SELECTION OF PRONUNCIATION VARIANTS IN SPONTANEOUS SPEECH: COMPARING THE PERFORMANCE OF MAN AND MACHINE

Mirjam Wester, Judith M. Kessens, Catia Cucchiarini & Helmer Strik
A^2RT, Dept. of Language & Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
{wester, kessens, catia, strik}@let.kun.nl, http://lands.let.kun.nl/

RESUME

Dans cet article, les performances d'un outil de transcription automatique sont évaluées. L'outil de transcription est un reconnaisseur de parole continue (CSR) fonctionnant en mode de reconnaissance forcée. Pour l'évaluation les performances du CSR ont été comparées à celles de neuf auditeurs experts. La machine et l'humain ont effectué exactement la même tâche: décider si un segment était présent ou non dans 467 cas. Il s'est avéré que les performances du CSR étaient comparables à celle des experts.

1. INTRODUCTION

Recently, various authors have pointed out that since much of the work in linguistic research has been based on laboratory speech, it is questionable whether the knowledge gathered so far generalizes to less formal types of speech, like spontaneous speech [1, 2]. This feeling has generated a growing interest in studying spontaneous speech. Moreover, the fact that large databases of spontaneous speech have now been created for the purpose of Automatic Speech Recognition (ASR) has given the impression that analyzing spontaneous speech is within reach of any linguist. However, the way in which information is stored in such databases is not always the most suitable representation for linguistic research. In other words, adequate instruments are needed to make it possible for linguists to access and effectively use the speech data contained in these databases.

The type of representation that is probably used most often in linguistic research is phonetic transcription. So in order to be useful for linguistic research, the speech material contained in the databases should be available in the form of phonetic transcriptions. However, since making phonetic transcriptions is extremely time-consuming, linguists often decide not to transcribe whole utterances, but only those parts of the utterance where the phenomenon under study is expected to take place. Even with this restriction, making phonetic transcriptions remains a tedious and costly task. Therefore, it would seem that developing an instrument for automatically transcribing speech would contribute to facilitating linguistic research and to making the large spontaneous speech databases accessible to many linguists.

In ASR, tools have been developed that go some way toward obtaining adequate phonetic representations of speech in an automatic manner. In order to find out whether these tools are useful to obtain phonetic transcriptions automatically, their performance should be studied. However, this is not straightforward because, as for human phonetic transcription, it is impossible to obtain a reference representation that can be assumed to be correct [3: pp. 11-13]. The most usual procedure is to take a consensus transcription [4] as the reference. A consensus transcription is made by a group of transcribers after they have reached a consensus on each transcribed symbol. Another possibility consists in having several transcribers transcribe the same material, and in constructing a reference transcription on the basis of the response of the various transcribers, by using a 'majority vote' procedure. The latter procedure will be adopted in this study. By comparing the automatically obtained transcriptions with the reference transcriptions, it is possible to determine whether the automatic transcription tool performs satisfactorily.

The aim of this paper is to show that the performance of our automatic transcription tool compares to that of expert linguists who carried out the same task, and that therefore, this tool can be used to obtain information on spontaneous speech processes from speech databases.

2. METHOD

In this experiment, a number of utterances were judged both by a panel of expert linguists and by a CSR. Both the linguists and the CSR had to carry out the same task: selecting the variant that had been realized for some of the words contained in the utterances.

2.1. Phonological Rules

For the current experiment, pronunciation variants were generated with five phonological rules concerning the following speech processes: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion (we use Sampa notation in this paper). All these rules describe either insertion or deletion processes (i.e. alterations in the number of segments) within words. The main reasons for selecting these five phonological rules are that they are frequently applied in Dutch and are well described in the literature. A more detailed description of the phonological rules can be found in [5, 6]. These rules were used to automatically generate pronunciation variants for the words being studied. Sometimes, more than one rule could apply in the same word. However, in selecting the speech material we decided to limit the number of rules which could apply in one word to two, in order not to make the task too complex for the listeners.

2.2. The Speech Material

The speech material used in this experiment was selected from a database named VIOS, which contains a large number of telephone calls recorded with an on-line version of the spoken dialogue system called OVIS [7]. OVIS is employed to automate part of an existing Dutch public transport information service. Currently, OVIS can be used to obtain information about Dutch train times. The speech material consists of interactions between man and machine, and therefore, it can be described as spontaneous speech.

From the VIOS corpus, 186 utterances were selected, which contain 379 words to which one or two rules apply. For 88 words two rules applied and four pronunciation variants were generated. For the other 291 words only one rule applied and two variants were generated. Consequently, the total number of instances in which a rule could be applied is 467 (/n/-del: 155, /r/-del: 127, /t/-del: 84, /@/-del: 53, /@/-ins: 48).

