another observation that can be made from Figure 3

is that for the consensus transcriptions an HMM topology length of 2 segments - or 20 ms – performs slightly better than the two other lengths. This is in line with the results of Brugnara et al. (1993), since they also found an optimal topology length of 20 ms.

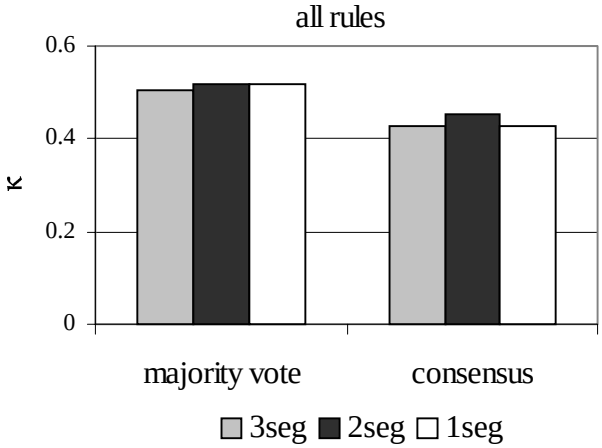

Figure 3: Total agreement values for HMMs with various topology lengths

The question that arises is why the agreement values for the /@/-insertion rule do not increase for the majority vote transcriptions while they do increase for the consensus transcriptions. This difference in result can probably be attributed to the listeners. It is striking that the listeners in the majority vote transcription task tend to choose for considerably more /@/-insertions (56%) than the humans who make the consensus transcriptions (37%) (see Figure 1). This difference in result might be explained by a difference in experience level between the majority vote and consensus transcribers. Furthermore, the different way in which the two kinds of transcriptions are made is probably another factor that is playing a role (see section 2.3.4).

To conclude, using a shorter topology length for the phone /@/ improves the total agreement values, but the improvements are very small. Furthermore, agreement is improved for the /@/-insertion rule of the consensus material, whereas this is not the case for the majority vote transcriptions. As mentioned in section 2.3.4, it seems reasonable to assume that the linguists who made the majority vote transcriptions will be much more aware of the various phonological processes that can occur in Dutch than the students who made the consensus transcriptions. This bias through expectation might be strong for the /@/-insertion rule as it is a frequently occurring process (Kuipers & van Donselaar, 1997). Another factor mentioned in section 2.3.4 is that the majority vote transcribers were aware of the purpose of this investigation. This difference of focus might bias the majority vote transcribers towards more /@/-insertions.

3.2 Degree of contamination of the HMMs

The speech material used for training contains much variation in pronunciation, whereas the baseline training lexicon contains only one canonical transcription for each word. Therefore, some of the transcriptions used for training the baseline HMMs

will be incorrect, e.g. a phone is present in the transcription but has not been realized. Through this mismatch between transcription and pronunciation the HMMs get contaminated. Subsequently, the contamination can lead to errors in the automatic transcriptions. The effect of contamination of the HMMs on automatic transcription will probably be that the APTs are more biased towards the transcriptions on which the HMMs are trained. To better illustrate our point, we can look at the following example: We train our baseline HMMs on the basis of canonical transcriptions of the training corpus in which /@/-insertion is not applied. Consequently, if the /@/ is present ('inserted') in the pronunciation, the HMM of the adjacent phones get contaminated with acoustic signal of the/@/. Through this contamination, the baseline CSR probably tends to choose less easily for /@/-insertion: If the /@/ is pronounced it can still be transcribed as not since the HMM for the adjacent phones contains acoustic information of the /@/.

The effect of contamination of our baseline HMMs probably will be that they are biased towards the transcriptions on which the HMMs are trained, i.e. the canonical transcriptions. By removing (some of) the mismatch between the transcription on which the HMMs are trained and the actual pronunciation, the bias can be reduced. Saraçlar (2000a) reported that this is indeed the case: The baseline HMMs that are trained on canonical transcriptions produce more canonical APTs than HMMs that are trained on the basis of automatic or manual transcriptions of the training material in which pronunciation variation is transcribed.

In this section, we will investigate whether using less contaminated HMMs is beneficial to automatic transcription. To this end, we used two kinds of HMMs that we expect to be less contaminated than the baseline HMMs, namely HMMs from pronunciation variation modeling research and HMMs that are trained on read speech material.

3.2.1 Modeling of pronunciation variation

One of the approaches we used to minimize the mismatch in the training corpus consists of modeling pronunciation variation (Wester, Kessens & Strik, 1998). In this research, automatic transcriptions of pronunciation variation are made by means of forced recognition. The new automatic transcriptions are then used to train new HMMs. From this pronunciation variation research, two sets of HMMs were taken that were used in addition to the baseline HMMs for making automatic transcriptions:

1. HMMs trained on a corpus for which automatic transcriptions of within-word pronunciation variants are made ('within HMMs'). These variants are generated using the same five within-word phonological rules as mentioned in section 2.1.

2. HMMs trained on a corpus for which also cross-word variation is transcribed ('within + cross HMMs'). For more details on the cross-word variation modeled, see Wester et al., 1998.

Figure 4 shows the total agreement values for the baseline HMMs and the HMMs from pronunciation variation research. It can be seen that for both data sets the total agreement values increase when less contaminated HMMs are used. These results are in line with the findings of Saraçlar (2000a).

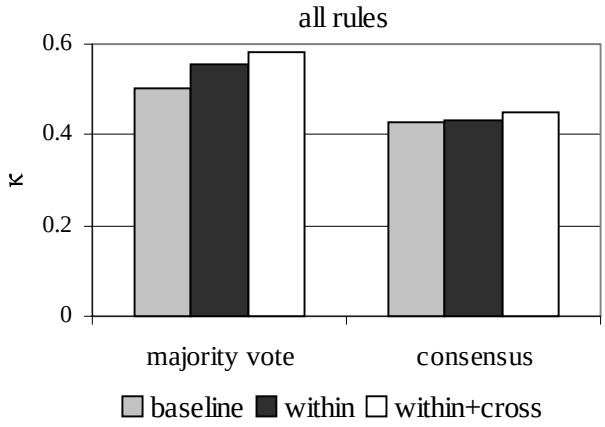

Figure 4: Total agreement values for the baseline HMMs and for HMMs from pronunciation variation research

Saraçlar (2000a) also showed that the pronunciation variation HMMs are less biased towards the canonical transcriptions than baseline HMMs. Closer inspection of our data reveals that also in our material the CSR tends to choose less often for canonical transcriptions; the percentage of canonical APTs for all rules decreases from 57.9% for the baseline HMMs, to 50.6% and 50.8% for respectively the 'within' and the 'within+cross' HMMs. This tendency is also observed per rule (see Appendix 1).

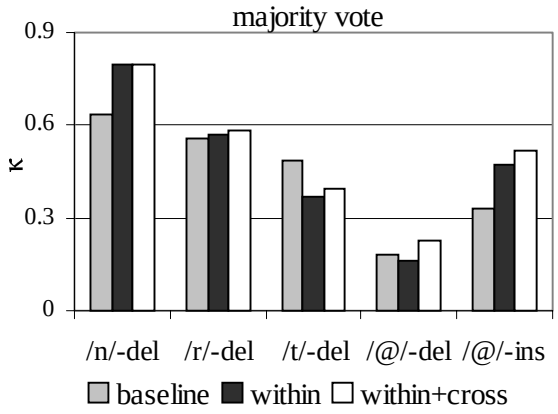

Figure 5a: Agreement values per rule for the majority vote transcriptions

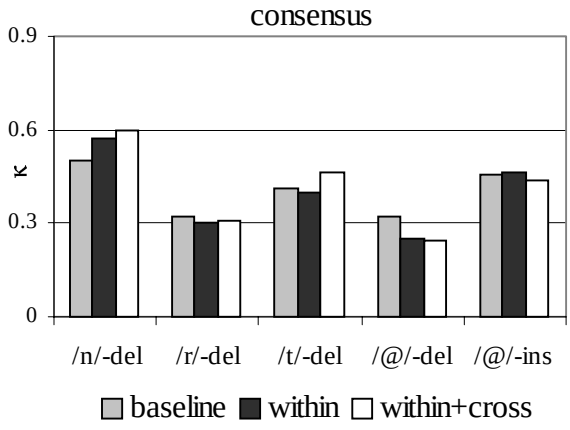

Figure 5b: Agreement values per rule for the consensus transcriptions

As the total agreement values increase if pronunciation variation HMMs are used, one could easily conclude that the contamination of the baseline HMMs indeed leads to transcription errors. However, the results per rule do not confirm this hypothesis unconditionally since agreement is not increased for all rules (see Figure 5).

Another observation that can be made from Figure 5 is that the agreement values are considerably increased for the /@/-insertion rule of the majority vote material, whereas this is not the case for the consensus material. The fact that we again find discrepancies in the results of the majority vote and consensus transcriptions of the /@/-insertion rule confirms our hypothesis that the way in which the transcribers decide on the application of this rule is different in the two transcription tasks.

Finally, it should be noted that the increase in agreement values is mainly caused by an increase in agreement for the /n/-deletion rule (see Figure 5). Another way of modeling pronunciation variation is to take the most frequent transcription of a word as the transcription in the lexicon (Cohen, 1989). If the most frequent transcription of a word is used in the training lexicon, the number of words for which there is a mismatch between the transcription and the realized pronunciation is reduced, which probably leads to better transcription quality. The /n/-deletion rule is the only rule for which the canonical transcription is not the most frequent one according to the human transcribers: The baseline HMMs are trained on transcriptions in which /n/-deletion is applied, whereas in our speech material the percentage of /n/-deletions according to the transcribers is less than 50% (see Figure 1). In order to investigate whether HMMs trained on the most frequent transcription of a word is beneficial to automatic transcription quality, we trained new HMMs on the basis of transcriptions in which /n/-deletion is not applied. To this end, we re-inserted all the /n/s in the transcriptions of the training material and we then train new HMMs. The new HMMs are referred to as ‘/@n#/’ (‘#’= word boundary) whereas the baseline HMMs are referred to as ‘/@-#/’ (‘-‘ = deletion).

Besides the effect that the ‘/@n#/’ HMMs are probably less contaminated, they are also contaminated in a different way. For the ‘/@n#/’ HMMs, the HMM for the phone /n/ will be contaminated with acoustic signal of the /@/, whereas for the ‘/@-#/’ HMMs the HMM for the phone /@/ will be contaminated with acoustic signal of the /n/. This different kind of contamination will probably bias the CSR to choose more transcriptions containing the /n/. As expected, by using the new ‘/@n#/’ HMMs considerably more /n/s are detected by the CSR; 71 more /n/s for the majority vote transcriptions, and 60 more /n/s for the consensus transcriptions (see Appendix 1). Figure 6 shows that the total agreement values increase using the ‘/@n#/’ instead of the ‘/@-#/’ HMMs.

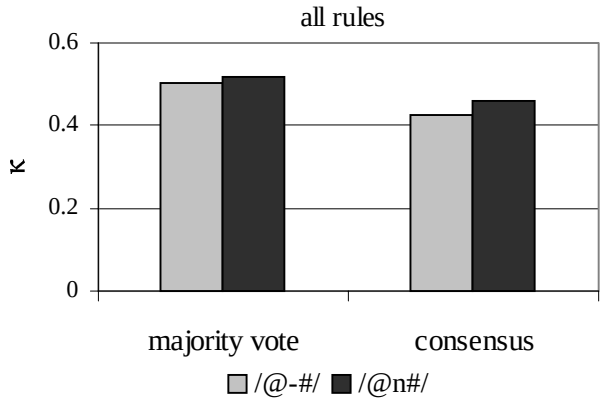

Figure 6: *Total agreement values for '/@-#/' and '/@n#/' HMMs*

Figure 7 shows the agreement values per rule. It can be seen that using the '/@n#/' HMMs instead of '/@-#/' HMMs increases agreement not only for the /n/-deletion rule, but also for some of the other rules. Furthermore, Figure 7 shows that especially for the /n/-deletion rule of the consensus transcriptions, the agreement values are improved. The agreement values for the '/@n#/' HMMs are even higher than for the 'within+cross' HMMs. Since we expect that the amount of contamination in the 'within+cross' HMMs is smaller than that in the '/@n#/' HMMs, one should expect that the agreement values are also higher for the 'within+cross' HMMs. A factor that might partly explain this result is that the '/@-#/' HMMs that were used for automatic transcription of the within- and cross-word pronunciation variation are contaminated. For this reason, the 'within+cross' HMMs that are trained on the basis of these partly incorrect automatic transcriptions are also (indirectly) contaminated.

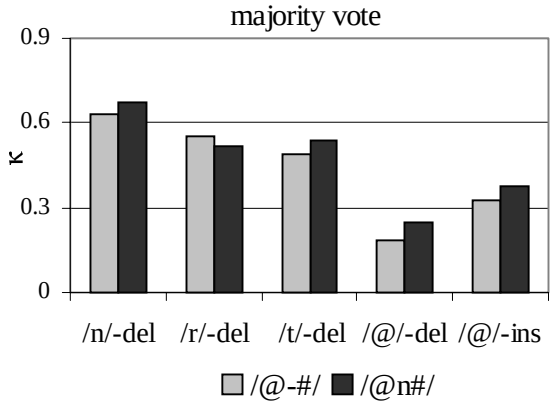

Figure 7a: Agreement values per rule for the majority vote transcriptions

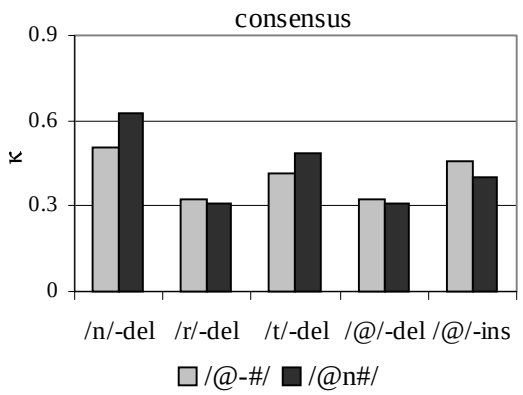

Figure 7b: Agreement values per rule for the consensus transcriptions

3.2.2 Spontaneous versus read speech for model training

It is well known that the amount of pronunciation variation tends to be larger in spontaneous than in read speech. Consequently, fewer mismatches should be found

between the speech signal and the transcriptions in read speech. Thus, it is to be expected that HMMs trained on read speech will be less contaminated than those trained on spontaneous speech. Since in the previous section it was shown that less contaminated HMMs can yield better results, we decided to use HMMs trained on read speech for automatic transcription. The HMMs were trained on 18,000 phonetically rich read sentences of the Dutch Polyphone corpus (den Os, 1995) containing about twice as many words as the VIOS training material.

Figure 8 shows that the total agreement values are higher when we use HMMs trained on read speech (Polyphone) instead of on spontaneous speech (VIOS). The total agreement values are also improved compared to the '/@n#/' HMMs.

