COMPARISON BETWEEN EXPERT LISTENERS AND CONTINUOUS SPEECH RECOGNIZERS IN SELECTING PRONUNCIATION VARIANTS

Mirjam Wester, Judith. M. Kessens, Catia Cucchiarini & Helmer Strik
A²RT, Dept. of Language & Speech, University of Nijmegen, The Netherlands
{M.Wester, J.Kessens, C.Cucchiarini, W.Strik}@let.kun.nl, http://lands.let.kun.nl/

ABSTRACT

In this paper, the performance of an automatic transcription tool is evaluated. The transcription tool is a continuous speech recognizer (CSR) which can be used to select pronunciation variants (i.e. detect insertions and deletions of phones). The performance of the CSR was compared to a reference transcription based on the judgments of expert listeners. We investigated to what extent the degree of agreement between the listeners and the CSR was affected by employing various sets of phone models (PMs). Overall, the PMs perform more similarly to the listeners when pronunciation variation is modeled. However, the various sets of PMs lead to different results for insertion and deletion processes. Furthermore, we found that to a certain degree, word error rates can be used to predict which set of PMs to use in the transcription tool.

1. INTRODUCTION

In [1] we reported on an experiment in which the performance of an automatic transcription tool was evaluated. The transcription tool is a continuous speech recognizer (CSR) which can be used to detect whether a phone is present or not (deletions and insertions of phones). It was shown that the CSR's performance is comparable to that of expert linguists who carried out the same task, i.e. to determine whether a phone was present or not in 467 cases. On average, the degree of agreement between the CSR and the listeners was only slightly lower than that between listeners, but comparisons with a reference transcription revealed that the machine's degree of performance was within the range of the linguists'. This means that the automatic tool proposed in [1] can effectively be used to obtain phonetic transcriptions of deletion and insertion processes.

It should be noted that, in the experiments in [1] we simply employed the CSR which we use in our pronunciation variation research [2] without trying to optimize it so as to make the CSR's transcriptions more similar to the human transcriptions. However, it is likely that properties of the CSR, like the speech material used for training, the procedure used to calculate the phone models (PMs) and the internal parameters of the CSR all influence the choice of variants on the part of the CSR.

For example, if the speech material used for training contains much variation in pronunciation and the lexicon contains only one baseline transcription for each word, then some of the transcriptions will be incorrect, e.g. a phone is present in the transcription but has not been realized. This type of mismatch between speech signal and transcription leads to contaminated PMs. Subsequently, the contamination can lead to errors in recognition. Therefore, it is important to minimize the mismatch between the acoustic signal and the transcriptions. One of the approaches we use to minimize the mismatch in the training corpus is by modeling pronunciation variation [2].

Another way of obtaining models which are less contaminated is to train PMs on read speech. It is well known that the extent of variation in spontaneous speech is larger than in read speech. So, for read speech there will be fewer mismatches between the speech signal and the transcriptions. Thus, it is to be expected that PMs which are trained on read speech will be less contaminated than those trained on spontaneous speech.

One can imagine that PMs with varying degrees of contamination may cause the CSR to select different pronunciation variants. As a consequence, the degree of agreement between the CSR and the reference transcription may vary as a function of the PMs employed. The purpose of the present study is to investigate to what extent the degree of agreement between the listeners and the CSR is affected by various sets of PMs.

Furthermore, if the agreement between CSR and listeners is affected by the various sets of PMs, it would be efficient to have a method to estimate how well the PMs will perform beforehand. In a normal situation, judgments given by listeners will not be available (and if they are it defeats the purpose of an automatic transcription tool) whereas different sets of PMs may very well be available. The easiest way of measuring the PMs' performance is by carrying out a standard recognition task. Therefore, we investigated whether word error rates (WER) can predict the degree of agreement between man and machine in selecting pronunciation variants.

This paper is organized as follows, in section 2, the method we used to investigate the performance of different sets of PMs is described. In section 3.1, we will show how different sets of PMs affect the degree of agreement between man and machine. Next, we will concentrate on the degree of agreement between different sets of PMs and the listeners for a number of phonological processes separately (section 3.2). Following on that, the results which indicate whether agreement between man and machine can be predicted on the basis of WER will be given (section 3.3). Finally, in section 4, we will discuss the implications of the results.