2.3. Experimental Procedure

Nine listeners and the CSR carried out the same task, i.e. deciding for the 379 words which variant best matched the word that had been realized in the spoken utterances (forced choice). For 88 words four variants were present, as mentioned above. For each of these words two binary scores were obtained, i.e. for each of the two underlying rules it was determined whether it was applied (1) or not (0). For each of the remaining 291 words with two variants one binary score was obtained. Thus, 467 binary scores were obtained for each listener and for the CSR.

The nine expert linguists 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 for the kind of people that may have to make phonetic transcriptions and that can be interested in automatic ways of obtaining such transcriptions from spontaneous speech databases.

The 186 utterances were presented to the listeners 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 a screen. 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 presented phonemic transcriptions best corresponded to the spoken word. The listener had the possibility of listening to an utterance as often as he/she felt was necessary in order to judge which pronunciation variant had been realized.

The utterances presented to the listeners were also used as input for the CSR, which is part of the spoken dialogue system OVIS [7]. In this CSR, one context-independent HMM is used for most phonemes, except for the /l/ and the /r/, for which separate models are trained for prevocalic and postvocalic position in the syllable. For automatic transcription purposes, the CSR is used in forced recognition mode, which means that the recognizer does not choose between all the words in the lexicon, but only between the different pronunciation variants of the same word. In this way, the CSR carries out the same task as the listeners, i.e. for each of the 379 words it determines which of the variants presented best matches the actual realizations.

The phone models we used were iterated models, which means they were trained on a corpus in which pronunciation variants of the five phonological rules had been added by means of a forced recognition. For a more detailed description of this iterative process see [8].

3. RESULTS

In order to determine whether the CSR performs in a way that is comparable with that of the nine listeners, two types of analyses were conducted. First, we checked whether the degree of agreement between the CSR and the nine listeners is comparable to that computed for the various listener pairs (section 3.1.). Second, on the basis of the responses of the nine listeners a reference transcription was composed. Subsequently, the responses of the CSR and those of the nine listeners were compared with the reference transcription. A comparison was made for all rules together (section 3.2.), and for each of the rules separately (section 3.3.).

3.1. Percentage Agreement

For all pairs of listeners, a percentage agreement score was calculated. Subsequently, the percentage of agreement between each of the nine listeners and the CSR was also calculated. The results are presented in Fig. 1. For instance, shown in 'column 1' are the percentage agreement scores of listener 1 with the CSR (■), with the other 8 listeners (x), and the average of these 8 between-listener agreement scores (●).

Percentage agreement for the listener pairs varies between 75% and 87%, and the average over all listener pairs is 82%. The average agreement over the nine listener-CSR pairs is 78%. So on average, the degree of agreement between the CSR and the listeners is only 4% lower than the degree of agreement between the listeners. In order to test the influence of the phone models, we repeated this type of analysis with the baseline phone models, i.e. phone models which were trained on a corpus in which no variants were added (see [8]). The results obtained with these phone models show that the average agreement between the CSR and the listeners is 74%. Since this is 4% lower than for the iterated phone models, we decided not to use the baseline phone models for the other tests.

In Fig. 1, it can also be seen that for each of the nine listeners percentage agreement with the CSR is lower than the average percentage agreement between listeners, however, the differences are small. In four cases, the CSR score is within the listener range (i.e. for listeners 1, 4, 7 and 8), and in the remaining five cases, the CSR score is maximally 2% below the range.

To summarize, these analyses show that although percentage agreement between the listeners and the machine is lower than percentage agreement between the listeners, the differences are so small that we can conclude that the performance of the CSR is comparable to that of the listeners.

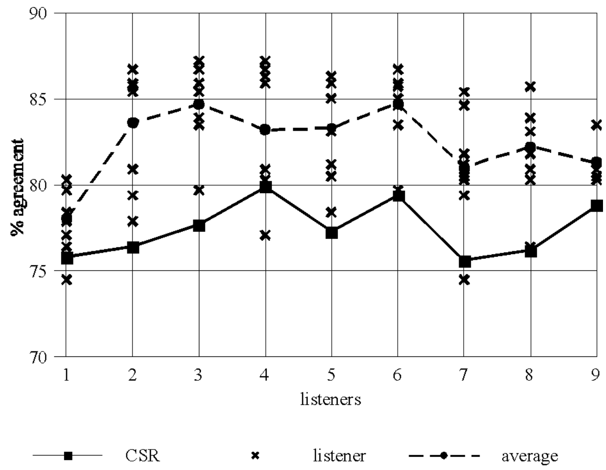

Figure 1: Percentage agreement between the CSR and each listener, and between all listener pairs plus an average over all listeners.

