Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 18 (2004) 123–141
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

On automatic phonetic transcription quality: lower word error rates do not guarantee better transcriptions

Judith M. Kessens, Helmer Strik *

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

Received 2 February 2003; received in revised form 24 June 2003; accepted 14 July 2003

Abstract

The first goal of this study was to investigate the effect of changing several properties of a continuous speech recognizer (CSR) on the automatic phonetic transcriptions generated by the same CSR. Our results show that the quality of the automatic transcriptions can be improved by using 'short' hidden Markov models (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, we found that context-dependent HMMs should preferably not be trained on baseline transcriptions if there is a mismatch between these baseline transcriptions of the speech material and the realized pronunciation. Finally, we found that by combining the changes in the properties of the CSR, the quality of automatic transcription can be further improved.

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

1. Introduction

Phonetic transcriptions (PTs) of speech are used in many disciplines. They can be obtained in two ways: manually or automatically. Manual phonetic transcriptions (MPTs) are made by experts who listen to an utterance and transcribe it into a sequence of speech units represented by

* Corresponding author. Tel.: +31-24-361-61-04; fax: +31-24-361-29-07.
E-mail address: strik@let.kun.nl (H. Strik).
URL: http://lands.let.kun.nl.

0885-2308/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0885-2308(03)00034-2

phonetic symbols. Making MPTs is extremely timeconsuming and therefore costly. Moreover, MPTs tend to contain an element of subjectivity (Cucchiarini, 1993; Shriberg & Lof, 1991). PTs can also be obtained automatically, e.g., by means of an automatic speech recognizer. This results in what we will call automatic phonetic transcriptions (APTs) in this paper. APTs are much faster to make, and are therefore much cheaper than MPTs.

APTs can be made in various ways. One approach is to perform phone recognition. In this kind of recognition phones are recognized, instead of words, as is the case during a normal recognition task. The recognizer is often constrained by a phone N-gram (a sort of phonotactic constraint), and by penalties on the generation of many short sequences of phones. When the content of an utterance (the orthographic transcription) is available, a different kind of APTs can be made. The orthographic transcription is first converted to a canonical phonetic transcription, e.g., by looking up the words in a lexicon or by means of a grapheme–phoneme converter. Next, a number of possible pronunciation variants are generated on the basis of the canonical phonetic transcription, e.g., by applying phonological rules (e.g. Adda-Decker & Lamel, 1998), data-derived rules (e.g. Kessens & Strik, 2003) or by means of decision trees (e.g. Riley et al., 1998). Then, the task of the recognizer is to decide for each word which of the variants best matches the acoustic signal. This is usually called forced recognition or forced alignment. In this study, forced recognition is employed. The number of transcription variants is restricted by allowing only pronunciation variants that are generated by applying five phonological rules.

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

1. *Consensus transcription*. A transcription made by several transcribers after they have agreed on each individual symbol (Shriberg, Kwiatkowski, & Hoffman, 1984).
2. *Majority vote principle*. The material is transcribed by more than one transcriber and the final transcription is the one chosen by the majority of the transcribers (Wester, Kessens, Cucchiarini, & Strik, 2001).

In this paper, both these strategies to obtain RTs were used. As a quality measure for the various APTs, we used agreement between the APTs and the human RTs: the higher the agreement with the human RTs, the better the quality of the APTs.

It is likely that the quality of APTs made with a continuous speech recognizer (CSR) will depend on the properties of this CSR. Therefore, it is important to know how the properties of CSRs influence the resulting APTs. Although CSRs have often been used for making APTs (see, e.g., the many references mentioned on p. 230 of Strik & Cucchiarini, 1999), there are very few studies in which this relationship has been studied in a systematic way. The few exceptions are the studies by Cox, Brady, and Jackson (1998) and Saraçlar and colleagues (Saraçlar, 2000; Saraçlar, Nock, & Khudanpur, 2000). The main goal of Cox et al. (1998) was to automatically generate annotations that result in synthetic speech of the same quality as that produced from hand-labeled speech. They observed the highest agreement between APTs and MPTs when the acoustic models were trained using only the phones for which the confidence level exceed a certain threshold. The main goal of Saraçlar (2000) and Saraçlar et al. (2000) was to reduce word error rates (WERs) by modeling pronunciation variation. A part of their study was concerned with trying to improve the

agreement between APTs and MPTs. They obtained the best results when manual transcriptions were used to train the acoustic models.