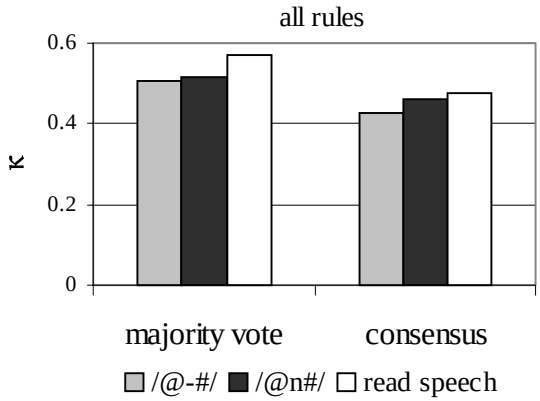

Figure 8: Total agreement values for read speech HMMs

Figure 9 shows that the trends in the results of the read speech HMMs are very similar to those obtained for the pronunciation variation HMMs: Agreement values mainly increase for the /n/-deletion rule and for the /@/-insertion rule. Furthermore, also the read speech HMMs are less biased towards the canonical transcriptions than the baseline HMMs: The percentage of canonical transcriptions for all rules decreases from 57.9% for the baseline HMMs, to 51.0% for read speech HMMs (see Appendix 1). The increase in overall agreement values could be caused by the larger amount of training material used to train the read speech HMMs. However, since the trends for the pronunciation variation HMMs and the read speech HMMs are very similar, the kind of contamination that is contained in the baseline HMMs is probably absent in both the pronunciation variation HMMs and the read speech HMMs.

The results presented in this section show that contamination of the HMMs due to pronunciation variation affects the overall quality of automatic transcription. The quality of automatic transcription can be improved by reducing the amount of contamination. This can be achieved by using pronunciation variation modeling HMMs or HMMs trained on read speech.

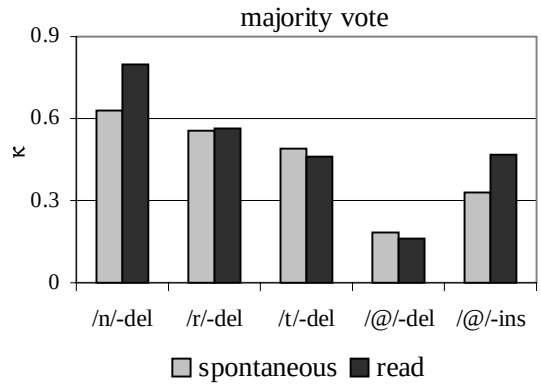

Figure 9a: Agreement values per rule for the majority vote transcriptions

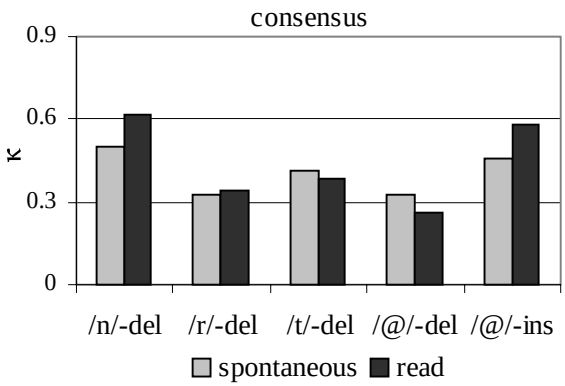

Figure 9b: Agreement values per rule for the consensus transcriptions

3.3 Context-independent vs. context-dependent HMMs

As CD-HMMs take account of the context in which a phone occurs, CD-HMMs are better equipped for modeling context effects such as transitions and co-articulation between phones. For this reason, CD-HMMs generally yield lower WERs (see e.g. Schwartz, Chow, Roucos, Krasner, Makhoul, 1984) and one could expect that CD-HMMs also produce better quality transcriptions. However, we hypothesize that CD-HMMs do not necessarily generate better transcriptions. As mentioned in section 3.2, the effect of contamination of the HMMs on automatic transcription is that the APTs are more biased towards the transcriptions on which the HMMs are trained. This means that if CD-HMMs are trained on the basis of canonical transcriptions of the training material, the CD-HMMs will produce APTs that are biased towards the canonical transcriptions. We hypothesize that the bias towards the canonical transcriptions is sometimes larger for the CD-HMMs than for the CI-HMMs. To illustrate this point, let us consider the following example. Suppose we train CD-HMMs on the basis of transcriptions of the training corpus in which /r/-deletion is not applied. In these transcriptions of the VIOS training corpus 30,018 /r/s are transcribed, of which 1,813 occur in the context /@rd/. However, a large part of these /r/s are not realized since for all words in our material that contain /@rd/, the rule conditions for the /r/-deletion rule are met. According to the human listeners, /r/-deletion is applied in about 1/3 of the cases (see Figure 2), thus of the /r/s in the context /@rd/ about 1/3 are not pronounced. This percentage corresponds to 2% of all /r/s in the training material. Consequently, if a CD-HMM is trained for /@rd/, then the /r/ is not present in 1/3 of the training tokens, which corresponds to 2% of the training tokens for the CI-HMM. This means that the CD-HMM for the context /@rd/ is more contaminated than the CI-HMM for the /r/. For this reason, the bias towards canonical transcriptions is larger for the CD-HMM (for the context /@rd/) than for the CI-HMM for the /r/. As the results in section 3.2 show that removing (part of) the bias towards the canonical transcriptions is beneficial for automatic transcription, we expect that enlarging the bias towards the canonical transcriptions will reduce the agreement values.

In order to investigate the effect of CD-HMMs on automatic transcription, state-tied CD-HMMs are trained on the basis of the canonical transcriptions of our training material. Since our HMMs have a tripartite structure and each of the three parts (or segments) consists of two identical states, state-tying is performed by tying segments. For state-tying it is assumed that all first segments are dependent on the left context of the phone, all middle segments are independent of the context, and all last segments are dependent on the right context. For this reason, all middle segments of each phone are clustered to train a CI-model for all middle segments of the same phone. Left and right CD-models are trained for clusters of first and last segments with equal left or right contexts. Each cluster of first and last segments contains at least 200 observations. All left and right contexts with less than 200 observations are clustered to train two backing off models: one for all first and one for all last segments with less than 200 observations. In total, 237 left CD-models and 227 right CD-models are trained. If we then look at the training corpus consisting of 326,494 phones, and thus 326,494 left and right contexts, we see that 94.3% and 94.4% of these contexts are covered by the right and left CD-models, respectively.

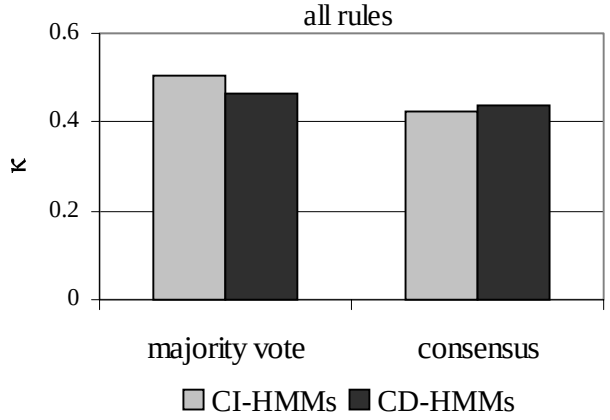

Figure 10: Agreement values for CI- versus CD-HMMs

Figure 10 shows that the agreement values are lower if CD-HMMs are used to obtain automatic transcriptions of the majority vote material, whereas for the consensus transcriptions a small improvement in the total agreement values is found. Figure 11 shows that agreement increases for some of the rules (viz. the /t/- and /@/-deletion rules), but decreases for others (viz. the /r/-deletion and /@/-insertion rules). Especially for the /r/-deletion rule of the majority vote material a large decrease in agreement is found. Due to this large decrease in agreement for the /r/-deletion rule, the total agreement value decreases for the majority vote transcriptions. If we look at the numbers of detected /r/s using CD-HMMs, it is striking that this number is extremely large for the majority vote transcriptions (see Appendix 1). The large deterioration in agreement for the /r/-deletion rule is indeed mainly caused by a large

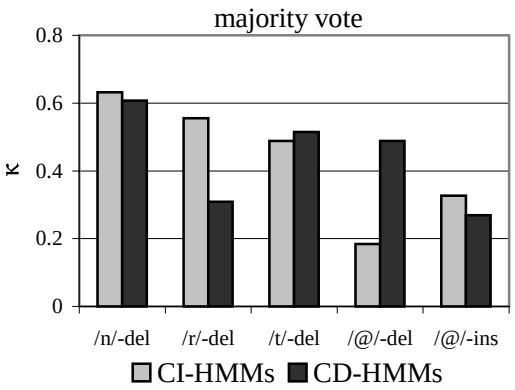

Figure 11a: Agreement values per rule for the majority vote transcriptions

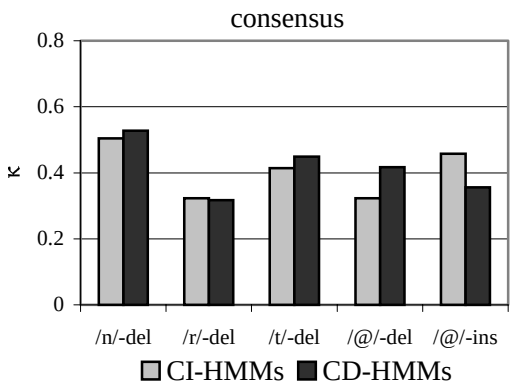

Figure 11b: Agreement values per rule for the consensus transcriptions

bias towards the canonical transcriptions (i.e. the transcription in which the /r/ is present): For the majority vote material, more /r/s are unjustly denoted as present by the CSR using CD-HMMs instead of CI-HMMs. The different /r/-deletion results for the majority vote and the consensus material is probably related to the fact the /r/-deletion rule is the only rule for which there exist a considerably difference in the identity and frequency of the words that are contained in the two types of material. Therefore, for obtaining the automatic transcription of /r/-deletion, CD-HMMs are used that concern other contexts. Probably, the amount of contamination in these different contexts varies and thus, also agreement varies.

Another observation that can be made from Figure 11 is that there is a large increase in agreement values for the /@/-deletion rule. Closer inspection of the /@/-deletion transcriptions reveal that this increase is caused by an increase in the number of detected /@/s: Nearly all extra /@/s that are now detected by the CSR concern /@/s that are denoted as present by the human transcribers.

To conclude, using CD-HMMs for automatic transcription causes a deterioration in the total agreement for the majority vote transcriptions, whereas a small improvement is found for the consensus transcriptions. The deterioration in agreement values for the majority vote transcriptions is mainly caused by an increase in the number of /r/s that are unjustly detected by the CSR. The difference in result for the two sets of materials with respect to the /r/-deletion rule can probably by explained by the fact that the identity and frequency of the words that are contained in the two types of material are different for this rule.

3.4 Combinations of properties

In this section, we will investigate the effect of two combinations of properties, on the assumption that some properties will be (partly) complementary in terms of their ability to improve automatic transcription quality.

3.4.1 Combination of pronunciation variation modeling and a short /@/ HMM

First, we investigate a combination of using a shorter topology length for the phone /@/ (see section 3.1) and pronunciation variation modeling (see section 3.2). It can be expected that these properties benefit from each other. On the one hand, pronunciation variation modeling removes the /@/-transcriptions for the /@/s that are not pronounced, thus making the short /@/ HMMs less contaminated. On the other hand, the short /@/ HMM will probably make better automatic transcriptions of the within- and cross-word pronunciation variation (recall that the agreement values are higher using the short /@/ HMM).

In order to train the combination HMMs, we make a new transcription of the within- and cross-word variation in the training material using the set of HMMs that contains the short /@/ HMM with the highest total agreement values (the short /@/ HMM consisting of 2 segments). Next, the new transcriptions are used to train a new set of HMMs.

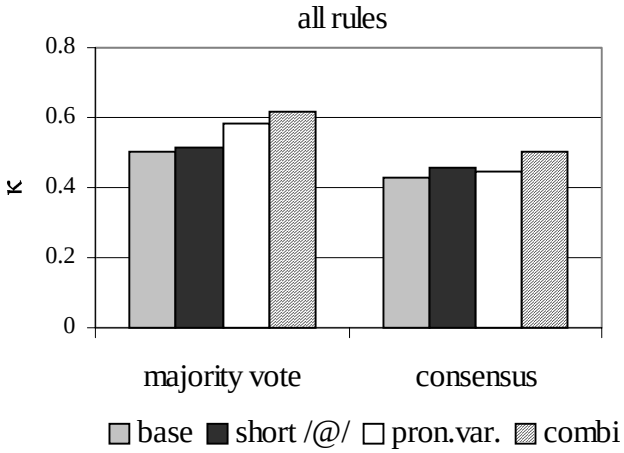

Figure 12: Total agreement for the combination of pronunciation variation modeling and short /@/ HMMs

Figure 12 shows the total agreement values for the baseline HMMs (‘base’), the agreement values for changing the separate properties (‘short /@/’ and ‘pron.var.’), and the agreement values for changing the two properties simultaneously (‘combi’). It can be seen that the combination of the two properties results in higher agreement values than each property separately.

Figure 13 shows that - compared to the baseline - agreement is largely improved for the /n/-deletion rule and the /@/-insertion rule. This increase in agreement for the /n/-deletion rule can be attributed to the pronunciation variation modeling since the increase in agreement was also found for the pronunciation variation HMMs. The two rules that especially benefit from the combination of the two properties are the /@/-deletion and /@/-insertion rules: For both rules the combination results are better than the results of the individual properties.

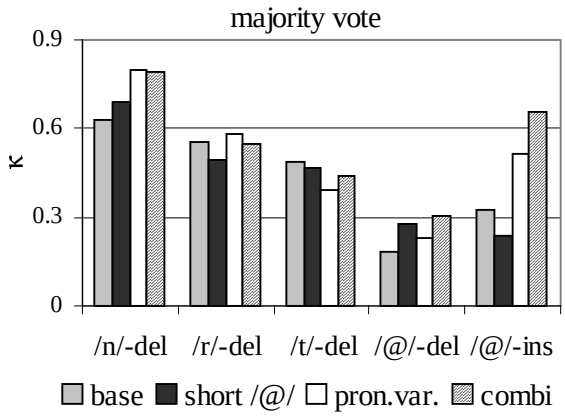

Figure 13a: Agreement values per rule for the majority vote transcriptions

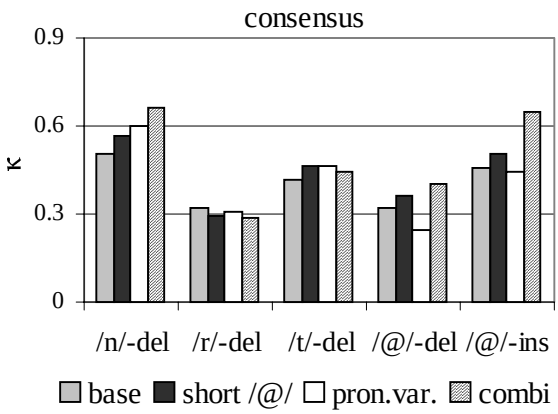

Figure 13b: Agreement values per rule for the consensus transcriptions

3.4.2 Combination of pronunciation variation modeling and CD-HMMs

Another combination of properties that could enhance the system's transcription quality is pronunciation variation modeling (section 3.2) and CD-HMMs (section 3.3). Due to modeling of pronunciation variation, part of the mismatch between the phonetic transcriptions of the training material and the actual pronunciation is removed, thus the CD-HMMs are less contaminated.

In order to train the combination HMMs, we make an automatic transcription of the within- and cross-word variation in the training material using the baseline HMMs. On the basis of this transcription, state-tied CD-HMMs are trained (see section 3.3 for more details on the state-tying procedure). Figure 14 shows the agreement values for the pronunciation variation HMMs ('pron.var.') and CD-HMMs ('CD') and the combination of pronunciation variation modeling and CD-HMMs ('combi'). In general, the combination of the two properties results in higher total agreement values than the agreement values for each property separately.

