*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 A^2RT 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, A^2RT, 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

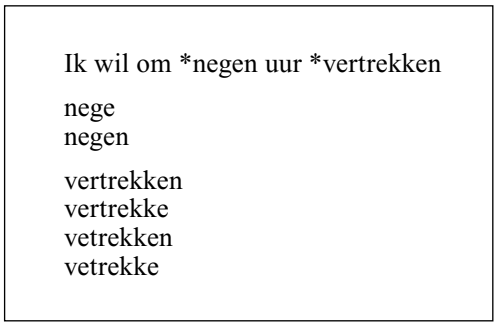

Figure 1

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:

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

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.

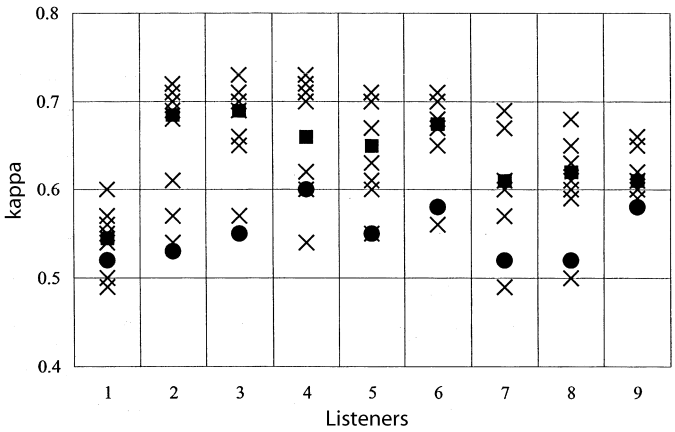

Figure 2
Cohen's κ for the agreement between the CSR and each listener (●), for listener pairs (×) and the median of the listeners (■)

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).

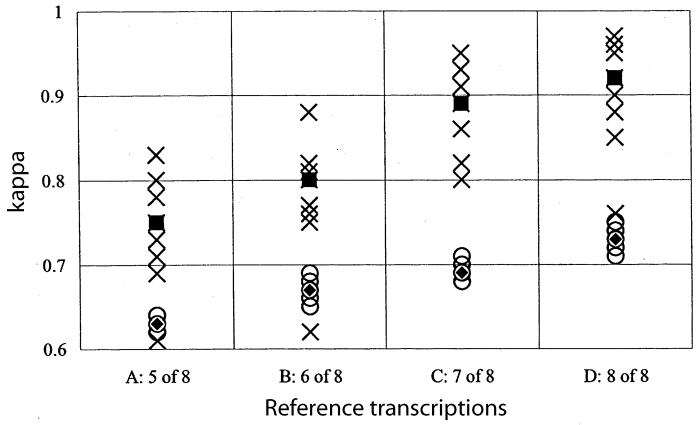

Figure 3

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

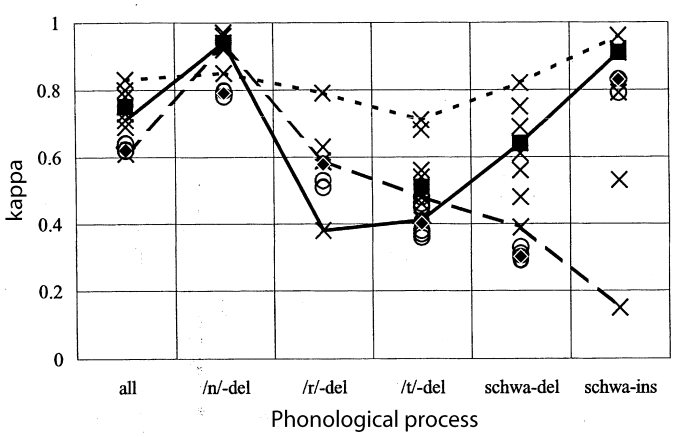

Figure 4
Cohen's κ for the listeners and the CSR compared to the sets of reference transcriptions (5 of 8) for the various phonological processes (○ = CSR, × = listener, ■ = median listeners, ◆ = median CSR, dotted line = listener A, solid line = listener B, and dashed line = listener C)

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.