Given the importance of the relationship between the properties of CSRs and APTs, and given the fact that there are very few studies in which this relationship has been studied systematically, we decided to carry out a study on this relationship. The first goal of this study, described in the current paper, was to investigate and compare a number of properties of CSRs for their effects on the quality of APTs. There are many properties of a CSR that can be changed and studied. The properties we studied were selected mainly on the basis of the following two criteria: (1) their potential to improve the agreement between APTs and MPTs and (2) their relevance for others. Regarding the second criterion, we did not select properties that are peculiar for our CSR, but instead studied general properties that are more likely to be useful for others. Since almost all CSRs use hidden Markov models (HMMs), we studied general properties of HMMs: topology of the HMMs (Section 3.1), degree of contamination of the HMMs (Section 3.2), context-independent versus context-dependent HMMs (Section 3.3), and combinations of these properties (Section 3.4). According to this line of reasoning, we left out of consideration other properties of CSRs, such as transition costs, because they are not used in all CSRs, and, if they are used, it is usually not straightforward how they should be changed. For similar reasons we did not investigate other factors, such as pronunciation variation consistency, because they rank lower given the two selection criteria mentioned above.

If changing properties of a CSR has an effect on the quality of the resulting APTs, then the choice of the CSR which will be used to make the APTs becomes an important one. In previous research on APT (Wester et al., 2001), we simply took the CSR that yielded the lowest WER that was available from our research on pronunciation variation modeling. In other research on APT, the choice of the CSR is usually not motivated explicitly. Intuitively, one might expect that the CSR that obtains the lowest WERs will also yield the best APTs. However, (automatic) speech recognition may well be quite a different task from (automatic) phonetic transcription. Therefore, it is worthwhile investigating whether lower WERs are indeed an indicator of higher quality APTs. This was the second goal of the research reported here.

This paper is organized as follows: in Section 2, the method that we employed is described. Subsequently, in Section 3, we present the results for each of the properties of the CSR that is investigated. The relationship between degree of agreement and WER is examined in Section 4. Finally, in Section 5, our general conclusions are presented.

2. Method

Section 2.1 describes the speech material used in the experiments. The CSR used for making the automatic transcriptions and for performing the recognition experiments in Section 4 is described in Section 2.2. Next, Sections 2.3 and 2.4 explain how the automatic and manual transcriptions were obtained. Finally, Section 2.5 describes how the automatic transcriptions were evaluated.

2.1. Speech material

The speech material used in the experiments was taken from the Dutch database VIOS. This database 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. Although the speakers in the VIOS database come from all over the Netherlands, this database contains only a very small number of dialectal or non-native speakers of Dutch.

Two sets of data were selected and for each data set a different kind of human RT was obtained. For the first set, a reference transcription based on a *majority vote* procedure was employed. This data set is equal to the one that was used in Wester et al. (2001). For the second set, a *consensus transcription* was made. The statistics of the two sets of transcription material are given in Table 1. In the columns '# utts', '# words', and '# words/utt', the number of utterances and words and the average number of words per utterance are given. Furthermore, in the last two columns ('total duration' and 'average duration'), numerical information on the total duration and the average duration per utterance are presented.

2.2. CSR

We used the CSR that is part of 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 were calculated for frames with a width of 16 ms. The HMMs were 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 were trained: 35 context-independent HMMs, one HMM for non-speech sounds, one HMM for filled pauses, and one HMM (consisting of one state) to model silence. The baseline lexicon contains one transcription for each word. These canonical transcriptions were obtained using the grapheme–phoneme converter that is part of a Text-to-Speech system for Dutch (Kerkhoff & Rietveld, 1994), followed by a manual correction. For the speech recognition experiments described in Section 4, a unigram and bigram language model was used, which was trained on the same 25,104 VIOS utterances used to train the acoustic models.

2.3. Automatic transcriptions

As explained in Section 1, the focus of this study is a restricted form of automatic transcription. The recognizer could only choose pronunciation variants that were automatically generated from canonical transcriptions by applying five phonological rules: /n/-deletion, /r/-deletion, /t/-deletion,

Table 1
Statistics of transcription material

Material	# utts	# words	# words/utt	Total duration (min)	Average duration (s)
Majority vote	186	1208	6.5	6:52	2:13
Consensus	296	2035	6.9	11:20	2:18
Total	482	3243	6.7	18:12	2:16

/@/-deletion, and /@/-insertion (SAMPA [1] notation is used throughout this paper). For more details and a description of the five phonological rules, see Wester et al. (2001). The main reasons for selecting these five phonological processes are that they are well described in the linguistic literature and that they occur frequently in Dutch. Moreover, these phonological processes typically occur in fast, extemporaneous speech. Therefore, it is to be expected that these processes occur often in the speech material that we used (see Section 2.1). In Kessens, Wester, Cucchiarini, and Strik (1997), we investigated the /n/-deletion, /t/-deletion, /@/-deletion, and /@/-insertion rules, and found that, on average, one of the non-canonical variants was chosen in 45% of the tokens for which pronunciation variants are present in the lexicon; this amounts to 8% and 10% of the total number of words in the test and training material, respectively.

The task of the CSR was 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 from 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).

2.4. Manual transcriptions

As already mentioned, two kinds of human RTs were obtained. The following two subsections give more details on the procedures to obtain these transcriptions.