2. METHOD & MATERIAL

2.1. Speech Material

The phonological processes under investigation concern insertions and deletions of phones. Pronunciation variants were generated using the following five phonological rules: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion (SAMPA-notation is used throughout this paper). The speech material used in the experiments was selected from the VIOS

database, which contains a large number of telephone calls recorded with the on-line version of a spoken dialogue system called OVIS [3]. OVIS is employed to automate part of an existing Dutch public transport information service. The speech material consists of interactions between man and machine.

From the VIOS corpus, 186 utterances were selected, which 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.

2.2. Experiments

The listeners and the CSR carried out the same task, i.e. deciding which variant best matched the word that had been realized in the spoken utterances for the 379 words (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.

2.3. CSR

2.3.1. Characteristics. The CSR uses phone models (continuous density hidden Markov models (HMMs)), language models (unigram and bigram), and a lexicon. The HMMs consist of three segments of two identical states, one of which can be skipped. In total 38 HMMs were trained. For each of the phonemes /l/ and /r/ two models were trained, a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/). One model was trained for non-speech sounds, and for each of the other 33 phonemes. In addition, a silence model consisting of a one state HMM was employed. For more details on the characteristics of the CSR see [3].

2.3.2. Lexica. Two different lexica were used for training the various sets of PMs: a baseline lexicon and a multiple pronunciation lexicon. The baseline lexicon contains one transcription for each word which was automatically generated using a Text-to-Speech system for Dutch [4]. The multiple pronunciation lexicon was automatically generated by applying the set of phonological rules listed in section 2.1. to the transcriptions in the baseline lexicon. The rules were applied to all words in the lexicon wherever it was possible and in no specific order. All of the generated variants were added to the baseline lexicon, thus creating the multiple pronunciation lexicon.

2.3.3. Forced Recognition. For the automatic transcription task, 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 that are present in the multiple pronunciation lexicon. Forced recognition is imposed through the language model (LM). For each utterance, the LM is derived on the basis of 100,000 repetitions of the same utterance. This means that it is virtually impossible for the CSR to choose other words than those present in the utterance. In this way, the CSR determines for each of the 379 words which of the present variants best matches the actual realization.

The training corpus is re-transcribed by carrying out forced recognition using the lexicon with multiple pronunciation variants. The chosen variants are then included in the training corpus. In this way, an updated transcription of the corpus is obtained which includes pronunciation variation. The updated transcriptions are then used to retrain the PMs.

Our training material, selected from the VIOS database, consisted of 25,104 utterances (81,090 words). The test material, which is used to test the different sets of PMs (section 3.3), consisted of 6,267 utterances (21,106 words).

2.3.4. Phone Models. As we explained in the introduction, we expect that the degree of contamination in the type of PMs used for performing forced recognition will influence which pronunciation variant is chosen. To investigate this, we trained the following five sets of PMs.

1. Baseline PMs: no pronunciation variation modeled, trained on VIOS material (spontaneous speech).
2. FP-model: baseline PMs with an extra model for filled pauses. Filled pauses in the baseline system are transcribed as /@m/ and /@/ thus causing the PM for /@/ to be contaminated by filled pauses. We trained a new model, /@=/, for all filled pauses to minimize contamination of the PM for /@/.
3. Pronunciation variation PMs: PMs in which pronunciation variation was modeled by training them on updated transcriptions, as explained above.
4. Optimized PMs: a combination of the previous two sets of PMs: pronunciation variation and an extra model for filled pauses.
5. Polyphone PMs: no pronunciation variation modeled, trained on Polyphone [5], a corpus which contains read speech.

2.4. Evaluation

For our evaluation we used reference transcriptions which were based on the judgments of the nine listeners in [1]. The reference transcriptions were 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.

Furthermore, the results are presented using Cohen's κ which is a measure of agreement in which a correction for chance agreement is made [6]:

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

P_o = observed proportion of agreement

P_c = proportion of agreement on the basis of chance

$-1 \leq \kappa \leq 1$

The reason we decided to use κ instead of percentage agreement is that the 0/1 distribution differs for the various rules. Due to these differences in the 0/1 distribution, the chance agreement for the various rules may differ. Consequently, the rules cannot simply be compared with one other unless a correction for chance agreement is made. Qualifications for different values of κ are: .00 - .20 slight, .21 - .40 fair, .41 - .60 moderate, .61 - .80 substantial and .81 - 1.00 almost perfect [6].

3. RESULTS

In order to determine whether the performance of the CSR is influenced by using different sets of PMs, two comparisons between man and machine were made. First, we calculated the overall agreement for each of the sets of PMs with the listeners (section 3.1.). Second, we looked at agreement between the listeners and the PMs for each of the rules separately (section 3.2.). Finally, we also wanted to know whether the performance of a set of PMs in a forced recognition task could be predicted on the basis of recognition results. These results are presented in section 3.3.