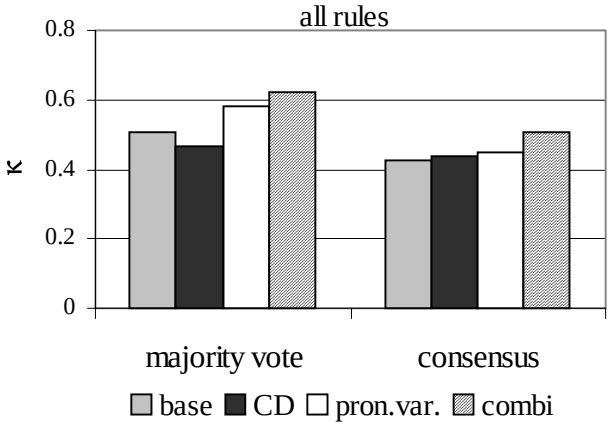

Figure 14: Total agreement values for combination of pronunciation variation and CD-HMMs

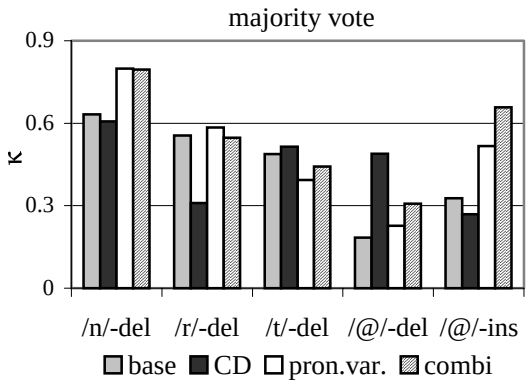

Figure 15a: Agreement values per rule for the majority vote transcriptions

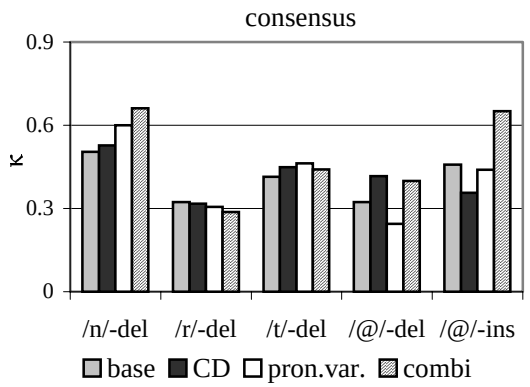

Figure 15b: Agreement values per rule for the consensus transcriptions

Figure 15 reveals that CD-HMMs can indeed benefit from pronunciation variation modeling: The large decreases in agreement values that were found for the majority vote transcriptions of the /r/-deletion rule disappear. Furthermore, for the /@/-insertion rule the combination results are better than the individual results. Clearly, context-dependent modeling improves upon pronunciation variation modeling. The /@/-deletion rule is the only rule for which the combination results substantially deteriorates compared to the results of each individual property. This result might be explained as follows: As the automatic transcriptions of the /@/-deletion variants are obtained with the baseline HMMs and as low agreement values are found for the /@/-deletion rule (the κ values are qualified as 'slight' and 'fair' for the baseline HMMs), the pronunciation variation modeling might deteriorate the context-dependent modeling.

4 Agreement and WER

In other research on APT, the choice of the speech recognizer is usually not clearly motivated. Most probably, one generally takes the speech recognizer with the lowest WER. Obviously, the assumption on which this choice rests is that a recognizer with a lower WER will produce better APTs. To investigate whether a recognizer with lower WERs indeed produces better quality transcriptions, we looked at the relation between WER on the one hand, and the agreement values between the APTs and human RTs on the other hand. We measured WER on the total transcription material (majority vote + consensus) for all sets of HMMs that are used in this article. The lexicon used in the recognition experiments contains 1,154 words, to which 1,119 pronunciation variants were added. The variants were automatically generated by applying the five phonological rules (see section 2.1) to the canonical transcriptions of the words, thus obtaining a lexicon containing 2273 entries. A language model is employed that distinguishes between different variants of the same word. For more details on this kind of language model, see Kessens et al. (1999). For more details on the CSR, see section 2.2. The WER is defined as follows:

$$\mathrm{WER} = \frac{\mathrm{S}+\mathrm{D}+\mathrm{I}}{\mathrm{N}} \mathrm{x}100\% \qquad (3)$$

where S is the number of substitutions, D the number of deletions, I the number of insertions, and N the total number of words. As a measure of agreement we used the total κ, which is the κ for the two data sets pooled together. In Figure 16, the scatter plot of the total κ as a function of WER is given.

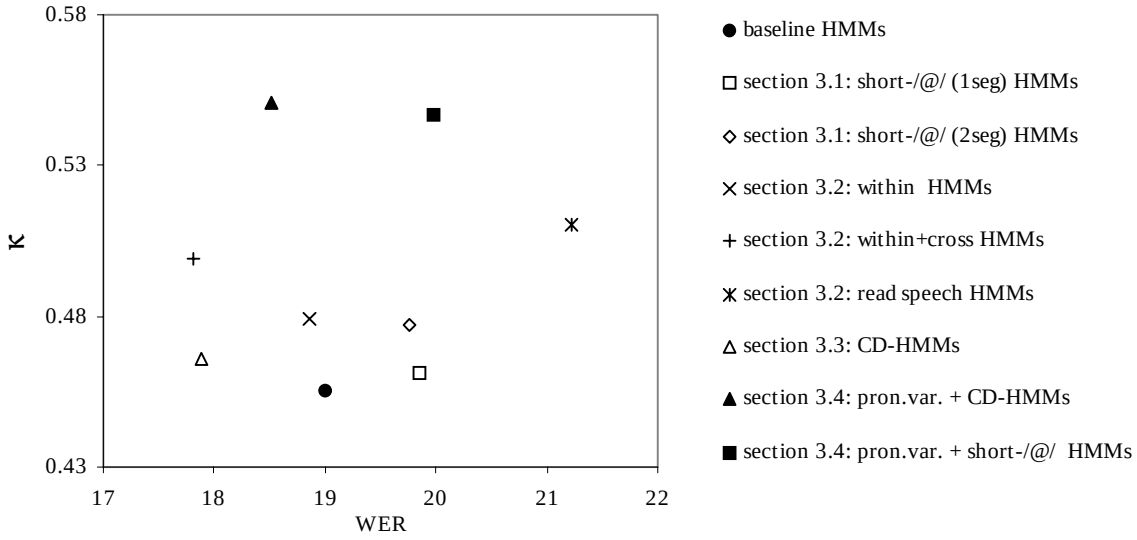

Figure 16: Scatter plot of total κ and WER on transcription data

The relation between κ and WER that we would expect is: The lower the WER, the higher κ. Figure 3 shows that this trend is not present. If we had selected the HMMs with the lowest WER ('within + cross HMMs') for automatic transcription, we would not have obtained the most optimal APTs. Furthermore, the HMMs that produce the optimal APTs (combination of pronunciation variation modeling and CD-HMMs) do not yield the lowest WERs. One could argue that it is not correct to use the total κ as a measure of agreement, since the agreement values are different for the two data sets. However, even when the two sets of data sets are treated separately, the expected relation between WER and κ is not found either. For both data sets the best agreement values are found for the combination of pronunciation variation modeling and CD-HMMs, whereas the lowest WERs are found for the pronunciation variation HMMs.

Saraçlar (2000a) and Saraçlar et al. (2000b) also reported results showing that a better transcription accuracy does not imply that the WER is also improved. They found that HMMs trained on automatic transcriptions of pronunciation variation improve transcription accuracy by 4.5% compared to using baseline HMMs, whereas the WER deteriorates by 1.4%. They conclude that this result can be explained by an increased lexical confusion: "Since our decision tree pronunciation model allows words to have a large number of pronunciations, many of which overlap with pronunciations of other words, 'recovering' the right word strings from more accurate phone recognition is difficult". We think that also another factor also plays a role. The sequences of phones that can be recognized during a normal recognition task are

constrained by the lexicon and the (word-level) language model. Through these constraints, it is impossible or less likely to recognize some sequences of phones during a conventional recognition task. During automatic transcription, however, the lexicon and the language model do not influence the resulting transcription and thus, these constraints do not influence transcription quality.

These results illustrate that recognition and automatic transcription are different tasks and should be optimized in different ways. For this reason, for automatic transcription one should not select the speech recognizer with the highest recognition performance in a conventional recognition task, but one should rather concentrate on the properties that the recognizer should have for making optimal APTs.

5. Discussion

In this paper, we evaluated the quality of APTs by measuring agreement with two kinds of human RTs, namely RTs based on a majority vote principle and consensus RTs. The agreement values for the consensus transcriptions were, in general, lower than the agreement values for the majority vote transcriptions. There are various possible explanations for the difference in absolute agreement values. First of all, the experience level of the students who made the consensus transcribers is probably lower than that of the majority vote transcribers. It seems reasonable to assume that linguists will be much more aware of the five phonological processes that were the focus of this study. As a consequence they may be more attentive to details that might be ignored by the students. Second, as the number of subjects involved in making the consensus transcriptions is smaller than the number of transcribers that made the majority vote transcriptions, the consensus transcriptions are probably less accurate. Third, the consensus transcribers were not specifically instructed to decide whether one of the five optional phonological rules under investigation was (or was not) applied in specific words in the utterance, whereas the majority vote transcribers were aware of the purpose of the investigation. Through this difference in focus, the reliability of the majority vote transcriptions might also be improved.

Although the absolute agreement values vary for the two types of human RTs, the general trends that we observe are very similar. There are two exceptions. First of all, for the /@/-insertion rule, the majority vote and consensus transcriptions reveal contradictory trends: Using a short /@/ HMM, agreement values deteriorate for the majority vote material, whereas agreement values increase for the consensus material. Furthermore, pronunciation variation HMMs lead to lower agreement values for the consensus transcriptions, whereas it is the other way around for the majority vote transcriptions. A possible explanation for the opposite results is that the majority vote transcribers are more biased towards /@/-insertion since they expect this process to occur. Since the consensus transcribers are probably less familiar with the /@/-insertion rule, they are probable less biased. Second, we found contradictory results for the CD-HMMs: Compared to using CI-HMMs, agreement values are lower for the majority vote material whereas the agreement values are higher for the consensus material. Probably, this difference is not related to the way the manual transcriptions

are made, but is probably due to the fact that the two types of material contain different words for the /r/-deletion rule.

The results that we obtained in this investigation are in concordance with the results reported by other authors. We observed that a topology length of 20 ms for the phone /@/ results in better quality APTs than the baseline topology length of 30 ms, which confirms the results reported by Brugnara et al. (1993). Furthermore, we found that pronunciation variation HMMs yield better quality APTs, a result that has also been reported by Saraçlar (2000a) and Saraçlar et al. (2000b). These authors also showed that the pronunciation variation HMMs were less biased towards canonical transcriptions than the baseline HMMs; this trend was also observed in our data.

Closer inspection of our data reveals that the results vary per rule. For this reason, we now discuss the results per rule. For the /n/-deletion rule, the baseline HMMs seem to contain contamination which results in a bias towards the deletion of /n/. This contamination can be reduced by training HMMs on the basis of a transcription of the training material in which the /n/-deletion rule is not applied since according to the humans /n/-deletion is applied in less than half of the cases. Furthermore, the contamination can be reduced by training HMMs on automatic transcriptions of pronunciation variation. Finally, by training the HMMs on read speech material the amount of contamination in the HMMs is also reduced. For the /r/-deletion rule, the most striking result is that agreement is considerably reduced by context-dependent modeling. This is mainly caused by a large bias of the CD-HMMs towards the canonical transcriptions. Consequently, more /r/s are unjustly detected using the CD-HMMs compared to using the CI-HMMs. The deterioration in agreement for the CD-HMMs can be reduced if the context-dependent modeling is combined with pronunciation variation modeling. For the /t/-deletion rule the only clear trend is that context-dependent pronunciation modeling seems to give the highest agreement values. For the /@/-deletion rule, the discrepancy between the number of detected phones by humans and CSR seems to be partly of a durational nature, since using a short /@/ HMM improves agreement values. The /@/-deletion rule is the only rule for which the combination of pronunciation modeling and CD-HMMs results in considerably lower agreement values than the agreement values for CD-HMMs without pronunciation variation modeling. A possible explanation for this result is that the quality of the automatic transcriptions of the /@/-deletion variants is low as they are obtained with the baseline HMMs (the κ values are qualified as ‘slight’ and ‘fair’ for the baseline HMMs). Finally, for the /@/-insertion rule we find differences in results for the majority vote and consensus transcriptions. In general, the number of phones that are denoted as present is higher for the humans than for the CSR. For the majority vote transcriptions this difference in detected phones becomes smaller if the amount of contamination contained in the HMMs is reduced, whereas for the consensus transcriptions, the topology length of the /@/ HMM seems to play a role. The fact that we find discrepancies in the results of the majority vote and consensus transcriptions of the /@/-insertion rule probably means that the way in which the humans decide on the application of this rule is different in the two transcription tasks. A way of limiting the discrepancy between the number of detected phones by CSR

and humans and increasing agreement for both materials is to use a combination of pronunciation variation modeling and a short /@/ HMM.

6 Conclusions

In this study, we have shown that changing the properties of a CSR does influence the degree of agreement between the automatic transcriptions and the reference transcription: For the majority vote transcriptions, the overall κ-value varies between 0.464 and 0.629. For the consensus transcriptions, the overall κ-value varies between 0.426 and 0.505. Although the absolute agreement values for the two kinds of human RTs differ, the general trends are very similar. Our results indicate that the quality of the automatic transcriptions can be improved by using 'short' HMMs. The quality of automatic transcription can also be improved by reducing the amount of contamination due to pronunciation variation. This can be achieved by using HMMs trained on the most frequently observed transcription, by using HMMs trained on automatic transcriptions of pronunciation variation, or by using HMMs trained on read speech. Furthermore, we found that CD-HMMs should not be trained on the baseline transcriptions, since for these transcriptions there is a mismatch between the phonetic transcriptions of the speech material and the realized pronunciation. If CD-HMMs are trained on automatic transcriptions of pronunciation variation, the mismatch is reduced, resulting in better quality transcriptions. The combination of two other properties, namely pronunciation variation modeling and 'short' HMMs, results in higher agreement values than those obtained with the individual properties. Finally, we found that by combining properties the quality of automatic transcription can be improved even further. For both data sets, the lowest total agreement values are obtained for the baseline HMMs, whereas the highest values are obtained for a combination of pronunciation variation modeling and CD-HMMs.

Finally, we observed that there is no clear relation between the WER of a CSR and the κ-values. Therefore, we can conclude that for obtaining automatic transcriptions, using the CSR with the lowest WER is not always the optimal solution. It appears that for this specific purpose, CSRs should be used that have been specially optimized for automatic transcription.

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). Grateful appreciation is extended to Loe Boves who gave useful comments on previous versions of this paper. Furthermore, we kindly thank Catia Cucchiarini for her useful comments on the parts of this paper concerning manual phonetic transcription. Finally, we would like to thank several members of the research group *A²RT* for their useful comments on a previous version of this paper.