2.4.1. 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 linguists who listened to the speech signal and decided which pronunciation variant best matched the realization that they had just heard. In this sense, their task was exactly the same as the CSRs, i.e., deciding which pronunciation variant matched the speech signal best. The listeners were selected to participate in this experiment because they all had carried out similar tasks for their own research. For this reason, they are representative of the group 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 (five or more out of nine) is taken to be the RT.

2.4.2. Consensus reference transcriptions

The transcribers who made the consensus reference transcriptions were Speech and Language Pathology students at the University of Nijmegen. They had all attended the same transcription course including 32 h contact time. The transcriptions used in this experiment were made as part of their final examination. The IPA transcription alphabet was used in this course. The transcribers all worked in one of 12 groups of two or three people (11 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

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

two (or three) listeners had to agree on each symbol in the utterance. The speech material was distributed among the groups in such a way that the number of words that each group had to transcribe was about equal. There was no overlap between the transcription material of the different groups.

The consensus transcriptions could not directly be used for analysis, as they were produced using the whole range of IPA symbols and diacritics, whereas the CSR used a limited set of SAMPA symbols. For this reason, the diacritics were discarded and the IPA-symbols were mapped to SAMPA symbols, as is shown in Table 2.

The different IPA symbols shown in Table 2 are all allophonic variants of the phone that is represented by the corresponding SAMPA symbol. Whenever the consensus transcription was not an allophonic variant but a different phone, the transcription was excluded from further analysis. In total, 22 consensus transcriptions were excluded: 1 /n/-deletion, 16 /r/-deletions, 2 /t/-deletions, 2 /@/-deletions, and 1 /@/-insertion. This results in the number of transcribed phones presented in the next section (Table 4).

2.5. Evaluation of the APTs

For analysis purposes, we treated the transcription task as a binary decision task: a binary score was obtained for each phone that could either be deleted or inserted as a result of the application of one of the five phonological rules: 1, if the rule was applied and 0, if this was not the case. To clarify this, let us consider the following example: for the word /dELft/ ('Delft') the conditions for application of the /t/-deletion and the /@/-insertion rules are met; thus, four pronunciation variants were generated. Table 3 shows the four variants (column 1), the rules that were applied (column 2), and the corresponding binary scores (column 3).

In total, 1237 binary decisions had to be made during automatic transcription. Table 4 shows the distributions of the number of binary decisions across the various rules and for the two data sets. The two sets of material were selected in such a way that the relative frequencies of potential and actual application of the rules correspond more or less to the relative frequencies of rule application in the training material. For the /@/-deletion and /@/-insertion rules, the relative frequencies of potential application are higher than the frequencies in the training material.

Table 2
Mapping of IPA-to-SAMPA symbols

IPA	n, m*	r, ʀ, ʁ, ɾ, ɹ, ɽ, ɻ	t	ə, ɜ
SAMPA	n	r	t	@

* The /m/ is only allowed in case of nasal assimilation.

Table 3
Example of pronunciation variants and corresponding binary scores

Pronunciation variant	Rules 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

Table 4
Numbers of binary decisions in the transcription material

Material	/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins	All
Majority vote	155	127	84	53	48	467
Consensus	287	230	109	41	103	770
Total	442	357	193	94	151	1237

However, this was necessary in order to obtain a sufficiently high number of observations for analysis.

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

$$\text{Cohen's } \kappa = \frac{P_o - P_c}{100 - P_c}, \quad -1 < \kappa < 1, \tag{1}$$

where P_o is the percentage observed agreement, P_c is the percentage chance agreement, and

$$P_o = 100\% \times \frac{\text{No. of agreements}}{\text{No. of agreements} + \text{No. of disagreements}}. \tag{2}$$

From the definition above it follows that $\kappa = 0$ indicates chance agreement, and $\kappa = 1$ indicates perfect agreement. A more detailed indication of how the κ-values should be interpreted is provided in Landis and Koch (1977): below 0.0, poor; 0.00–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect.

The APTs were compared to the two types of RTs (majority vote and consensus), and κ-values were calculated. Although the absolute κ-values were somewhat different for the two types of RTs, the general trends that we observed were similar. For reasons of brevity and clarity we therefore present only κ-values for all of the transcription material (see row 4 in Table 4). Furthermore, the results are presented for all rules pooled together. If necessary, we point out in what respect the results are rule dependent. For the sake of completeness, the κ-values per rule are given in Appendix B.

3. Results

The first aim of this investigation was to determine how various properties of CSRs affect the quality of APTs. The properties of the CSR that were investigated are all related to the HMMs. The general procedure was to take our baseline CSR and to substitute the baseline HMMs with a different set of HMMs in which each of the following properties was 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)–(3) (Section 3.4)

Each subsection in this section starts with a description of the investigated property of the CSR. Next, the agreement values are presented. Finally, each subsection 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 than human listeners do. Fig. 1 shows the percentages 'phone present' in the human RTs and in the APTs made with the baseline HMMs (see also Appendix A). Fig. 1 shows that for all rules the listeners tend to detect more phones than the CSR. Especially, for the /@/-deletion and /@/-insertion rules the differences are large.