3.2. Reference Transcriptions for All Rules

On the basis of the responses of the nine listeners, a reference transcription was composed by using a 'majority vote' procedure. When nine listeners are involved, as in this experiment, a reference transcription of this kind can be made by using different degrees of strictness: ❶ a majority of at least 5 out of 9, ❷ 6 out of 9, ❸ 7 out of 9, ❹ 8 out of 9 and, eventually, by taking only those cases in which ❺ all nine listeners agree. It is obvious that in going from 1 to 5, the number of cases involved is reduced (1: 467, 2: 435, 3: 385, 4: 335, 5: 246). Furthermore, it is to be expected that if we compare the performance of the CSR with the reference transcriptions of type ❶, ❷, ❸, ❹, and ❺, the degree of agreement between the CSR and the reference transcription will also increase when going from 1 to 5. The rationale behind this is that the cases for which a greater number of judges agree should be easier to judge than the other ones. Therefore, it can be expected that they should be easier for the CSR too.

In Fig. 2, we see that the degree of agreement between the reference transcriptions and the listeners is higher than that between the reference transcriptions and the CSR. This is not surprising if we consider that the reference transcriptions are based on the listeners' responses and not on those of the CSR. In Fig. 2, we also see that percentage agreement between the CSR and the reference transcription gradually increases from 81% to 90%, as expected. We may therefore conclude that the CSR shows similar behavior to the humans in the sense that for cases in which the agreement between listeners is higher, the agreement of the listeners with the CSR is also higher.

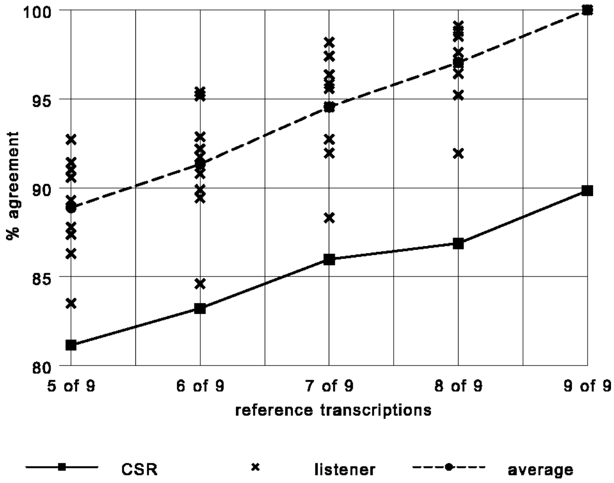

Figure 2: Percentage agreement between listeners and the various reference transcriptions, and between CSR and the reference transcriptions

3.3. Reference Transcription for Various Phonological Rules

In the previous section, we have compared the various reference transcriptions with the responses of the nine listeners and those of the CSR for all the cases pooled together. However, it is possible that the CSR and the nine listeners perform differently for the various phonological rules. Therefore, we will now break down the results for the five phonological rules. Since chance agreement differs for the various conditions, percentage agreement is not the most suitable measure to compare between the rules. That is why for this comparison we used Cohen's κ, in which a correction for chance agreement is made [9]:

$\kappa = (P_o - P_c) / (1 - P_c)$

P_o = observed proportion of agreement

P_c = proportion of agreement on the basis of chance

In order to calculate Cohen's κ, the reference transcription of type ❶ was used, i.e. the transcription obtained by taking the 'majority vote' of the nine listeners (5 out of 9). The results are shown in Fig. 3.

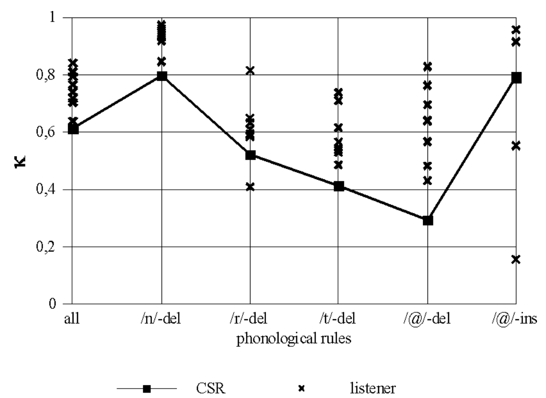

Figure 3: Cohen's κ for the listeners and the CSR compared to the reference transcriptions for the various phonological rules.