References

Adda-Decker, M. & Lamel, L. (1998). Pronunciation variants across systems, languages and speaking style. In *Proceedings of the workshop on modeling pronunciation variation for ASR, Rolduc*, pp. 131-136.

Brugnara, F., Falavigna, D. & Omologo, M. (1993). Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication*, **12**, 357-370.

Cucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study*. Ph.D. thesis, University of Nijmegen, Nijmegen, The Netherlands.

Cohen, J.A. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213-220.

Cohen, M. H. (1989). *Phonological Structures for Speech Recognition*. Ph. D. thesis, University of California, Berkeley.

Cox, S., Brady, R. & Jackson, P. (1998). Techniques for accurate annotation of speech waveforms. In *Proceedings of ICSLP'98*, pp. 1947-1950.

Kerkhoff, J. & Rietveld, T. (1994). Prosody in Niros with Fonpars and Alfeios. In *Proceedings of the Department of Language & Speech, Univ. 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 variation. *Speech Communication*, **29**, 193-207.

Kessens, J.M. & Strik, H. (2001). A data-driven method for modeling pronunciation variation, *submitted to Speech Communication*.

Kuipers, C. & Donselaar, W. van (1997). The influence of rhythmic context on schwa epenthesis and schwa deletion in Dutch. *Language and Speech*, **41 (1)**, 87-108.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

Ljolje, A., Hirschberg, J. & van Santen, J. P. H. (1997). Automatic Speech Segmentation for Concatenative Inventory Selection. *Progress in Speech Synthesis* (J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg eds.), pp. 305-311, Springer-Verlag, New York.

den Os, E.A., Boogaart, T.I., Boves, L. & Klabbers, E. (1995). The Dutch Polyphone Corpus. In *Proceedings of Eurospeech'95*, pp. 825-828.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H. , Saraçlar, M., Wooters, C. & Zavaliagkos, G. (1998). Stochastic Pronunciation Modelling from Hand-labelled Phonetic Corpora. Speech Communication, **29**, 209-224

Saraçlar, M. (2000a). Pronunciation modeling for conversational speech. P.h.D. thesis, John Hopkins University, Baltimore, Maryland.

Saraçlar, M., Nock, H., Khudanpur, S. (2000b). Pronunciation modeling by sharing Gaussian densities across phonetic models. In *Computer, Speech & Language*, **14**, 137-160.

Schwartz, R., Chow, Y., Roucos, S., Krasner, M. & Makhoul, J. (1984). Improved hidden Markow modeling of phonemes for continuous speech recognition. In *Proceedings of ICASSP'84*, pp. 35.6.1-35.6.4.

Shriberg, L.D. & Lof, L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics*, **5**, 225-279.

Shriberg, L.D., Kwiatkowski, J. & Hoffman, K. (1984). A Procedure for Phonetic Transcription by Consensus. *Journal of Speech and Hearing Research*, **27**, 456-465.

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 *Proceeding of Eurospeech'97*, pp. 2125-2128.

Strik, H., Russel, A.J.M., van den Heuvel, H. Cucchiarini, C. & Boves, L. (1997). A spoken dialog system for the Dutch public transport information service. *International Journal of Speech Technology*, **2-2**, 119-129.

Strik, H. & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, **29**, 225-246.

van de Velde, H. (1996). *Variatie en verandering in het gesproken Standaard-Nederlands (1935-1993)*. Ph. D. thesis, University of Nijmegen, Nijmegen, The Netherlands.

Wester, M., Kessens, J.M, & Strik, H. (1998). Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation, In *Proceedings of the workshop Modeling Pronunciation Variation for ASR, Rolduc*, pp. 145-150.

Wester, M., Kessens, J.M., Cucchiarini, C. & Strik, H. (2001). Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer. *Accepted for publication in Language & Speech.*

Appendix 1: *Numbers APTs in which the relevant phone is 'present'. For the /r/-, /t/- and /@/-deletion rules the APTs in which the relevant phone are 'present' are the canonical transcriptions. For the /n/-deletion rule and /@/-insertion rule, the APTs in which the relevant phones are 'present' are the non-canonical transcriptions. Therefore, for the /n/-deletion rule and /@/-insertion rule the numbers of canonical transcriptions are given between brackets. In the last row, the numbers are given for the human RTs.*

Table A1.1: Numbers of 'phone present' for the majority vote material.

section	HMMs	/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins	all
3.1	3seg (baseline)	72 (83)	66	64	17	14 (34)	233 (264)
3.1	2seg	77 (78)	73	63	30	16 (32)	259 (276)
3.1	1seg	79 (76)	83	66	36	18 (30)	282 (291)
3.2	within	90 (65)	58	62	16	18 (30)	244 (231)
3.2	within+cross	90 (65)	57	63	19	17 (31)	246 (235)
3.2	/@n#/	110 (45)	68	66	20	11 (37)	275 (236)
3.2	read speech	96 (59)	69	63	19	20 (28)	267 (238)
3.3	CD	83 (72)	107	71	29	8 (40)	298 (319)
3.4	pron.var. & short /@/	96 (59)	61	62	31	29 (19)	279 (232)
3.4	pron.var & CD	95 (60)	67	66	19	19 (29)	266 (241)
human RTs		95 (60)	78	70	38	27 (21)	308 (267)

Table A1.2: Numbers of 'phone present' for the consensus material.

section	HMMs	/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins	all
3.1	3seg (baseline)	121 (166)	123	64	20	24 (79)	352 (452)
3.1	2seg	136 (151)	129	62	21	33 (70)	381 (433)
3.1	1seg	137 (150)	148	64	25	34 (69)	408 (456)
3.2	within	146 (141)	109	57	18	33 (70)	363 (395)
3.2	within+cross	148 (139)	110	59	15	32 (71)	364 (394)
3.2	/@n#/	181 (106)	131	66	17	27 (76)	422 (396)
3.2	read speech	164 (123)	120	62	21	36 (67)	403 (393)
3.3	CD	150 (137)	172	72	25	15 (88)	434 (494)
3.4	pron.var. & short /@/	151 (136)	123	63	22	41 (62)	400 (406)
3.4	pron.var & CD	158 (129)	124	69	19	35 (68)	405 (409)
human RTs		181 (106)	155	83	28	38 (65)	485 (437)

Appendix 2: *Agreement values for all sets of HMMs*

Table A2.1: Agreement values for majority vote material

section	HMMs	/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins	all
3.1	3seg (baseline)	0.632	0.555	0.488	0.184	0.327	0.504
3.1	2seg	0.691	0.495	0.464	0.279	0.240	0.517
3.1	1seg	0.689	0.509	0.462	0.285	0.152	0.518
3.2	within	0.799	0.568	0.372	0.163	0.475	0.559
3.2	within+cross	0.799	0.585	0.393	0.227	0.517	0.584
3.2	/@n#/	0.672	0.520	0.538	0.250	0.376	0.518
3.2	read speech	0.795	0.566	0.464	0.159	0.469	0.571
3.3	CD	0.607	0.309	0.515	0.488	0.269	0.464
3.4	pron.var. & short /@/	0.795	0.547	0.442	0.307	0.658	0.619
3.4	pron.var. & CD	0.810	0.633	0.538	0.294	0.517	0.629

Table A2.2: Agreement values for consensus material

section	HMMs	/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins	all
3.1	3seg (baseline)	0.504	0.323	0.414	0.323	0.458	0.426
3.1	2seg	0.567	0.292	0.466	0.360	0.507	0.455
3.1	1seg	0.545	0.256	0.374	0.310	0.531	0.429
3.2	within	0.573	0.300	0.398	0.253	0.464	0.432
3.2	within+cross	0.600	0.306	0.463	0.245	0.440	0.447
3.2	/@n#/	0.626	0.306	0.484	0.311	0.401	0.459
3.2	read speech	0.616	0.338	0.387	0.262	0.578	0.474
3.3	CD	0.528	0.317	0.449	0.417	0.356	0.438
3.4	pron.var. & short /@/	0.661	0.287	0.441	0.399	0.651	0.504
3.4	pron.var. & CD	0.664	0.312	0.488	0.192	0.640	0.505

Article 3

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

ELSEVIER

Speech Communication 29 (1999) 193–207

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

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

Judith M. Kessens *, Mirjam Wester, Helmer Strik

A²RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

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

Abstract

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

Zusammenfassung

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

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

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

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

Résumé

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

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

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

2. Method

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

2.1. Baseline

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

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

2.2. General procedure

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

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

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

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

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

2.3. Method for modeling within-word pronunciation variation

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

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

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

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

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

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

reizen /rEiz@n/ → /rEiz@/

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

Amsterdam /Amst@rdAm/ → /Amst@dAm/

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

stressed: *Arnhem* /ARnEm/ → /AnEm/

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

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

Haarlem /ha:RlEm/ → /ha:lEm/

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

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

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

'*s avonds* /sa:vOnts/ → /sa:vOns/

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

Utrecht /ytrExt/ → /ytrEx/

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

latere /la:t@r@/ → /la:tr@/

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

Delft /dELft/ → /dEl@ft/

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

2.4. Modeling cross-word pronunciation variation

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

2.4.1. Method 1 for modeling cross-word pronunciation variation

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

2.4.2. Method 2 for modeling cross-word pronunciation variation

The second method which we adopted for modeling cross-word variation was to make use of multi-words. Multi-words are word sequences which are joined together and added as separate entities to the lexicon. In order to be able to compare the results of this method to the results of the previous one, the same cross-word processes were modeled in both methods. On the basis of the seven words from cross-word method 1, multi-words were selected from the list of 50 word sequences. Only those word sequences in which at least one of the seven words was present could be chosen. Thus, 22 multi-words were selected. Subsequently, these multi-words were added to the lexicon and the language model. It was necessary for us to also add the multi-words to the language model, because effectively, for our CSR they are "new" words. Next, the cross-word variants of the multi-words were also added to the lexicon, and the remaining steps of the general procedure were carried out (see Section 2.2).

All of the selected multi-words have at least two pronunciations. If the parts of the multi-words are counted as separate words, the total number of words which could have a pronunciation variant covers 6% of the total number of words in the training corpus. This percentage is lower than that for cross-word method 1 due to the contextual constraints imposed by the multi-words. Table 3 shows the multi-words (column 1), the total number of times the multi-word occurs in the training material (column 2), their baseline transcriptions (column 3) and their added cross-word variants (column 4).

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

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

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

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

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

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

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

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

3. CSR and material

3.1. CSR

The main characteristics of the CSR are as follows. The input signals consist of 8 kHz, 8 bit A-law coded samples. Feature extraction is done every 10 ms for 16 ms frames. The first step in feature analysis is an FFT analysis to calculate the spectrum. In the following step, the energy in 14 mel-scaled filter bands between 350 and 3400 Hz is calculated. Next, a discrete cosine transformation is applied to the log filterband coefficients. The final processing stage is a running cepstral mean substraction. Besides 14 cepstral coefficients ($c_0 - c_{13}$), 14 delta coefficients are also used. This makes a total of 28 feature coefficients.

The CSR uses acoustic models, word-based language models (unigram and bigram) and a lexicon. The acoustic models are continuous density hidden Markov models (HMMs) with 32 Gaussians per state. The topology of the HMMs is as follows: each HMM consists of six states, three parts of two identical states, one of which can be skipped (Steinbiss et al., 1993). In total, 39 HMMs were trained. For each of the phonemes /l/ and /r/, two models were trained, because a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/). For each of the other 33 phonemes context-independent models were trained. In addition, one model was trained for non-speech sounds and a model consisting of only one state was employed to model silence.

3.2. Material

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

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

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

The baseline test lexicon contains 1154 entries, which are all the words in the test corpus, plus a number of words which must be in the lexicon because they are part of the domain of the application, e.g. station names. The test corpus does not contain any out-of-vocabulary words. This is a somewhat artificial situation, but we did not want the CSR's performance to be influenced by words which could never be recognized correctly, simply because they were not present in the lexicon. Adding pronunciation variants generated by the five phonological rules (within-word method) leads to a lexicon with 2273 entries (also an average of about 2 entries per word). For cross-word methods 1 and 2, the same variants were added to the test lexicon as those which were added to the training lexicon.

4. Results

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

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

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

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

4.1. Modeling within-word pronunciation variation

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

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

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

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

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

4.2. Modeling cross-word pronunciation variation

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

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

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

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

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

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

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

4.4. Comparing methods in isolation and in combination

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

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

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

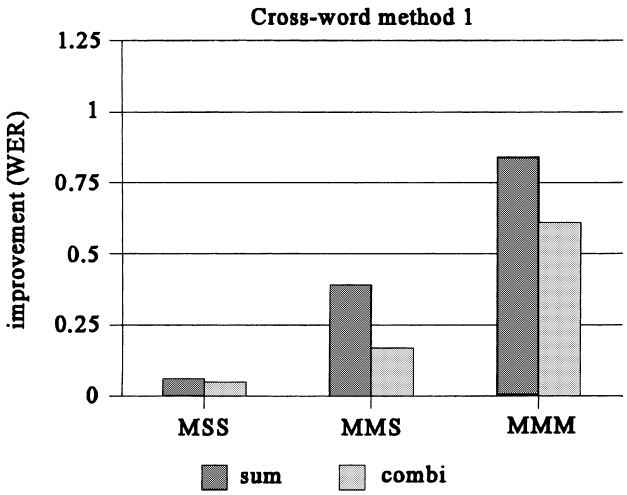

Fig. 1. Improvements (WER) for cross-word method 1 combined with the within-word method and the sum of the two methods in isolation.

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

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

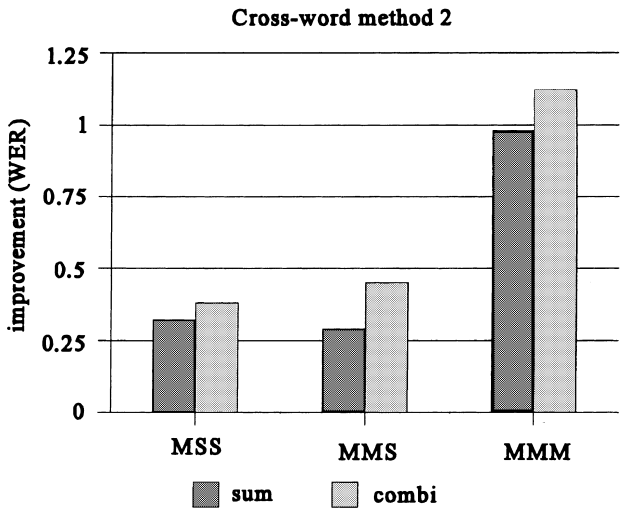

Fig. 2. Improvements (WER) for cross-word method 2 combined with the within-word method and the sum of the two methods in isolation.

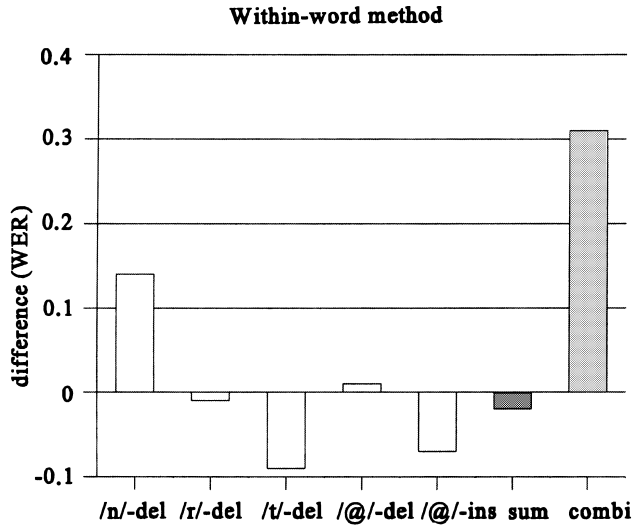

Fig. 3. Difference in WER between the baseline result and results of adding variants of separate rules to the lexicon, sum of those results, and combination result of all rules.