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 listeners 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 consist of six states of which three can be skipped, the topology length of the baseline /@/ HMM is three states, or 30 ms.

Brugnara, Falavigna, and Omologo (1993) pointed out that topology length is a critical point for the automatic segmentation of speech. In order to investigate the optimal topology length, these authors compared various HMM topologies, with phone recognition rate as an evaluation criterion. They found an optimal accuracy with HMMs that have a minimum duration of 20 ms. This result corroborates our expectation that our HMM topology length of 30 ms is suboptimal for the task of automatic transcription.

As the results of Wester et al. (2001) and Brugnara et al. (1993) show that topology length is important for automatic transcription, we decided to investigate the effect of changing the HMM topology length. As the differences in the numbers of phones that are detected by the CSR and by the listeners are largest for the rules concerning /@/, we decided to focus on the topology length

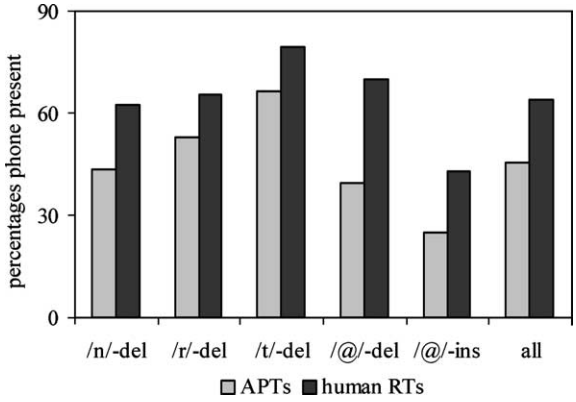

Fig. 1. Percentages 'phone present' for APTs versus human RTs.

for the /@/ HMM. In order to determine the duration of the phone /@/ in the training material, the /@/ had to be present in the transcriptions used for segmentation. Therefore, the /@/-insertion rule was applied to the baseline transcriptions. Subsequently, we made an automatic segmentation of the training material using the baseline HMMs. On the basis of this segmentation, we determined the number of frames that was assigned to each /@/. Next, we divided the /@/s into two categories:

1. *Short* /@/. The duration in the segmentation was exactly 3 frames, or 30 ms (1796 /@/s).
2. *Long* /@/. The duration in the segmentation was >3 frames, or longer than 30 ms (18,640 /@/s).

Two sets of HMMs were trained. For the first HMM set, all short /@/s were used to train an HMM consisting of one segment (two identical states of which one can be skipped), or with a topology length of 10 ms. For the second HMM set, the short /@/s were used to train an HMM consisting of two segments, or with a topology length of 20 ms. For both model sets, the long /@/s were used for training an HMM consisting of three segments.

We expect that by using the short /@/ HMM, more /@/s will be detected by the CSR. Table 5 shows the percentage of /@/s that were denoted as 'present' by the CSR. The following abbreviations are used: '3 seg' denotes the baseline HMMs with a three-segment topology for the phone /@/, '2 seg' denotes the two-segment topology, and '1 seg' denotes the one-segment topology. In Table 5, it can be seen that the percentages '/@/ present' indeed increased when using the short-@/ HMM.

Furthermore, we expect that by using the short /@/ HMM the agreement values will increase for the /@/-deletion and /@/-insertion rule. As the number of extra detected /@/s was larger for the "1 seg" HMMs compared to the "2 seg" HMMs, one should expect that the agreement value is also larger for the "1 seg" HMMs compared to the "2 seg" HMMs. Fig. 2 shows that this was not the case, which means that not all of the extra /@/s increased agreement. Furthermore, Fig. 2

Table 5
percentages '/@/ present' for HMMs with various topology lengths and human RTs

	APTs (%)			Human RTs (%)
	3 seg	2 seg	1 seg	
/@/-deletion	39	54	65	70
/@/-insertion	25	32	34	43

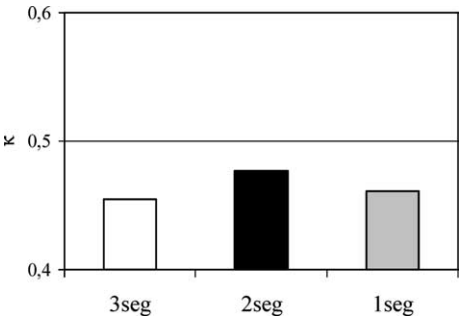

Fig. 2. Agreement values for HMMs with various topology lengths.

shows that the optimal HMM topology length is two segments, or 20 ms. This result is in line with the results of Brugnara et al. (1993), since they also found an optimal topology length of 20 ms.

To conclude, using a shorter topology length for the phone /@/ decreased the discrepancy between the number of phones detected by listeners and CSR. Furthermore, the highest agreement value was found for a short /@/ HMM with a topology length of 20 ms.