3.1. Agreement for all Rules with Reference Transcription

In Figure 1, the κ values for the various sets of PMs compared to the five different reference transcriptions are shown. It can be seen that for all of the sets of PMs the κ values increase as the reference transcription becomes stricter.

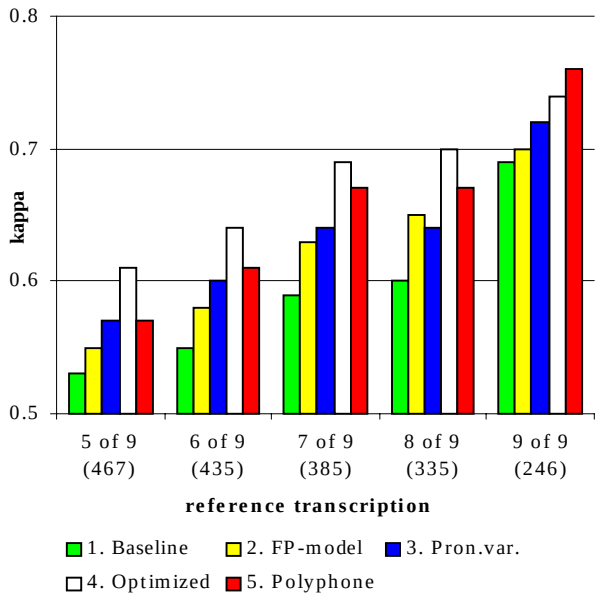

Figure 1: Agreement between PMs and reference transcriptions for different degrees of strictness. (Numbers between brackets indicate number of items.)

Figure 1 also shows that both approaches to minimizing the contamination in the PMs have a positive effect on agreement with the reference transcriptions. First of all, by adding a separate model for filled pauses and modeling pronunciation variation by using five phonological rules both lead to a higher degree of agreement with the listeners than the baseline PMs do. Moreover, the optimized set of PMs, which is the combination of adding a model for filled pauses and modeling pronunciation variation, leads to even higher κ values. Secondly, training PMs on read speech instead of spontaneous speech also leads to a higher degree of agreement with the reference transcription compared to the baseline.

3.2. Agreement for Different Rules

In the previous section, we compared the performance of the different sets of PMs to the reference transcription with all the rules pooled together. In Figure 2, the agreement is shown for the five phonological rules separately. Only the results of agreement with reference transcription type ❸ (6 out of 9) are shown here.

Figure 2 shows that the κ values for the four deletion rules roughly stay the same for the different sets of PMs, whereas for the /@/-insertion rule there is a gradual increase when going from PMs 1 to PMs 4. Thus, the gradual increase in κ values seen in Figure 1 when going from PMs 1 to PMs 4, is mainly a result of the increase in κ values for the /@/-insertion rule.

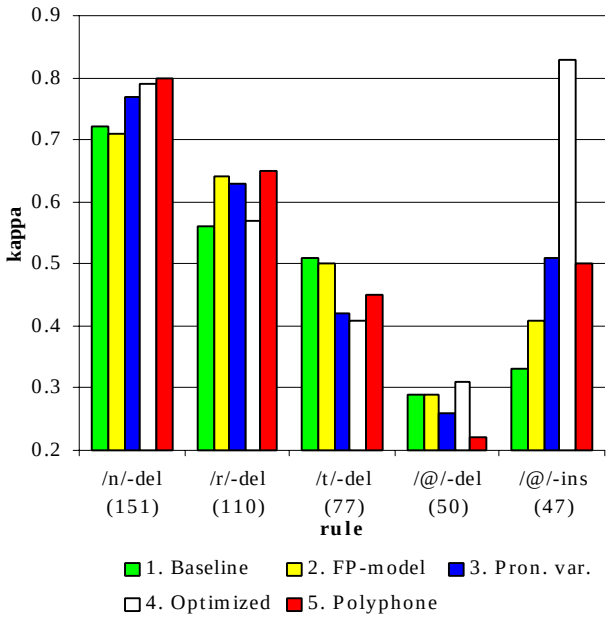

Figure 2: Agreement between PMs and a reference transcription based on at least 6 of 9 listeners agreeing, for the five rules separately (Numbers between brackets indicate number of items).