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

4.5. Overall results

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

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

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

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

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

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

* Significant improvements.

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

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

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

Type of change	Number of utterances
Same utterance, different mistake	474 (7.6%)
Improvements	267 (4.3%)
Deteriorations	183 (2.9%)
Net result	+84 (1.3%)

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

5. Discussion

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

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

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

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

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

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

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

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

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

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Article 4

J. M. Kessens, C. Cucchiarini and H. Strik. A data-driven method for modeling ponunciation variation, submitted to Speech Communication.

Abstract

This paper describes a rule-based data-driven (DD) method to modeling 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 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, also deteriorations are 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 8.2% relatively, is found.

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 (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 is 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 to a small number of variants. For example, in our knowledge-based approach to modeling pronunciation variation, maximally 16 variants per 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 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 data-driven 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.

The aim of the present paper is threefold. First, we analyse whether the DD method of modeling pronunciation variation that we have adopted leads to a reduction in the WER. Second, 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 an error analysis at word level. The goal of this error analysis is to determine how the changes in WER came about. It should be noted that this kind of analysis is seldom done in pronunciation variation modeling research (but Ravishankar and Eskenazi, 1997; Kessens et al., 1999; and Wester et al., 2000b), despite its indisputable importance for understanding what is really going on. However, limitations of these error analyses are

that they are performed manually, with the consequence that only limited numbers of variants/rules can be analyzed. Since the present error analysis is performed automatically, much larger amounts of material can be analysed. The third goal of this paper is examine the adequacy of three criteria for rule selection. In this way 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, 4 and 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).

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%

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. In the baseline system, for each word, one transcription is present in the lexicon. This so-called *'canonical transcription'* was obtained using a Text-to-Speech system (TTS) 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.

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 the 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. As was mentioned in the introduction, in this research we limited ourselves by looking only at deletions of phones, and thus only deletion rules were obtained. The following five steps describe the whole procedure of automatic extraction of the candidate rules:

1. For each word in an utterance, the so-called *'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 which pronunciation variant best matches the acoustic signal. In this way, data-driven 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} | v @ R b I n d I N | Y t r E x t | ('|' = word boundary)
T_{dd} | 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:

   /L F R/$_{can}$ → /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 in the literature as '*rewrite rules*', 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.

6. For each candidate rule, we calculate three frequency measures:
   - F_{cond}: the number of times the condition for the rule (/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} ($0 \leq F_{rel} \leq 1$).

3.1.2 Motivations for performing rule selection

Before using the rules in order to generate variants for the recognition experiments, we made a selection on the set of 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 the addition of variants. The large increase 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 the decoding time, since decoding time is directly related to the size of the lexicon. Third, in data-driven approaches, the data-derived variants are usually selected or filtered, as the variants might be based on artefacts of the CSR (e.g. contamination of the acoustic models) instead of 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 may 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), confusability measures (Wester and Fossler-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 (1998) and Lehtinen et al (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 $F_{abs}>100$. This was done for various reasons. First, the data-driven 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 2,951 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 ($F_{abs}>100$) 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 2,951 rules are selected. In Appendix 1, 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 rules 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_{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 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.).

Table 2: Statistics of the multiple pronunciation lexica

	Frel >	# 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

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:

$$\mathrm{WER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}} \qquad (1)$$

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 Figure 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. Figure 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 lower than the WER for the baseline testing condition. The reduction in WER is significant (t-test, α=0,05) for 1.25 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.

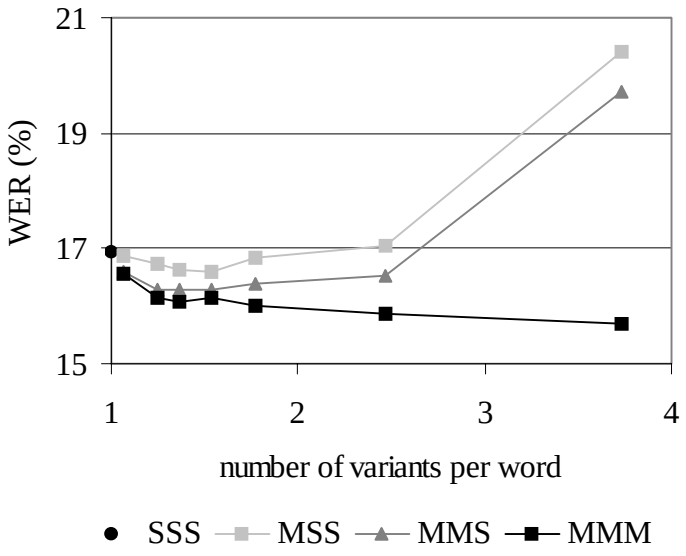

Figure 1: WERs for the different testing conditions

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 per 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 per 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 the 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 this method is 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 will discuss the results and summarize our conclusions.

4.1.1 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:

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=incorrect), as *'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 applied, the count was equally distributed over the rules: If N rules applied to the recognized word, each of these rules was assigned a score of $\frac{1}{N}$.

Table 3: 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
5	Elst	Delft<dELft>	Ede<e:d@>	different error	-	-

4.1.2 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 Figure 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.

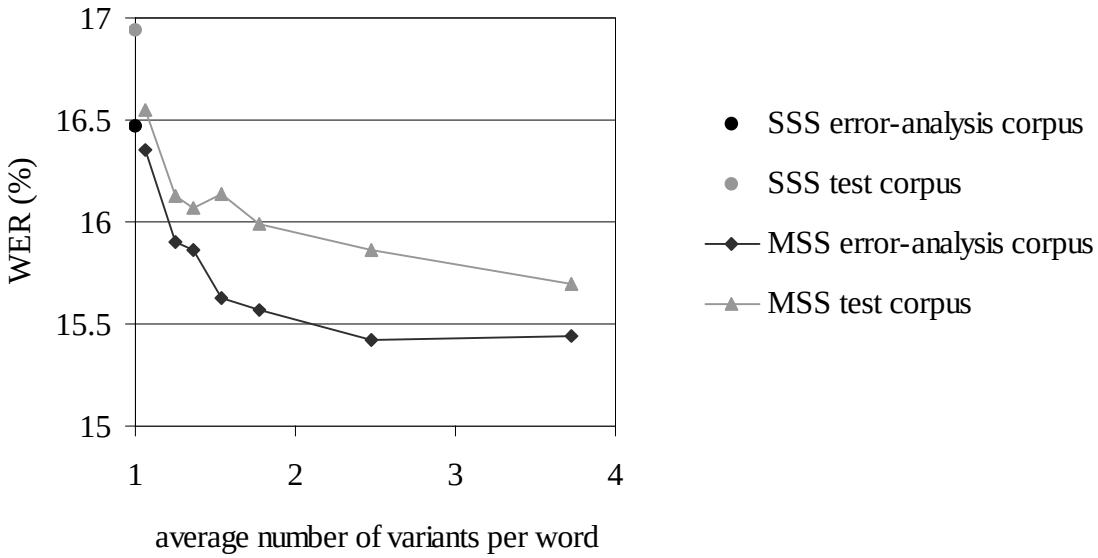

Figure 2: WERs for testing condition `MMM` measured on test- and error analysis corpus

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. Figure 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 reduction in WER:

$$reduction\ in\ WER = WER_{SSS} - WER_{MMM} = 100\%\ x \frac{total\ net\ result}{total\ number\ of\ words} \qquad (2)$$

Figure 3 shows that many changes occur, whereas the total net result or the reduction in WER 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 reduction in WER is (3.8%-2.8%=) 1%. These results show that it is in principle 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.

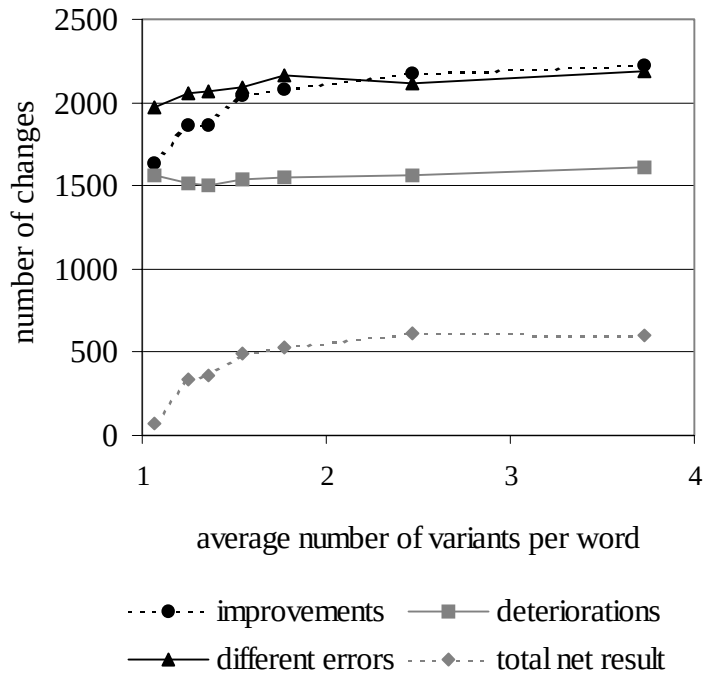

Figure 3: Different types of changes for testing condition MMM compared to SSS measured on error analysis corpus

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

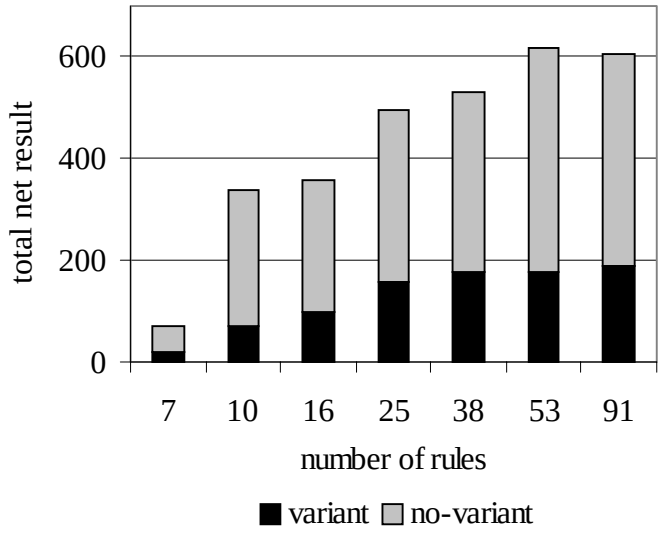

Figure 4: Distribution of the total net result over the two categories of changes for all rule sets

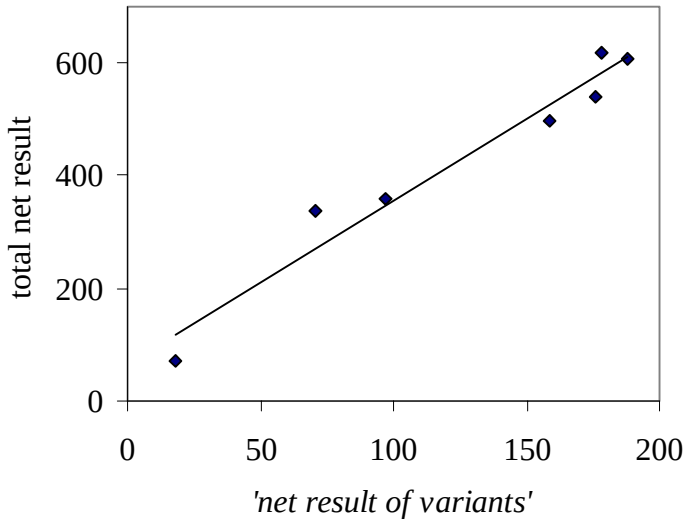

Figure 5: Regression line for correlation between 'net result of variants' at rule set level and the total net result

The category changes with the label *'variant'* (black bars in Figure 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. Figure 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'*.

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.

Figure 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 Figure 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 Figure 6. This is not the case, since Pearson's correlation is still fairly high (0.77) if these six data-points are excluded. Figure 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 reduction in WER, see (2)).

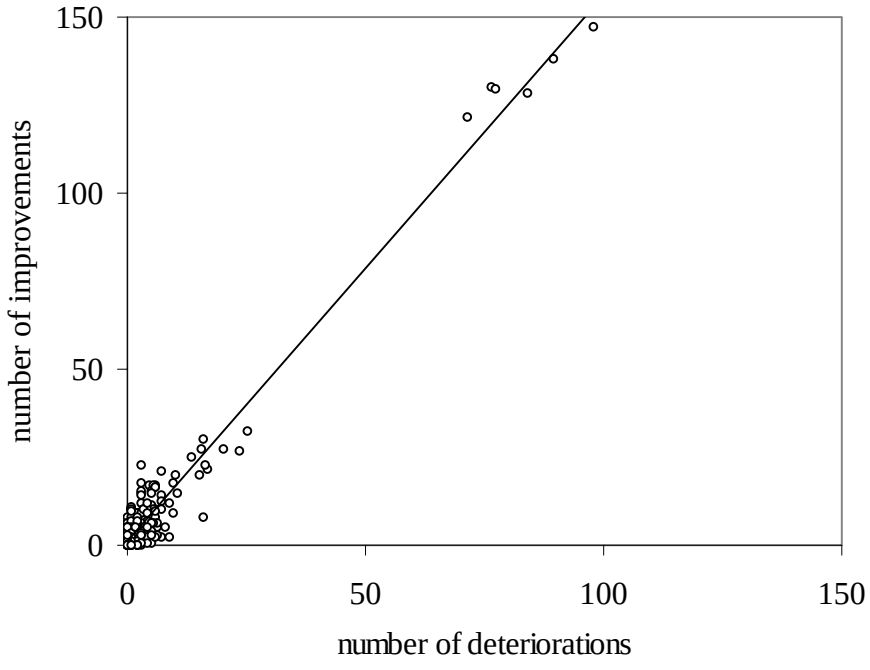

Figure 6: Correlation between improvements and deteriorations

In Figure 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 ≥ 5 in one of the 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).