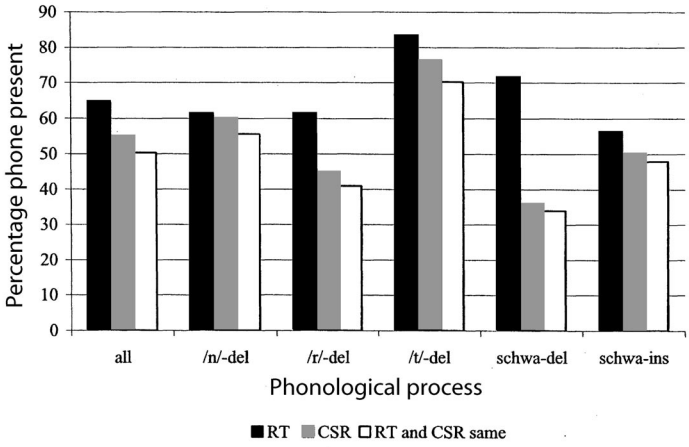

Figure 5

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

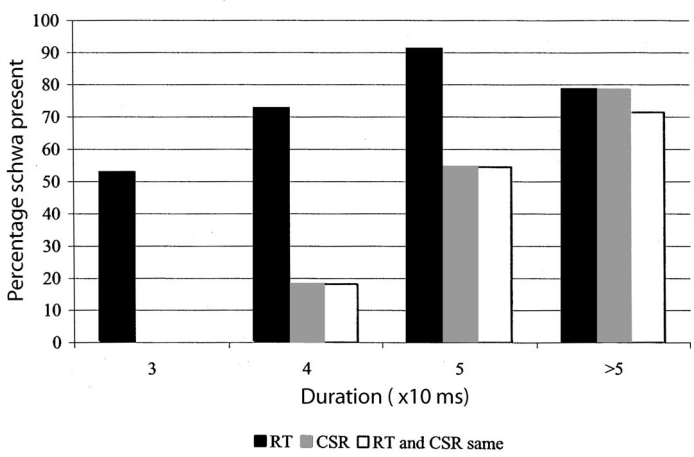

Figure 6

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

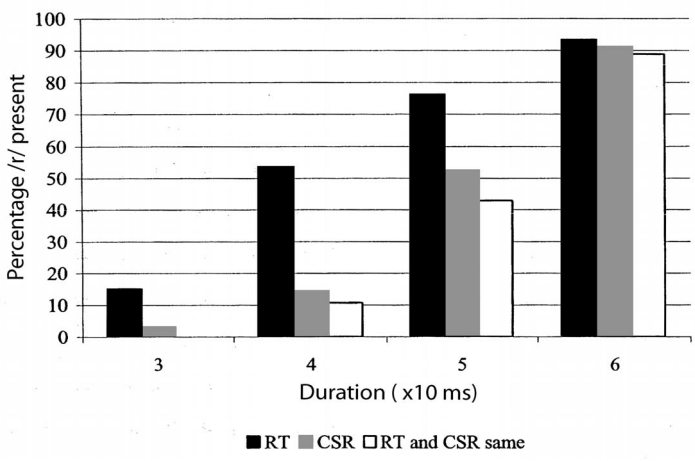

Figure 7

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, ɜ̆, ə̲, ə̟, əˠ, and other vowels: ɛ̆, ɜ̈. There were fewer transcriptions pertaining to schwa-insertion, viz.: not present: ø, different realizations of schwa: ə, ɜ̆ 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, ɜ̆, ə̲, ə̟, əˠ, ɛ̆, ɜ̈, ɐ, ɪ]

II. deletions (ø) and short schwas (ɜ̆) are classified as "phone not present" and the rest is classified as "phone present": [ə, ə̲, ə̟, əˠ, ɛ̆, ɜ̈, 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	ɜ̆	ə̲	ə̟	əˠ	ɛ̆	ɜ̈	**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	ɜ̆	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

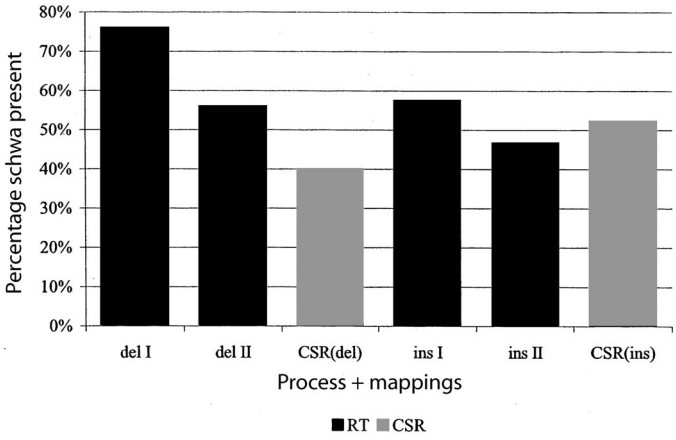

Figure 8

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)