3.2. Degree of contamination of the HMMs

The speech material used for training contains a lot of 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. Because of these mismatches between the transcriptions of the training material and the pronunciation, the HMMs become 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. 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 (2000) reported that this is indeed the case: baseline HMMs that were trained on canonical transcriptions produced more canonical APTs than HMMs that were trained on the basis of automatic or manual transcriptions of the training material in which pronunciation variation has been transcribed.

In this section, we 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 were trained on read speech material.

3.2.1. Modeling pronunciation variation

One approach that we used to minimize the mismatch in the training corpus was modeling pronunciation variation (Wester, Kessens, & Strik, 1998). In our previous research, we used an approach whereby automatic transcriptions were made by means of forced recognition. The new automatic transcriptions were then used to train new HMMs. For the current experiments, we took the following two sets of HMMs from our research on pronunciation variation modeling and used them in addition to the baseline HMMs for making automatic transcriptions:

1. HMMs trained on a corpus for which within-word pronunciation variants were transcribed ('within HMMs'). These variants were automatically obtained 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 was transcribed ('within + cross HMMs'). For more details on the cross-word variation modeled, see Wester et al., 1998.

Fig. 3 shows that the total agreement values increase if the HMMs from pronunciation variation research ('within' and 'within + cross') are used instead of the baseline HMMs (base). These results are in line with the findings of Saraçlar (2000) on the Switchboard corpus. Another observation from Saraçlar (2000) is that the pronunciation variation HMMs are less biased towards the canonical transcriptions than the baseline HMMs. Closer inspection of our data revealed that, as expected, also in our material the CSR tends to choose canonical transcriptions less often

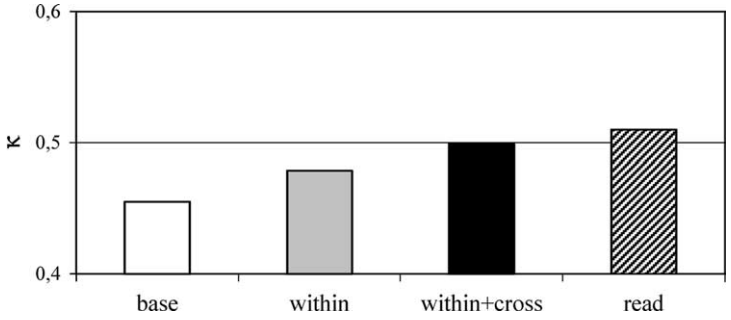

Fig. 3. Agreement values for the baseline HMMs and for HMMs from pronunciation variation research.

(22–23% less often) when using the HMMs from pronunciation variation research compared to when using the baseline HMMs (see Appendix A).

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, in read speech, fewer mismatches should be found between the speech signal and the transcriptions. Thus, it is to be expected that HMMs trained on read speech will be less contaminated than those trained on spontaneous speech. Since it was shown in the previous section that less contaminated HMMs could 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, Boogaart, Boves, & Klabbers, 1995) containing about twice as many words as the VIOS training material. Both corpora contain telephone speech. Fig. 3 shows that the total agreement values are higher when we use HMMs trained on read speech (read) than when we use HMMs trained on spontaneous speech (base).

The results presented in this section show that 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.

3.3. Context-independent versus context-dependent HMMs

As context-dependent HMMs (CD-HMMs) take account of the context in which a phone occurs, they 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 et al., 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, contamination of the HMMs causes a bias of the APTs towards the transcriptions on which the HMMs are trained. This means that if HMMs are trained on the basis of canonical transcriptions of the training material, the HMMs will produce APTs that are biased towards the canonical transcriptions. We hypothesize that the bias towards the canonical transcriptions is in some cases larger for the CD-HMMs than for the context-independent HMMs (CI-HMMs). To illustrate this point, let us consider the following example: suppose we train CD-HMMs on the basis of transcriptions of the VIOS training corpus in which

/r/-deletion is not accounted for. In these transcriptions, 30,018 /r/s are transcribed, of which 1813 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 /r/-deletion rule can be applied. According to the human listeners, /r/-deletion is applied in about 35% of the cases (see Fig. 2); thus of the /r/s in the context /@rd/ about 35% are probably not pronounced. Consequently, if a CD-HMM is trained for /@rd/, then the /r/ is not present in 35% of the training tokens. This percentage corresponds to 2% of all /r/s in the training material. This means that if a CI-HMM is trained for the /r/, then the /r/ is not present in 2% of the training tokens. Consequently, 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/).

In order to investigate the effect of CD-HMMs on automatic transcription, state-tied CD-HMMs were 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 were clustered to train a CI-model for all middle segments of the same phone. Left and right CD-models were trained for clusters of first and last segments with equal left or right contexts. Each cluster of first and last segments contained at least 200 observations. All left and right contexts with fewer than 200 observations were 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 were trained. If we then look at the training corpus consisting of 326,494 phones, we observe that 94.3% of the left context tokens and 94.4% of the right context tokens are covered by the left and right CD-models, respectively.