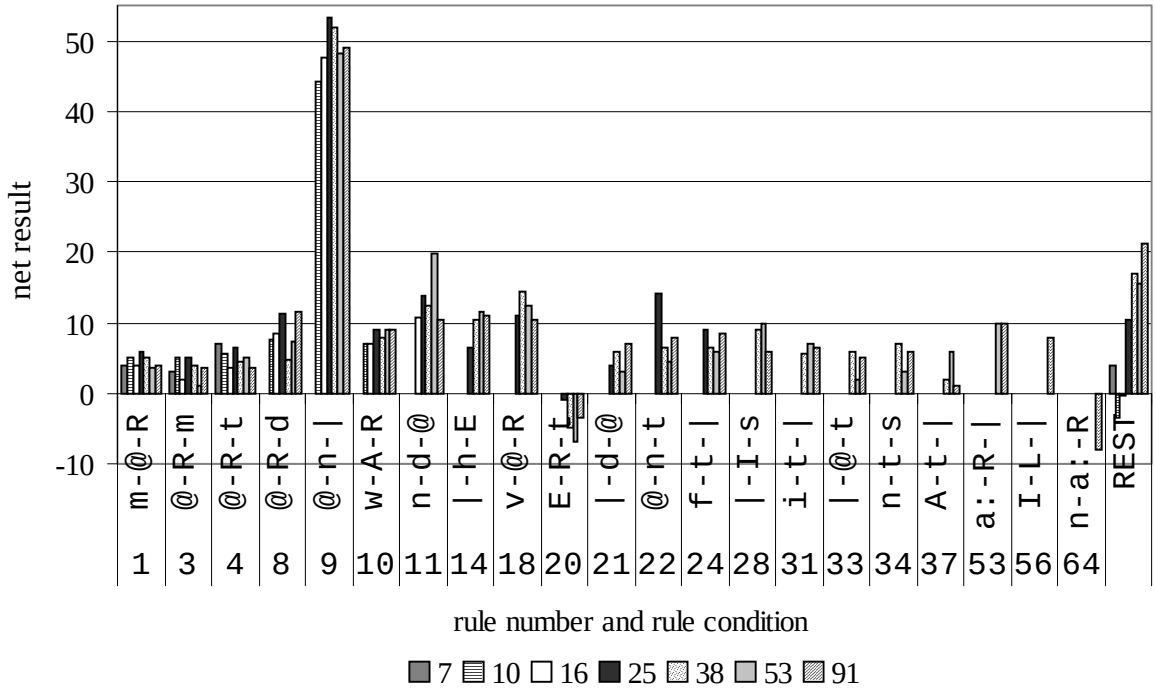

Figure 7: Contributions of each individual rule to the net result

In Figure 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 Figure 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, 2000b) revealed that the differences in 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 summed up. 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 negate 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 (1999). These authors found that the number of errors corrected through the modeling of pronunciation variation are quite significant, but at the same time also new errors were introduced, substantially or completely negating the gains.

The current error analysis also revealed some new results. We found that about 1/3 of the reduction in WER was obtained because a variant was recognized. For this category of changes we can directly determine which rules caused the changes. For the other 2/3 of the reduction in WER 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 reduction in WER (see formula 2), this indicates that the *'net result of variants'* could be used to predict the reduction in WER. 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 reduction in WER, 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 reduction in WER 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 Figure 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 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.

Figure 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. Furthermore, in general, the reduction in WER becomes smaller if the average number of variants per word is increased.

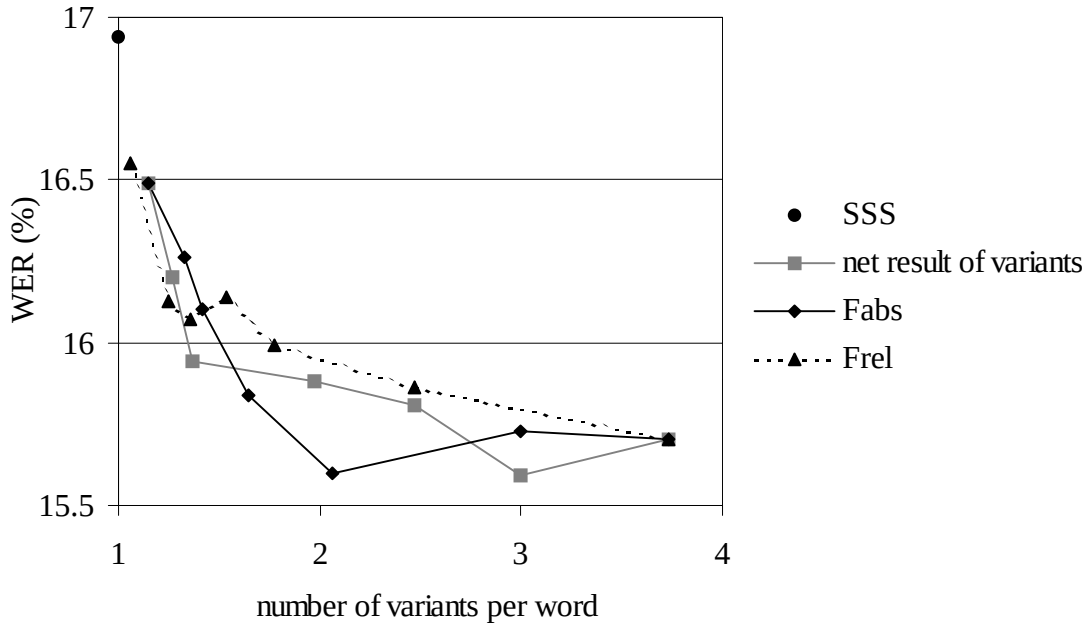

Figure 8: WERs for rules selected on the basis of F_{abs}, F_{rel} and 'net result of variants'

5.2.2 Correlations at rule set level

The reduction in WER was calculated by subtracting all the WERs plotted in Figure 8 from the WER measured for the baseline (16.94%). Since correlations are calculated at rule set level, it was necessary to determine the values of the three criteria at rule set level. 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 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 ($F_{abs\text{-}rule\ set}$) was obtained by summing the values of F_{abs} for all the rules in the set. F_{rel} at rule set level ($F_{rel\text{-}rule\ set}$) was obtained by dividing $F_{abs\text{-}rule\ set}$ by $F_{cond\text{-}rule\ set}$. F_{cond} (see Section 3.1.1, step 6) at rule set level ($F_{cond\text{-}rule\ set}$) was obtained by summing the values of F_{cond} for all the rules in the set.

Figure 9 shows the values of the reduction in WER and the corresponding measures at rule set level, together with the regression lines based on all 19 data points. In Figure 9, '▲' indicates the rule sets selected on the basis of '*net result of variants'*, '◆' indicates the rule sets that are selected on the basis of F_{abs} and '■'

indicates the rule sets selected on the basis of F_{rel}. In Figure 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 reduction in WER increases as the number of rules is increased.

In Figure 9a, it can be seen that if '*net result of variants*$_{\text{- rule set}}$' is increased, the reduction in WER becomes larger, and the correlation is high (0.86). Figure 9b shows that if $F_{abs\text{-}rule\ set}$ is increased, the reduction in WER is also larger, and the correlation is even higher (0.93). The strong correlation between $F_{abs\text{-}rule\ set}$ and reduction in WER 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 Figure 6). This means that the more rules are used, and thus the higher $F_{abs\text{-}rule\ set}$, the larger is the total net improvement.

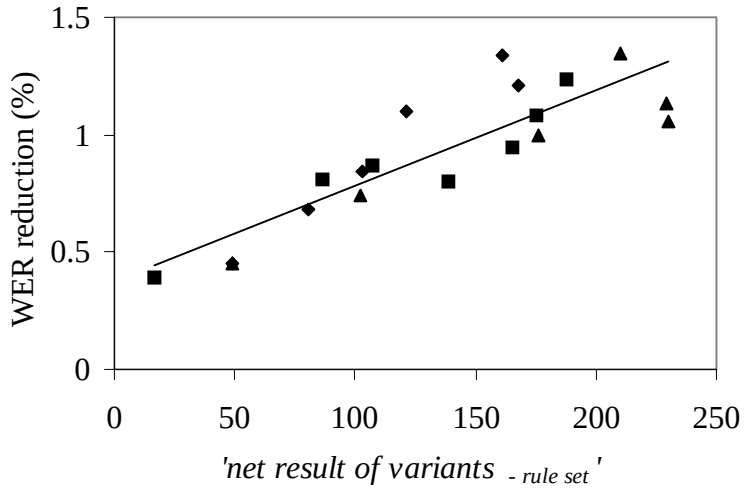

Figure 9a: Relation between 'net result of variants$_{\text{-rule set}}$' and reduction in WER

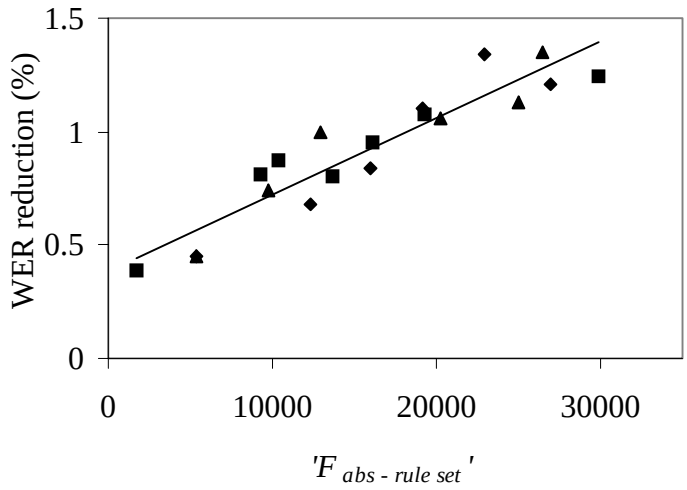

Figure 9b: Relation between $F_{abs\text{-}rule\ set}$ and reduction in WER

In Figure 9c, it can be seen that the reduction in WER is increased if $F_{rel\text{-}rule\ set}$ becomes smaller (Pearson's correlation is -0.83). This is against expectation, as one would expect the reduction in WER 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, namely $F_{rel\text{-}rule\ set}$ and $F_{abs\text{-}rule\ set}$: If $F_{rel\text{-}rule\ set}$ becomes smaller, $F_{abs\text{-}rule\ set}$ increases, and we observed that the reduction in WER is larger if $F_{abs\text{-}rule\ set}$ is increased.

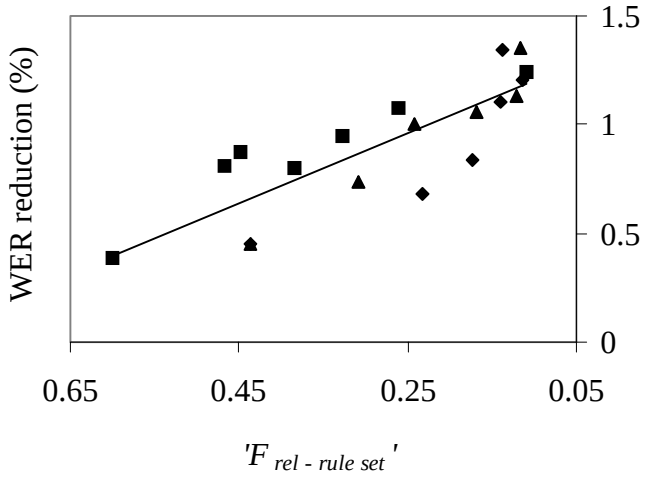

Figure 9c: Relation between $F_{rel\text{-}rule\ set}$ and reduction in WER

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}. Let us try to understand why F_{abs} is probably a better predictor of the reduction in WER 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.

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 reduction in WER.

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 reduction in WER 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 Figure 8). Since decoding time is correlated with the number of entries in the lexicon, this means that the 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.

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 the recognition performance could be further improved by using more rules. If indeed a linear relation exists between $F_{abs\text{-}rule\ set}$ and reduction in WER, as plotted in Figure 9a, then recognition performance could be further improved by increasing $F_{abs\text{-}rule\ set}$. Two remarks should be made about this point. The first remark concerns the linear relationship between $F_{abs\text{-}rule\ set}$ and the reduction in WER. We expect that the relationship between $F_{abs\text{-}rule\ set}$ and reduction in WER cannot be modelled by a simple straight line. For higher values of $F_{abs\text{-}rule\ set}$ we expect the straight line to flatten out. It might even be the case that recognition performance decreases for very high $F_{abs\text{-}rule\ set}$ 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 for our material, the relation between $F_{abs\text{-}rule\ set}$ and the average number of variants per word in the lexicon is not linear, as is shown in Figure 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 for 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.

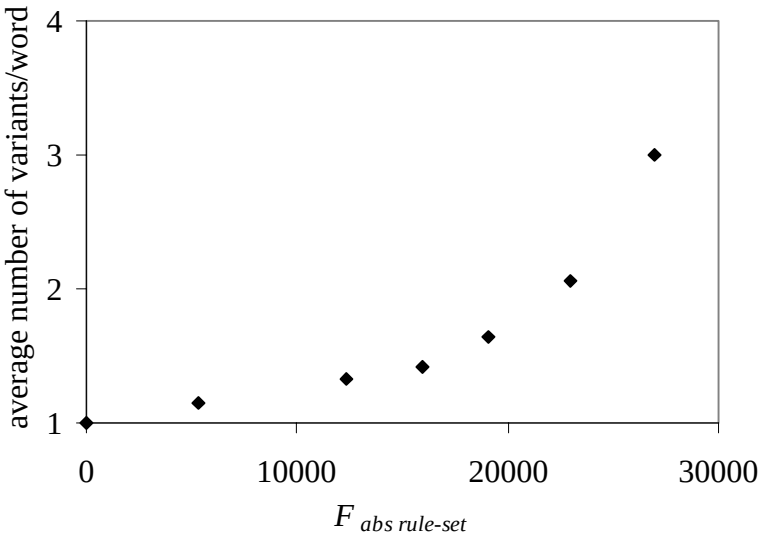

Figure 10: Relation between $F_{abs\text{-}rule\ set}$ and the average number of variants per word in the lexicon. $F_{abs\text{-}rule\ set}$ = 0 corresponds to the baseline system (SSS)

7. GENERAL CONCLUSIONS

As mentioned in the introduction, the aim of the present paper was threefold. First, we analysed whether the data-driven 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 data-driven 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 negate the improvements, resulting in a small total net improvement in WER. These results show that it is in principle 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 reduction in WER 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 reduction in WER 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'*.

8. ACKNOWLEDGMENTS

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.

5. REFERENCES

Amdal, I., Korkmazskiy, F., Surendran, A. C., 2000. Joint pronunciation modeling of non-native speakers using data-driven methods. In: Yuan, B., Huang, T., Tang, X. (eds.), *Proceedings of ICSLP*, Beijing, China, October 16-20 2000, Vol. 3, pp. 622-625.

Cremelie, N., Martens, J.-P., 1999. In search of better pronunciation models for speech recognition. In: Strik (eds.), *Speech Communication* **29**, 115-136.

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

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

Fukada, T., Yoshimura, T., Sagisaka, Y., 1999. Automatic generation of multiple pronunciations based on neural networks. In: *Speech Communication* **27**, 63-73.

Holter, T., Svendsen, T., 1999. Maximum likelihood modelling of pronunciation variation. In: Strik (eds.), *Speech Communication* **29**, 177-191.

Kessens, J.M., Wester, M., and Strik, H., 2000. Automatic Detection and Verification of Dutch Phonological Rules, *PHONUS5: Proceedings of the "Workshop on Phonetics and Phonology in ASR"*, Saarbrücken: Institute of Phonetics, University of the Saarland, december 2000, pp. 117-128.

Kessens, J.M., Wester, M., and Strik, H., 1999. Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation. In: Strik (eds.), *Speech Communication* **29**, 193-207.

Kessens, J.M., Wester, M., Cucchiarini, C. and Strik, H., 1998. The selection of pronunciation variants: comparing the performance of man and machine. In: *Proceedings of ICSLP*, Sydney, Australia, November 30 - December 4 1998, Vol. 6, pp. 2715-2718.

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

Lehtinen, G., Safra, S., 1998. Generation and selection of pronunciation variants for a flexible word recognizer. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), *Proceedings of the ESCA Workshop "Modeling Pronunciation Variation for Automatic Speech Recognition"*, Rolduc, Kerkrade, The Netherlands, 4-6 May 1998, *A^2RT*, University of Nijmegen, pp. 67-72.