Figure 2 shows that the PMs are affected more by modeling pronunciation variation for the /@/-insertion rule than the other rules. The obvious difference here is that the first four processes are all deletion processes whereas the last process is a process of insertion. The difference between these processes lies in the specific PMs that are contaminated during training of the baseline PMs. Table 1 shows examples for both processes (the contaminated phones highlighted). For deletion processes it is the PM for the phone which is deleted which is contaminated, whereas for an insertion process the PMs of the phones surrounding the phone which is inserted are contaminated. Furthermore, the process of /@/-insertion also causes post-vocalic /L/ and /R/ to become pre-vocalic /l/ and /r/. This may possibly also influence the results found for /@/-insertion, and this will be investigated in further detail in the near future.

	baseline	pron. var.
/@/-deletion	/la:t@r@/	/la:tr@/
/@-insertion	/dELft/	/dEl@ft/

Table 1: Examples of application of the rules for /@/-deletion and /@/-insertion with phones which cause contamination in the set of baseline PMs highlighted.

The combination of modeling pronunciation variation and adding an extra model for filled pauses, as far as /@/-insertion is concerned, is clearly the most obvious improvement. However, why additionally adding a model for filled pauses is beneficial to the CSR for /@/-insertion and not for /@/-deletion is not quite clear.

On the whole, we found that the nine listeners tend to say that a /@/ is present in more cases than the machine, i.e. the CSR chooses more /@/-deletion and less /@/-insertion (especially for baseline PMs) than the listeners do. This can partly be explained by the fact that listeners use information from context, transitions etc. to base their judgments on. Furthermore, it is very difficult for listeners to judge whether or not a /@/ is present in the words for which /@/-insertion and /@/-deletion can occur, because they can always still (imagine they) hear part of the /@/. As for the PMs, they are monophones with no explicit context modeling. Therefore, it may be better to use context dependent PMs instead of context independent PMs.

3.3 Recognition Performance of Phone Models

It would be most efficient to estimate beforehand what type of PMs should be used for a task such as the one described here. To find out if this is possible we carried out a number of recognition tests. We carried out standard recognition tests on our test material (VIOS) and calculated the best sentence word error rates (WER = ((S+D+I)/N)*100) for each set of PMs. The lexicon which was employed was the multiple pronunciation lexicon which was used for all of the tests with the different PMs. The results obtained are shown in column 2 of Table 2.

PM	WER	κ (6/9)
baseline	12.44	.55
FP-model	12.30	.58
pron. var.	12.22	.60
optimized	12.01	.64
polyphone	18.06	.61

Table 2: WERs for different types of PMs and κ for a reference transcription where 6 of the 9 listeners agree.

Table 2 shows that WERs decrease and the κ values increase as the PMs become less contaminated. The only exception is the set of polyphone PMs. The WER on the VIOS test set is significantly higher for this set of PMs than for all other sets whereas κ is almost the highest. This is not surprising as for polyphone there is a mismatch between training and test material, whereas for the other sets of PMs this is not the case.

4. CONCLUSIONS

From the results in Figure 1 it is clear that using different PMs in a forced recognition task leads to different results. Minimizing the contamination in the PMs leads to PMs which show a higher degree of agreement with listeners. However, this is not the case for each of the rules separately. Figure 2 showed that minimizing the contamination in the PMs does not have a pronounced effect on the performance of the forced recognition for the deletion processes whereas it certainly has an effect for the process of /@/-insertion.

The WERs obtained by performing a normal recognition on an independent test set give an indication as to how well the PMs will perform in a forced recognition test, as long as there is no mismatch between training and test material.

We can also conclude that the type of training material employed to train a set of PMs affects their performance in a forced recognition test, as well as in a normal recognition test. We found that PMs trained on read speech and tested on spontaneous speech perform substantially worse than PMs trained on the same type of spontaneous speech in a standard recognition task. However, in a forced recognition task these PMs outperform most of the PMs trained on spontaneous speech.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Kessens, J.M., Wester, M., Cucchiarini, C. and Strik, H. 1998. The Selection of Pronunciation Variants: Comparing the Performance of Man and Machine. *Proc. International Conference on Spoken Language Processing*, Vol. 6: 2715-2718.

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

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

[4] Kerkhoff, J. and Rietveld, T. 1994. Prosody in Niros with Fonpars and Alfeios. *Proc. Dept. of Language & Speech, University of Nijmegen*, Vol.18: 107-119.

[5] den Os, E.A., Boogaart, T.I., Boves, L. and Klabbers, E. 1995. The Dutch Polyphone Corpus. *Proc. Eurospeech 95*. 825-828.

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