As expected, the CD-HMMs were indeed more biased towards the canonical transcriptions than the CI-HMMs: 14% (97) more canonical transcriptions are chosen using CD-HMMs compared to using CI-HMMs. Fig. 4 shows that the agreement values are slightly lower for the CD-HMMs compared to the CI-HMMs. Closer inspection of the results per rule (see Appendix B) revealed that using CD-HMMs for automatic transcription causes a considerable decrease in the agreement value for the /r/-deletion and /@/-insertion rule.

To conclude, despite the fact that CD-HMMs yield lower WERs, when using this type of HMMs for obtaining automatic transcriptions, the quality of the transcriptions is not improved.

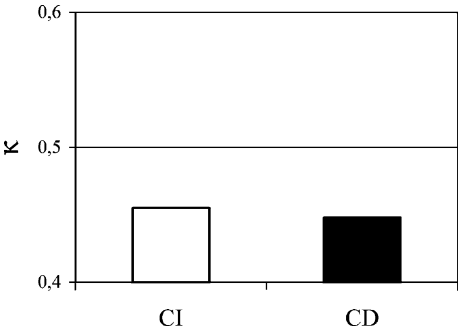

Fig. 4. Agreement values for CI- versus CD-HMMs.

3.4. Combinations of properties

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

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

First, we investigated 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 /@/ HMM 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 made 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 were used to train a new set of HMMs.

Fig. 5 shows the agreement values for the baseline HMMs ('base'), for changing the separate properties ('short /@/' and 'pron.var.'), and for changing the two properties simultaneously ('combi'). It can be seen that the combination of the two properties results in a higher agreement value than each property separately.

Closer inspection of the results per rule (see Appendix B) shows that – compared to the baseline HMMs – agreement is largely improved for the /n/-deletion rule and the /@/-insertion rule. The increase in agreement for the /n/-deletion rule can mainly be attributed to pronunciation variation modeling as for this rule a large increase in agreement was found using the pronunciation variation HMMs. The rule that especially benefits from the combination of the two properties is the /@/-insertion rule: the combination result is much better than the results for the individual properties.

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) combined with CD-HMMs (Section 3.3). Due to

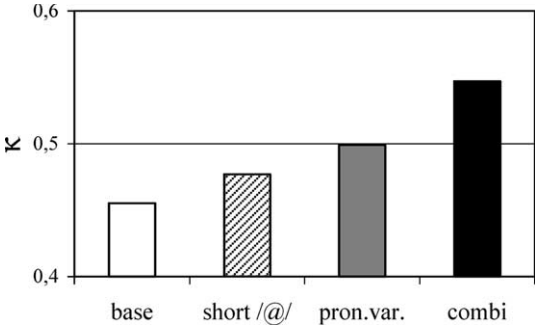

Fig. 5. Agreement for the combination of pronunciation variation modeling and short /@/ HMMs.

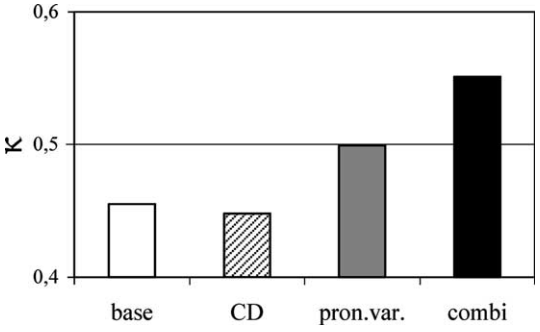

Fig. 6. Agreement values for the combination of pronunciation variation modeling and CD-HMMs.

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 should be less contaminated.

In order to train the combination HMMs, we made 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 were trained (see Section 3.3 for more details on the state-tying procedure).

Fig. 6 shows the agreement values for the pronunciation variation HMMs ('pron.var.'), CD-HMMs ('CD') and the combination of pronunciation variation modeling, and CD-HMMs ('combi'). The combination of the two properties results in a higher agreement value than the agreement value for each property separately.

Closer inspection of the results per rule (see Appendix B) reveals that CD-HMMs can indeed benefit from pronunciation variation modeling: the decrease in agreement values that was found for the /r/-deletion and /@/-insertion rule disappears. Furthermore, again, for the /@/-insertion rule the combination results are much better than the individual results. For the /@/-deletion rule, the combination result is not improved compared to the result obtained when using the CD-HMMs. This result can be explained as follows: as the automatic transcriptions of the /@/-deletion variants were obtained with the baseline HMMs and as low agreement values were found for the /@/-deletion rule [the κ-values are qualified as 'slight' and 'fair' for the baseline HMMs (Landis & Koch, 1977)], pronunciation variation modeling might have a negative effect on the context-dependent modeling.