Nock, H. J., Young, S. J., 1998. Detecting and correcting poor pronunciations for multiword units. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), *Proceedings of the ESCA Workshop "Modeling Pronunciation Variation for Automatic Speech Recognition"*, Rolduc, Kerkrade, The Netherlands, 4-6 May 1998, *A^2RT*, University of Nijmegen, pp. 85-90.

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. In: Strik (eds.), *Speech Communication* **29**, 209-224.

Ravishankar, M., Eskenazi, M., 1997. Automatic generation of context-dependent pronunciations, In: Kokkinakis, G., Fakotakis, N., Dermatas, E. (eds.), *Proceedings Eurospeech*, Rhodes, Greece, 22-25 September 1997, Vol. 5, pp. 467 - 2470.

Schiel, F., Kipp, A., Tillmann, H. G., 1998. Statistical modeling of pronunciation: It's not the model, it's the data. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), *Proceedings of the ESCA Workshop "Modeling Pronunciation Variation for Automatic Speech Recognition"*, Rolduc, Kerkrade, The Netherlands, 4-6 May 1998, *A^2RT*, 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: *Proceedings of the ESCA Third European Conference on Speech Communication and Technology: Eurospeech*, Berlin, pp. 2125-2128.

Strik, H., 2001. Pronunciation adaptation at the lexical level. In: *Proceedings of the 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. In: Strik (eds.), *Speech Communication* **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. In: *Int. Journal of Speech Technology*, Vol. 2, No. 2, 119-129.

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

Wester, M., Fosler-Lussier, E., 2000a. A comparison of data-derived and knowlegde-based modeling of pronunciation variation. In: Yuan, B., Huang, T., Tang, X. (eds.), *Proceedings of ICSLP*, Beijing, China, October 16-20th 2000, Vol. 4, pp. 270-273.

Wester, M., Kessens, J.M. and Strik, H., 2000b. Pronunciation variation in ASR: Which variation to model? In: Yuan, B., Huang, T., Tang, X. (eds.), *Proceedings of ICSLP*, Beijing, China, October 16-20th 2000, Vol. 4, pp. 488-491.

Yang, Q., Martens, J.-P., 2000a. Data-driven lexical modeling of pronunciation variations for ASR. In: Yuan, B., Huang, T., Tang, X. (eds.), *Proceedings of ICSLP*, Beijing, China, October 16-20th 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: *Proceedings 11th ProRisc Workshop*, November 29 – December 1, Veldhoven, The Netherlands, pp. 589-593.

Wiliams, G., 1999. *Knowing What You Don't Know: Roles for Confidence Measures in Automatic Speech Recognition*, PhD. thesis, Department of Computer Sciences, University of Sheffield, Sheffield, United Kingdom.

Appendix 1

Table 4: Statistics of the 91 selected rules, ordered according to descending F_{rel}. 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.

	Context	*Frel*	*Fabs*		Context	*Frel*	*Fabs*		Context	*Frel*	*Fabs*
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

Samenvatting

Spraak is voor mensen een zeer natuurlijke en efficiënte manier van communiceren. Tegenwoordig is het mogelijk om met behulp van een computer spraak automatisch om te zetten in tekst. Deze techniek wordt *automatische spraakherkenning* (ASH) genoemd. Sinds de komst van de eerste automatische spraakherkenners zijn de herkenprestaties en mogelijkheden van automatische spraakherkenners enorm verbeterd. In het verleden was het alleen mogelijk om een beperkte set van geïsoleerd uitgesproken woorden te herkennen (bijvoorbeeld de cijfers 0-10). Huidige spraakherkenners daarentegen hebben een veel grotere woordenschat en kunnen ook omgaan met *continue spraak.* Bij continue spraak gaat het om complete uitingen, waarbij de woorden niet los, maar aan elkaar uitgesproken worden. De herkenprestaties van huidige spraakherkenners zijn zo goed dat het mogelijk is om spraak in te zetten als communicatiemiddel tussen mens en machine. De mogelijkheden van het gebruik van ASH zijn echter beperkt, omdat ASH niet altijd foutloos werkt.

Ondanks de snelle ontwikkeling van ASH blijken mensen nog steeds beter te zijn in spraakverstaan dan computers. Dit is echter niet zo verwonderlijk, aangezien mensen veel meer (en andere) informatie gebruiken dan spraakherkenners bij de decodering van spraak. Eén van de moeilijkheden van het herkennen van continue spraak is dat de manier waarop woorden worden uitgesproken erg variabel is. Als twee woorden bijvoorbeeld achter elkaar worden gesproken, kan het gebeuren dat sommige klanken niet (of niet volledig) worden uitgesproken, bijvoorbeeld, “dat is” kan uitgesproken worden als “da’s”. Het verschijnsel dat woorden op verschillende manieren kunnen worden uitgesproken wordt ook wel *uitspraakvariatie* genoemd. Mensen hebben meestal geen moeite om de verschillende uitspraken van een woord te herleiden tot één en hetzelfde woord, maar hoe ze dat precies voor elkaar krijgen is niet bekend. Van spraakherkenners weten we wel precies hoe ze werken en dus ook hoe ze zouden kunnen omgaan met uitspraakvariatie. De huidige spraakherkenners maken echter niet altijd expliciet gebruik van de verschillende manieren waarop woorden uitgesproken kunnen worden, waardoor uitspraakvariatie kan leiden tot herkenfouten. In dit onderzoek is daarom nagegaan of het modelleren van uitspraakvariatie in spraakherkenners de herkenprestaties ervan kan verbeteren.

Het proefschrift bestaat uit een viertal artikelen die onderzoek beschrijven dat gerelateerd is aan het modelleren van uitspraakvariatie. De vier artikelen worden voorafgegaan door zes inleidende hoofdstukken die een kader scheppen voor het onderzoek dat beschreven is in de artikelen. In Hoofdstuk 1 worden de basisprincipes van ASH in het kort uitgelegd aan de hand van werking van de automatische spraakherkenner die gebruikt is in dit onderzoek. In Hoofdstuk 2 wordt beschreven welke bronnen van uitspraakvariatie te onderscheiden zijn en wordt uitgelegd waarom uitspraakvariatie kan leiden tot herkenfouten. In Hoofdstuk 3 worden het doel van het onderzoek en de gebruikte onderzoeksmethodologie beschreven. Hoofdstuk 4 bestaat uit de samenvattingen van de vier artikelen. In Hoofdstuk 5 worden de onderzoeksresultaten bediscussieerd. Tenslotte worden in Hoofdstuk 6 de conclusies

van dit proefschrift beschreven, samen met aanbevelingen voor verder onderzoek. De inleidende hoofdstukken worden hieronder kort besproken, gevolgd door samenvattingen van de vier artikelen. Tenslotte worden de algemene conclusies van dit proefschrift en suggesties voor toekomstig onderzoek gegeven.

Automatische spraakherkenning

De werking van een automatische spraakherkenner kan in het kort als volgt uitgelegd worden. Een spraakherkenner kan opgebouwd gedacht worden uit drie modules:

1. Het *lexicon*. Het lexicon bestaat uit een lijst van alle woorden die de spraakherkenner kan herkennen (*orthografische* representatie), samen met een beschrijving van de klanken waaruit de standaarduitspraak van het woord is opgebouwd (*fonetische* representatie).
2. De *foonmodellen*. Dit zijn statistische modellen waarin de akoestische eigenschappen van de klanken (*fonen*) van de taal zijn vastgelegd.
3. Het *taalmodel*. Het taalmodel bevat statistische informatie over de taal, zoals de frequentie van voorkomen van woorden en sequenties van woorden.

Tijdens *training* worden de parameters van de foonmodellen geschat aan de hand van een grote hoeveelheid spraak met bijbehorende automatisch gegenereerde fonetische transcripties. De parameters van het taalmodel worden geschat op basis van een grote hoeveelheid tekst (orthografische transcripties). Tijdens *herkenning* wordt voor een groot aantal mogelijke sequenties van woorden (hypotheses) de waarschijnlijkheden bepaald. Hiertoe zijn twee scores bepalend:

1. De *akoestische score*; deze wordt geschat met behulp van de foonmodellen en geeft aan hoe waarschijnlijk het is dat het geobserveerde akoestische signaal is gegenereerd door het statistische model van ieder afzonderlijk foon, en
2. De *taalmodel score*; deze wordt geschat met behulp van het taalmodel en geeft de a priori waarschijnlijkheid voor iedere hypothese aan.

De hypothese met de hoogste totale waarschijnlijkheid is de sequentie van woorden die uiteindelijk wordt herkend.

De spraakherkenner die in dit onderzoek gebruikt is vormt een onderdeel van het gesproken dialoogsysteem OVIS (Openbaar Vervoer Informatie Systeem). Door met OVIS te bellen kan telefonisch informatie worden verkregen over binnenlandse treinreizen. Voor het trainen van de foonmodellen en het taalmodel en voor het uitvoeren van de herkenexperimenten is spraakmateriaal nodig. Het spraakmateriaal dat we hebben gebruikt in het in dit proefschrift beschreven onderzoek bestaat uit opnames van telefoongesprekken met OVIS.

Uitspraakvariatie

In onze *referentieherkenner* is slechts één uitspraak per woord aanwezig. Een uitspraak die afwijkt van de uitspraak in het lexicon kan op twee verschillende manieren herkenfouten veroorzaken. Ten eerste kan de afwijking in de uitspraak zo groot zijn dat er een ander woord in het lexicon is dat meer op het uitgesproken woord lijkt en dus ten onrechte wordt herkend. Ten tweede zorgt de afwijkende uitspraak er

tijdens training voor dat verkeerde stukken akoestisch signaal worden toegewezen aan een foonmodel, waardoor dit foonmodel wordt vervuild. Het gebruik van deze vervuilde foonmodellen tijdens herkenning kan vervolgens weer leiden tot herkenfouten.

Methodes voor het modelleren van uitspraakvariatie ter verbetering van ASH kunnen op verschillende manieren ingedeeld worden. Voor het onderzoek beschreven in dit proefschrift is het belangrijk om een onderscheid te maken tussen kennisgebaseerde en datagestuurde methoden. Bij een kennisgebaseerde methode wordt de informatie over uitspraakvariatie uit de literatuur gehaald, terwijl bij een datagestuurde methode deze informatie uit (een grote hoeveelheid) spraakdata wordt afgeleid.

Doel en onderzoeksmethodologie

Het doel van het modelleren van uitspraakvariatie is om de herkenprestaties van spraakherkenners te verbeteren. Voordat het mogelijk is om uitspraakvariatie adequaat te modelleren, is het nodig om te weten welke uitspraakvarianten voorkomen in de spraak die de herkenner moet kunnen verwerken en wat de frequenties van voorkomen van de uitspraakvarianten is. Deze informatie kan verkregen worden door fonetische transcripties te maken van zeer grote hoeveelheden spraakmateriaal. In dit onderzoek is ervoor gekozen om de transcripties automatisch te genereren. Dit houdt in dat de spraakherkenner zelf op basis van het akoestisch signaal beslist welke van een aantal mogelijk uitspraakvarianten het meest waarschijnlijk is uitgesproken. Aangezien *automatische transcriptie* een essentieel onderdeel vormt van onze onderzoeksmethodologie is eerst een uitgebreide studie verricht waarin de gebruikte automatische transcriptiemethode nader is onderzocht. Deze studie is beschreven in de eerste twee artikelen van het proefschrift. Het doel van dit deel van het onderzoek is om erachter te komen wat de kwaliteit van automatisch gegenereerde transcripties is en hoe de best mogelijke automatische transcripties verkregen kunnen worden. In de laatste twee artikelen worden twee studies beschreven waarin uitspraakvariatie wordt gemodelleerd. Het doel van deze studies is om te achterhalen of het mogelijk is de herkenprestaties van spraakherkenners te verbeteren door het modelleren van uitspraakvariatie. Bovendien hopen we ook meer inzicht te krijgen in hoe uitspraakvariatie het best gemodelleerd kan worden.

Onze onderzoeksmethodologie komt erop neer dat uitspraakvariatie wordt gemodelleerd in alle drie de modules van de spraakherkenner. Ten eerste worden er uitspraakvarianten toegevoegd aan het referentielexicon (dat één fonetische transcriptie voor ieder woord bevat), zodat er voor sommige woorden verschillende uitspraken mogelijk zijn. Op deze manier is er een betere overeenstemming tussen de gerealiseerde uitspraak van woorden en de fonetische transcriptie ervan in het lexicon. Ten tweede wordt een automatische fonetische transcriptie van het trainingsmateriaal gemaakt. Tijdens automatische transcriptie gebruikt de spraakherkenner een lexicon waaraan uitspraakvarianten zijn toegevoegd en beslist de herkenner zelf welke van de varianten het best past bij het akoestische signaal. Het is de verwachting dat deze automatisch verkregen fonetische transcripties de spraak beter beschrijven dan de

fonetische transcripties die verkregen zijn uit het referentielexicon. Op basis van deze nauwkeurigere transcripties van het trainingsmateriaal worden nieuwe foonmodellen getraind. Ten derde wordt uitspraakvariatie gemodelleerd in het taalmodel. Dit houdt in dat iedere uitspraakvariant een eigen a priori waarschijnlijkheid krijgt. Deze waarschijnlijkheden worden geschat op basis van de nieuwe automatisch verkregen transcripties van het trainingsmateriaal. Om te voorkomen dat uitspraakvarianten van onwaarschijnlijke woorden ten onrechte worden verward met andere woorden in het lexicon, wordt gebruik gemaakt van variantspecifieke waarschijnlijkheden.

Artikel 1

In artikel 1 is de kwaliteit van de automatische fonetische transcripties onderzocht door de automatische transcripties te vergelijken met transcripties gemaakt door ervaren transcribenten. Dit zijn mensen die ervaring hebben in het maken van fonetische transcripties van spraak. De transcriptietaak van de spraakherkenner bestond uit een gedwongen keuze uit een beperkt aantal mogelijke uitspraakvarianten voor een beperkt aantal woorden. De varianten werden automatisch gegenereerd door vijf optionele fonologische regels toe te passen op de woorden in het lexicon. Deze regels zijn gebaseerd op de volgende vijf frequent voorkomende fonologische processen: /n/-, /r/-, /t/-, /@/-deletie en /@/-insertie. Aangezien transcribenten ook fouten maken, is het niet mogelijk om een referentietranscriptie te verkrijgen waarvan aangenomen kan worden dat deze volledig correct is. Om deze reden hebben we twee verschillende strategieën gebruikt om menselijke referentietranscripties te verkrijgen in de twee experimenten die zijn uitgevoerd. In het eerste experiment gebruikten we een referentietranscriptie gebaseerd op het meerderheidsoordeel van negen ervaren transcribenten die onafhankelijk van elkaar werkten, terwijl in het tweede experiment twee (of drie) transcribenten consensus moesten bereiken over de referentietranscriptie. Als kwaliteitsmaat voor de automatische transcripties gebruikten we de mate van overeenstemming tussen de automatische transcripties en de referentietranscripties. Hoe groter de mate van overeenstemming tussen de automatische transcripties en de referentietranscripties, hoe hoger de transcriptiekwaliteit.