For each condition in Fig. 3, the degree of agreement between the reference transcription and the nine listeners (x) plus the CSR (■) is shown, first for all rules and then for the individual rules. As is clear from Fig. 3, the results do indeed differ for the five phonological rules. It is clear that both the CSR and the listeners perform best on the /n/-deletion rule. Furthermore, agreement is somewhat lower for the other three deletion rules, both for the CSR and the listeners. Finally, for /@/-insertion agreement is again higher for most listeners and the CSR. However, it can also be seen that for this rule the variability in the degree of agreement between the listeners is larger than the variability for the other rules. In general, it can be concluded that also for the individual rules the behavior of the CSR is similar to that of the listeners.

4. DISCUSSION AND CONCLUSIONS

The results presented in the previous section reveal that, for the task under study, the performance of the listeners and that of the CSR are similar, and that, on average, the degree of agreement between the CSR and the listeners is only slightly lower than that between listeners. This means that the automatic tool proposed in this paper can be used effectively to obtain phonetic transcriptions of deletion and insertion processes in spontaneous speech.

Although this tool cannot be used to obtain phonetic transcriptions of complete utterances from scratch, it clearly can be employed for hypothesis verification, which is probably the most common way of using phonetic transcriptions in various fields of linguistics, like phonetics, phonology, sociolinguistics, and dialectology. Another possible limitation of this tool is that so far it has been tested for deletions and insertions only, so that we do not know how it performs with substitutions.

However, in spite of these limitations an important contribution of automatic transcription to linguistics would be that it makes it possible to use existing speech database for the purpose of linguistic research. The use of these databases has at least two important advantages [10]. First, many of these databases contain spontaneous speech, a type of speech that is very under-researched at present. Second, these databases contain large amounts of speech material. 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.

At this point, it is important to note that in the current experiment we simply employed the CSR which we use in our ASR research. We did not try to adapt our CSR so as to make its transcriptions more similar to the human transcriptions. Still, the transcriptions made by the CSR do depend on the properties of the CSR. For instance, in section 3.1 we showed that by using other phone models, the average agreement between CSR and listeners drops from 78% to 74%. In the near future, we intend to study the effect of the CSR properties on the produced transcriptions. In this way, we hope to improve the quality of the automatic transcriptions.

To conclude, in this paper we have presented a tool that can be used effectively to obtain automatic transcriptions of deletion and insertion processes. Future research will indicate whether this tool can be used for other processes and whether its performance can be improved. For the time being, an instrument is available that makes it possible for linguists to access and effectively use the speech material contained in large speech databases for studying a number of spontaneous speech processes.

5. ACKNOWLEDGEMENTS

The research by Judith M. Kessens was carried out within the framework of the Priority Programme Language and Speech Technology, sponsored by NWO (Dutch Organisation for Scientific Research). The research by Helmer Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. We thank Lou Boves for thoroughly reviewing an earlier version of this paper.

6. REFERENCES

[1] J. Rischel, "Formal linguistics and real speech", *Speech Communication*, Vol. 11, pp. 379-392, 1992.

[2] M. Swerts, R. Collier, "On the controlled elicitation of spontaneous speech", *Speech Communication*, Vol. 11, pp. 463-468, 1992.

[3] C. Cucchiarini, "Phonetic transcription: a methodological and empirical study", PhD thesis, University of Nijmegen, 1993.

[4] L.D Shriberg, J. Kwiatkowski, and K. Hoffman, "A Procedure for Phonetic Transcription by Consensus", *Journal of Speech and Hearing Research*, Vol. 27, pp. 456-465, 1984.

[5] G. Booij, "The Phonology of Dutch", Clarendon Press, Oxford, 1995.

[6] C. Cucchiarini and H. van den Heuvel, " */r/ Deletion in Standard Dutch*", Proc. of the Dept. of Language & Speech, University of Nijmegen, Vol. 19, pp.59-65, 1995.

[7] H. Strik, A. Russel, H. van den Heuvel, C. Cucchiarini, and L. Boves, "A Spoken Dialogue System for the Dutch Public Transport Information Service", *Int. Journal of Speech Technology*, Vol. 2, No. 2, pp. 119-129, 1997.

[8] M. Wester, J.M. Kessens, and H. Strik, "*Improving the Performance of a Dutch CSR by Modeling Pronunciation Variation*", Proc. of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition, pp. 145-150, Kerkrade, 1998.

[9] T. Rietveld and van R. van Hout, "Statistical techniques for the study of language and language behaviour", Mouton de Gruyter, Berlin, 1993.

[10] P. Ladefoged, "Knowing enough to analyze spoken languages", *Speech Communication*, Vol. 11, pp. 1-4, 1992.