4. Agreement and WER

In most research on APT, the choice of the speech recognizer is usually not an explicit issue under investigation. Most probably, one generally takes the speech recognizer with the lowest WER. Obviously, the underlying assumption 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 relationship between WER on the one hand, and the agreement values between the APTs and human RTs on the other hand. We measured WERs on the total

transcription material (majority vote + consensus) for all sets of HMMs that are used in this paper. The lexicon used in the recognition experiments contains 1154 words, to which 1119 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 was employed that distinguishes between different variants of the same word. For more details on this kind of language model, see Kessens, Wester, and Strik (1999). The WER is defined as follows:

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100\%, \tag{3}$$

where S is the number of substitutions, D the number of deletions, I the number of insertions, and N the total number of words. In Fig. 7, the scatter plot of κ as a function of WER is given.

One would expect that a lower WER yields a higher κ. Fig. 7 shows that this relationship 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. However, at this point it should be noted that we do not think that WER and κ are completely uncorrelated. In order to make good quality transcriptions, a certain level of recognition performance is necessary, and vice versa, a CSR that performs badly in a conventional recognition task will not be very useful for making high-quality automatic transcriptions.

Saraçlar (2000) and Saraçlar et al. (2000) also reported results showing that better transcription accuracy does not imply that the WER also improves. 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

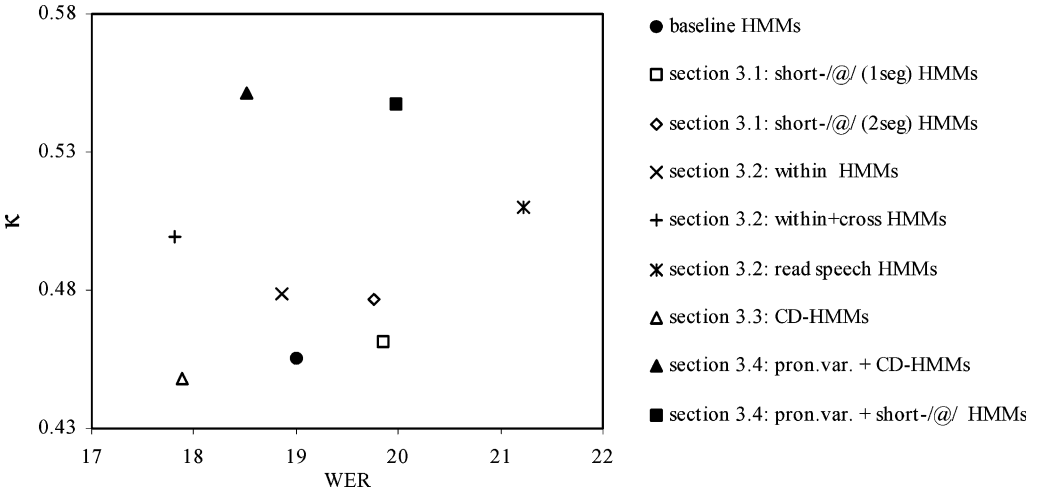

Fig. 7. Scatter plot of total κ and WER on transcription data.

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 another factor also plays a role. The sequences of phones that can be recognized during a conventional recognition task are constrained by the lexicon, the language model, and recognition parameters (e.g., word insertion penalties). Through these (word-level) constraints, it is impossible to recognize other sequences of phones than the phone sequences in the lexicon. During automatic transcription, however, the choice for a specific transcription variant is determined by the acoustics (this research) or by a combination of acoustics, phonotactic constraints, and/or insertion and transition penalties. These phone-level constraints are different from the word-level constraints that are used during a conventional (word) recognition task.

These results illustrate that recognition and automatic transcription are different tasks and the tools to perform these tasks should be optimized in different ways. For this reason, for making optimal automatic transcriptions 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 speech recognizer should have.

5. 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 reference transcriptions; the overall κ-values vary between 0.44 and 0.55. Our results indicate that the quality of the automatic transcriptions can be improved by using 'short' HMMs and by reducing the amount of contamination due to pronunciation variation. The latter can be achieved by using HMMs from pronunciation variation modeling research, or by using HMMs trained on read speech. We also compared CI-HMMs with CD-HMMs. We found that the κ-values obtained with CD-HMMs trained on baseline transcriptions are lower than the κ-values for CI-HMMs. A reason for this decrease in κ-values could be that the baseline transcriptions often do not provide an adequate description of the realized pronunciation. This mismatch can be reduced by using pronunciation variation modeling. And indeed, the κ-values obtained with CD-HMMs trained on these improved transcriptions (resulting from pronunciation variation modeling research) were higher. In fact, the κ-values obtained for this combination of properties were higher than the κ-values for each individual property (pronunciation variation modeling and CD-HMM). The same result (κ-values for combination higher than for each individual property) was also found for pronunciation variation modeling in combination with 'short' HMMs. This is a promising result, since it indicates that by taking the right combination of properties a substantial improvement in the κ-values might be obtained.