De belangrijkste conclusies van het eerste experiment is dat de mate van overeenstemming met de referentietrancripties significant lager is voor de spraakherkenner dan voor de transcribenten. Het is echter ook gebleken dat de verschillen niet voor alle vijf de regels significant zijn en dat voor één van de transcribenten de mate van overeenstemming ook significant lager was dan voor de overige transcribenten. De verschillen tussen automatisch en handmatig verkregen transcripties zijn echter niet groot; ze kunnen heel goed acceptabel zijn, afhankelijk van het doel waarvoor de transcripties gebruikt worden.

In het tweede experiment is specifiek gekeken naar de transcriptie van het foon /@/ in de context van de /@/-deletie en /@/-insertie regels. Hieruit blijkt dat de spraakherkenner en de transcribenten een andere drempel voor de duur van de /@/ gebruiken op grond waarvan besloten wordt of de /@/ uitgesproken is of niet.

Artikel 2

In artikel 2 hebben we nader onderzocht wat de relatie is tussen een aantal eigenschappen van de spraakherkenner en transcriptiekwaliteit. De uitspraakvarianten werden weer automatisch gegenereerd door dezelfde vijf optionele fonologische regels toe te passen op de woorden in het lexicon als in artikel 1. Als kwaliteitsmaat voor de automatische transcripties gebruikten we weer de mate van overeenstemming tussen de automatische transcripties en de referentietranscripties. Zowel referentietranscripties gebaseerd op het meerderheidsoordeel van de transcribenten als consensus transcripties werden gebruikt.

Ten eerste hebben we gekeken naar de invloed van verschillende eigenschappen van de foonmodellen op de transcriptiekwaliteit. Hiertoe hebben we vier experimenten uitgevoerd. Het eerste experiment toont aan dat de impliciete minimale duur van een foonmodel die gerelateerd is aan de topologie van de foonmodellen invloed heeft op de transcriptiekwaliteit. De minimale duur opgelegd door de topologie van de foonmodellen die wij gebruiken tijdens een normale herkentaak blijkt te lang te zijn voor automatische transcriptie waardoor het moeilijker is om zeer korte fonen te detecteren. Het tweede experiment laat zien dat voor automatische transcripties het best foonmodellen gebruikt kunnen worden die getraind zijn op spraakmateriaal waarvoor de transcriptie zeer nauwkeurig aansluit bij hetgeen gezegd is. Uit het derde experiment blijkt dat het gebruik van context-afhankelijke t.o.v. context-onafhankelijke foonmodellen niet altijd leidt tot een betere transcriptiekwaliteit. Het vierde experiment laat zien dat het gelijktijdig optimaliseren van bovengenoemde eigenschappen van de foonmodellen tot een nog hogere transcriptiekwaliteit leidt.

Ten tweede hebben we onderzocht of er een relatie bestaat tussen het percentage herkenfouten dat een spraakherkenner maakt tijdens een normale herkentaak en de kwaliteit van de automatische transcripties die met dezelfde spraakherkenner worden gegenereerd. Uit deze vergelijking blijkt dat er geen duidelijk verband is tussen het percentage herkenfouten en de transcriptiekwaliteit behaald met dezelfde spraakherkenner. Deze bevinding bevestigt het intuïtieve idee dat fonetisch transcriberen en spraakverstaan (net als voor mensen) twee verschillende processen zijn. Voor automatische fonetische transcriptie is het daarom noodzakelijk om spraakherkenners te ontwikkelen die geoptimaliseerd zijn voor deze taak.

Artikel 3

Artikel 3 beschrijft een studie waarin een kennisgebaseerde methode voor het modelleren van binnen- en tussenwoorduitspraakvariatie is onderzocht. De binnenwoorduitspraakvarianten werden automatisch gegeneerd door de vijf fonologische regels die gebruikt zijn in de eerste twee artikelen toe te passen op de woorden in het lexicon. Verder zijn ook tussenwoorduitspraakvarianten (t.g.v. reductie, contractie en cliticizatie) gegenereerd voor een aantal zeer frequent voorkomende woordsequenties. Vervolgens hebben we zowel de binnen- als de tussenwoorduitspraakvariatie in alle drie de modules van de spraakherkenner gemodelleerd en hebben we gemeten wat de invloed is op de herkenprestaties. Uit deze experimenten blijkt dat toevoegen van uitspraakvarianten aan het lexicon tot een

kleine verbetering in herkenprestaties leidt. Het hertrainen van de foonmodellen leidt ook tot een geringe verbetering. Als tenslotte a priori waarschijnlijkheden worden gebruikt voor de uitspraakvarianten wordt de grootste verbetering gevonden. Uit een vergelijking van twee methodes om tussenwoorduitspraakvariatie te modelleren blijkt dat tussenwoordvariatie het beste gemodelleerd kan worden door een aantal zeer frequente woordsequenties als aparte woorden op te nemen in het lexicon - zogenaamde *multiwoorden* - en vervolgens uitspraakvarianten te genereren voor deze multiwoorden. Uit de herkenexperimenten blijkt verder dat er interactie plaatsvindt tussen uitspraakvarianten: De verbetering in het percentage herkenfouten die je zou verwachten op basis van experimenten waarin de binnenwoord- en tussenwoordvarianten in isolatie worden getest is niet gelijk aan de verbetering die gevonden wordt als de varianten in combinatie worden getest. Tenslotte blijkt dat de grootste verbetering wordt gevonden als binnen- en tussenwoorduitspraakvariatie gelijktijdig worden gemodelleerd: T.o.v. onze referentieherkenner vinden we een significante verbetering in het percentage fout herkende woorden van 1.1% absoluut of 8.8% relatief.

Artikel 4

Artikel 4 beschrijft een studie waarin een datagestuurde methode voor het modelleren van een uitspraakvariatie is onderzocht. In continue spraak komt het vaak voor dat niet alle fonen waar een woord uit bestaat worden uitgesproken. In dit onderzoek concentreren we ons op deze zogenaamde *deleties* van fonen. Dit onderzoek bestaat uit drie deelstudies. De methode om de informatie over de deletieprocessen uit de data af te leiden is in deze drie deelstudies gelijk en werkt als volgt. Allereerst wordt een automatische transcriptie gemaakt van een grote hoeveelheid spraakmateriaal. Hiertoe worden een zeer groot aantal mogelijke uitspraakvarianten automatisch gegenereerd door ieder foon in de fonetische transcriptie optioneel te maken. De automatisch verkregen fonetische transcripties van het spraakmateriaal worden vervolgens opgelijnd met de transcripties die worden opgezocht in het lexicon van onze referentieherkenner. Uit de opgelijnde transcripties worden vervolgens deletieregels afgeleid. Een deletieregel beschrijft in welke context (linker- en rechterbuurfoon) een foon gedeleerd wordt. Tenslotte wordt op een aantal verschillende manieren regels geselecteerd.

In de eerste deelstudie worden de regels geselecteerd met de hoogste relatieve frequentie van toepassen. Vervolgens worden met deze regels uitspraakvarianten gegeneerd, die in alle drie de modules van de spraakherkenner worden gebruikt. Uit herkenexperimenten blijkt dat als uitspraakvarianten alleen toegevoegd worden aan het lexicon niet altijd een verbetering in herkenfouten wordt gevonden. Als het aantal toegevoegde varianten erg groot is wordt zelfs een verslechtering gevonden. Verder blijkt wederom dat het hertrainen van de foonmodellen de herkenprestaties slechts minimaal verbetert. Tenslotte laten onze experimenten zien dat het gebruik van a priori waarschijnlijkheden voor uitspraakvarianten van cruciaal belang is. Als de datagestuurde varianten worden gebruikt in alle modules van de spraakherkenner,

vinden we t.o.v. de referentieherkenner een significante verbetering in het percentage fout herkende woorden van 1.2% absoluut of 7.3% relatief.

In de tweede deelstudie hebben we geprobeerd te achterhalen hoe de veranderingen in het herkenresultaat precies tot stand zijn gekomen door een uitgebreide foutenanalyse uit te voeren. Deze foutenanalyse laat zien dat er naast verbeteringen ook verslechteringen optreden als gevolg van het modelleren van uitspraakvariatie. Door de introductie van deze verslechteringen wordt slechts een kleine netto verbetering in herkenresultaat gevonden. Verder blijkt er een sterke correlatie te bestaan tussen het aantal verbeteringen en verslechteringen op regelniveau. Dit betekent dat het niet mogelijk is de herkenprestaties te verbeteren door regels uit te sluiten die veel fouten introduceren, omdat deze regels ook fouten oplossen.

In de derde deelstudie hebben we drie maten onderzocht die gebruikt zouden kunnen worden om regels te selecteren. Twee van deze maten zijn gebaseerd op de toepassingsfrequentie van een regel: F_{abs} en F_{rel}, respectievelijk de absolute en relatieve frequentie waarmee een regel is toegepast. De derde maat ('netto resultaat') komt voort uit de eerder uitgevoerde foutenanalyse en geeft aan hoeveel woorden netto beter herkend worden ten opzichte van de referentieherkenner. Het blijkt dat F_{abs} en het 'netto resultaat' het percentage fout herkende woorden het best voorspellen. Van deze twee maten verdient F_{abs} de voorkeur, omdat F_{abs} praktisch gezien het makkelijkst te berekenen is. Als F_{abs} gebruikt wordt om de regels te selecteren vinden we ten opzichte de referentieherkenner een significante verbetering in het percentage fout herkende woorden van 1.4% absoluut of 8.2% relatief.

Conclusies

Op grond van dit proefschrift kunnen een aantal conclusies getrokken over automatische fonetische transcriptie van spraak. We hebben laten zien dat het mogelijk is om met een spraakherkenner automatische fonetische transcripties van spraak te maken. De kwaliteit van deze automatische transcripties is over het algemeen wel iets lager dan de kwaliteit van transcripties die gemaakt zijn door ervaren transcribenten. Of dit verschil in kwaliteit acceptabel is, hangt af van het doel waarvoor de transcripties gebruikt worden. Verder blijkt de kwaliteit van automatische transcripties niet direct gerelateerd te zijn aan de herkenprestaties van een spraakherkenner. Voor het verkrijgen van optimale transcripties is het daarom het beste om bepaalde voor de transcriptietaak specifieke eigenschappen van de spraakherkenner te optimaliseren. Zo blijkt het gebruik van een foonmodeltopologie met een korte impliciete minimale duur de transcriptiekwaliteit te verbeteren. Verder is het beter om foonmodellen te gebruiken die getraind zijn op spraakmateriaal waarvan de transcripties nauwkeurig aansluiten bij de uitspraak. Het gebruik van context-afhankelijke foonmodellen blijkt alleen nuttig te zijn als deze foonmodellen getraind zijn op basis van een nauwkeurige fonetische transcriptie. Tenslotte blijkt dat het combineren van bovengenoemde optimale eigenschappen van de foonmodellen in een nog hogere transcriptiekwaliteit resulteert.

Ten aanzien van het modelleren van uitspraakvariatie in ASH kunnen ook een aantal conclusies getrokken worden. Het blijkt mogelijk te zijn om de herkenprestaties van een spraakherkenner te verbeteren door uitspraakvariatie expliciet te modelleren. Voor zowel de kennisgebaseerde als de datagestuurde modelleermethodes werd een vergelijkbare, significante verbetering in herkenprestaties gemeten. Het opnemen van uitspraakvarianten in het lexicon zonder verdere aanpassingen aan de spraakherkenner blijkt niet altijd nuttig te zijn, vooral als het aantal toegevoegde varianten groot is. Het hertrainen van de foonmodellen op basis van een nauwkeurigere transcriptie van het trainingsmateriaal is slechts van beperkt nut. Tenslotte is een belangrijke conclusie dat als uitspraakvarianten toegevoegd worden aan het lexicon het van cruciaal belang is om gebruik te maken van a priori waarschijnlijkheden van deze varianten.

Verder onderzoek

Het proefschrift eindigt met suggesties voor toekomstig onderzoek. Ten aanzien van automatische fonetische transcriptie worden een aantal mogelijke richtingen aangegeven. Het is wenselijk om meer onderzoek te doen waarin nagegaan wordt in welke mate automatische gegenereerde transcripties verschillen van transcripties die gemaakt zijn door mensen. Verder is het belangrijk om maten te ontwikkelen waarmee de kwaliteit van automatische transcripties ingeschat kan worden zonder dat hiervoor een vergelijking met handmatig gegenereerde transcripties nodig is. In automatische spraakherkenning worden maten gebruikt waarmee kan worden geschat hoe zeker een spraakherkenner is van de uitkomst van het herkenresultaat. Dergelijke maten kunnen waarschijnlijk ook gebruikt worden om de kwaliteit van automatische transcripties te meten.

Op het gebied van het modelleren van uitspraakvariatie worden ook een aantal suggesties voor verder onderzoek gedaan. Het is bekend dat mensen veel meer (en andere) informatiebronnen gebruiken om spraak te herkennen dan de huidige generatie automatische spraakherkenners. Om deze reden is het wenselijk om meer onderzoek te doen naar mogelijke alternatieve informatiebronnen. Op basis van deze extra informatie (zoals spreeksnelheid, de voorspelbaarheid van een woord en de mate waarin een woord geaccentueerd is) kan de waarschijnlijkheid van varianten beter geschat worden. Eén van de redenen waarom de tot dusver gebruikte methoden voor het modelleren van uitspraakvariatie maar een kleine verbetering van het percentage fouten opleveren is dat de uitspraakvarianten die toegevoegd zijn aan het lexicon verward worden met andere woorden in het lexicon. Een methode om de verwarbaarheid van varianten te verkleinen is om uitspraakvariatie *dynamisch* te modelleren, d.w.z. dat de varianten alleen gebruikt worden als ze zeer waarschijnlijk zijn. Dynamisch modelleren van uitspraakvariatie wordt gezien als een veelbelovende onderzoeksrichting in uitspraakvariatieonderzoek. Tenslotte wordt aangegeven dat het voor het vergelijken van verschillende methodes om uitspraakvariatie te modelleren niet voldoende is om alleen herkenpercentages te rapporteren. Een uitgebreide analyse van de veranderingen in herkenresultaat geeft een beter beeld van het effect van het modelleren van uitspraakvariatie en maakt het mogelijk om de verschillende methodes beter met elkaar te vergelijken.

Curriculum Vitae

Judith Maria Kessens was born on February 19th, 1972 in Muiden, The Netherlands. She attended Haarlemmermeercollege in Badhoevedorp and Katholiek College Amsterdam-west (KCA) in Amsterdam, where she obtained her VWO diploma in 1990. From 1990 to 1996 she studied Technical Physics at the Technical University (TU) of Delft, with a specialization in Acoustics. For her M.Sc. thesis she carried out research on noise reduction in order to improve speech intelligibility for the hearing impaired. This research was carried out at the department of Experimental Audiology of the Academic Hospital of the Vrije Universiteit (VU) in Amsterdam. From December 1996 to July 2001 she was employed as a Ph.D. student by the Netherlands Organization for Scientific Research (NWO) for the 'Language and Speech Technology Priority Programme'. During this time she was stationed at the Department of Language and Speech at the faculty of Arts, at the University of Nijmegen (KUN). The present thesis reports on the research that was carried out within the framework of this project. In 1998, she was co-organizer of the ESCA workshop “Modeling Pronunciation Variation for Automatic Speech Recognition”, which took place from 4-6 May 1998 in Rolduc, Kerkrade, The Netherlands. Since November 2001 she is working on the EU-funded IST project (10651) ‘Multimedia Indexing and Searching environment’ (MUMIS).