Finally, we observed that there is no clear relationship between the WER of a CSR and the κ-values. Therefore, we can conclude that using the CSR with the lowest WER is not always the best solution for obtaining optimal automatic transcriptions. In order to obtain high-quality automatic transcriptions, one could start with using HMMs that are trained on read speech. However, if the automatic transcriptions need to be further improved, specialized CSRs should be developed which are optimized for the transcription task.

Acknowledgements

The research by Judith M. Kessens was carried out within the framework of the Priority Programme Language and Speech Technology, funded 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 A^2RT and two anonymous reviewers for their useful comments on a previous version of this paper.

Appendix A. Numbers of phone 'present' in the transcriptions

In the canonical transcriptions, the only rule (of the five rules) that is applied is the /n/-deletion rule, since the pronunciation without the /n/ is considered to be the most likely one according to the linguistic literature (van de Velde, 1996). Therefore, the phone /n/ in the 442 cases of the /n/-del rule is not present in the canonical transcription. Obviously, also for the 151 cases of the /@/-ins rule, the phone /@/ is not present in the canonical transcription.

Both man (human RTs) and machine (APTs) decided for the 1237 cases under study whether the target phone was present or not. For the /r/-del, /t/-del, and /@/-del rules, deciding that the phone is present implies choosing the canonical variant. However, for the /n/-del and /@/-ins rules it is exactly the other way round: deciding that the phone is present implies choosing the NON-canonical variant (in these cases NOT present is equivalent to canonical). Therefore, in Table 6, also the numbers of canonical transcriptions are given between brackets for the /n/-deletion and /@/-insertion rule, and for all rules. In the last row of Table 6, the numbers of 'phones present' are given for the human RTs.

"Percentages of 'phone present' for APTs versus human RTs" presented in Fig. 1 were calculated as follows (using numbers from the first and last row of Table 6, and the numbers in the last row of Table 4):

Table 6
Numbers of 'phone present' and canonical transcriptions for all sets of HMMs

APTs		/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins	All rules
Section	HMMs						
3.1	3 seg (baseline)	193 (249)	189	128	37	38 (113)	585 (716)
3.1	2 seg	213 (229)	202	125	51	49 (102)	640 (709)
3.1	1 seg	216 (226)	231	130	61	52 (99)	690 (747)
3.2	within	236 (206)	167	119	34	51 (100)	607 (626)
3.2	within + cross	238 (204)	167	122	34	49 (102)	610 (629)
3.2	/@n#/	291 (151)	199	132	37	38 (113)	697 (632)
3.2	read speech	260 (182)	189	125	40	56 (95)	670 (631)
3.3	CD	233 (209)	279	143	54	23 (128)	732 (813)
3.4	pron.var. & short /@/	247 (195)	184	125	53	70 (81)	679 (638)
3.4	pron.var. & CD	253 (189)	191	135	38	54 (97)	671 (650)
Human RTs		276 (166)	233	153	66	65 (86)	793 (704)

/n/-del – APT: 193/442 = 43.7%
/n/-del – RT: 276/442 = 62.4%
/r/-del – APT: 189/357 = 52.9%
/r/-del – RT: 233/357 = 65.3%
⋮
all – APT: 585/1237 = 47.2%
all – RT: 793/1237 = 64.1%

Appendix B. Agreement values per rule for all sets of HMMs

Agreement values per rule for all sets of HMMs

Section	HMMs	/n/-del	/r/-del	/t/-del	/@/-del	/@/-ins	All
3.1	3 seg (baseline)	0.55	0.41	0.45	0.24	0.42	0.45
3.1	2 seg	0.61	0.36	0.47	0.32	0.42	0.48
3.1	1 seg	0.60	0.35	0.41	0.30	0.40	0.46
3.2	within	0.65	0.40	0.40	0.20	0.47	0.48
3.2	within + cross	0.67	0.41	0.45	0.23	0.47	0.50
3.2	/@n#/	0.64	0.38	0.51	0.28	0.39	0.48
3.2	read speech	0.68	0.42	0.42	0.20	0.55	0.51
3.3	CD	0.56	0.31	0.48	0.46	0.33	0.44
3.4	pron.var. & short /@/	0.71	0.38	0.45	0.35	0.67	0.55
3.4	pron.var. & CD	0.71	0.43	0.51	0.25	0.60	0.55

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, 213220.

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 and Speech, vol. 18, University of Nijmegen, pp. 107–119.

Kessens, J.M., Wester, M., Cucchiarini, C., Strik, H., 1997. Testing a method for modelling pronunciation variation, In: Proceedings of the COST Workshop, Rhodos, pp. 37–40.

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., 2003. A data-driven method for modeling pronunciation variation. Speech Communication 40 (4), 517–534.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

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., 2000. Pronunciation modeling for conversational speech. Ph.D. thesis, John Hopkins University, Baltimore, ML.

Saraçlar, M., Nock, H., Khudanpur, S., 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer, Speech and 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 withinword and crossword 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. Language and Speech 44 (3), 377–403.