Lower WERs do not guarantee better transcriptions

Judith M. Kessens & Helmer Strik

A²RT, Dept. of Language and Speech, Univ. of Nijmegen, the Netherlands
{J.Kessens, Strik}@let.kun.nl
http://lands.let.kun.nl/

Abstract

The goal of this paper is to investigate the effect of various properties of the CSR on automatic transcription. To this end, we used various versions of a continuous speech recognizer (CSR) to make automatic transcriptions. Our results show that changing certain properties of the CSR affects the resulting automatic transcriptions. The best results were obtained when 'short' hidden Markov models (HMMs), and context-independent HMMs were used. Furthermore, we found that minimizing the amount of contamination in the HMMs improves the quality of the automatic transcriptions. Another important result is that there does not appear to be a straightforward relation between word error rate (WER) and the transcription quality. In other words: A CSR with a lower WER does not always guarantee better transcriptions.

1. Introduction

In the last few years, increasingly large speech corpora are being developed for speech technology. Since for many purposes it is necessary that these corpora be phonetically transcribed, there has been a growing interest in automatic transcription, because for such large-scale corpora manual transcriptions are infeasible (see e.g. [1]).

In [2], we reported on an experiment in which the performance of an automatic transcription tool was evaluated. It was shown that the performance of the CSR used for making the automatic transcriptions is comparable to that of expert listeners who carried out the same transcription task. On average, the degree of agreement between the CSR and the listeners was only slightly lower than that between the listeners.

In the experiment in [2], we simply employed the CSR which we used in other research [3] without trying to optimize it to make the CSR's transcriptions more similar to the human transcriptions. One is inclined to think that the best CSR available, i.e. the CSR with the lowest WER, will produce the most optimal transcriptions. However, it remains to be seen whether a CSR with a lower WER yields better quality transcriptions. It is likely that properties of the CSR, such as for instance the speech material used for training, the procedure to estimate the phone models and the internal parameters of the CSR all influence the CSR's transcriptions.

The goal of the current research is to investigate the effect of various properties of the CSR on automatic transcription. To this end, various versions of our CSR carried out exactly the same transcription task as the humans in [2]. As a quality measure of the various automatic transcriptions, we used agreement between the automatic transcriptions and a reference transcription based on the human transcriptions. Furthermore, we examined the relation between the degree of agreement and the WER measured on an independent test set.

This paper is organized as follows: In section 2, we describe the speech material and the method that we employed. Subsequently, in section 3, we present the results for each of the CSR's properties that was investigated. The relation between degree of agreement and WER is examined in section 4. Finally, in section 5, we discuss the results, and present some concluding remarks in section 6.

2. Material and method

2.1. Material

The speech material used in the experiments was taken from the VIOS database, which consists of interactions between man and machine. The extemporaneous (or spontaneous) speech in the VIOS database contains a lot of variation in pronunciation. The variation we investigated concerns the following five frequently observed processes (rules): /n/-, /r/-, /t/- and /@/-deletion, and /@/-insertion (SAMPA notation is used throughout this paper). From the VIOS material, 186 utterances were selected, which contain 379 words with relevant contexts for one or two of the above-mentioned rules to apply.

2.2. Transcriptions

The transcription task was to determine which of the variants (generated using the five rules) best matched the words that had been realized in the spoken utterance (forced choice). For 88 of the 379 words, the conditions for rule application were met for two rules simultaneously and for the remaining 291 words only one condition of rule application was relevant. For each underlying rule, we determined whether the rule was applied or not. Thus, in total 467 binary scores were obtained for each of the listeners and the CSR (for more details regarding this experiment, see [2]).

The 9 listeners who carried out the transcription task all had experience in making transcriptions. A reference transcription was obtained by taking the majority vote, i.e. the transcription on which 5 or more of the 9 listeners agree. Several versions of our CSR were used to carry out the same task, thus obtaining various automatic transcriptions. These automatic transcriptions were generated by using the CSR in forced recognition mode, which means that the recognizer has to decide which of the two (or more) variants better matches the acoustic signal: the one with or the one without the target phone.

Two measures of agreement between the automatic transcription and the reference transcription were calculated:

1. Percentage agreement:

$$P_o = 100\% \times \frac{\#\text{agreements}}{\#\text{agreements} + \#\text{disagreements}} \qquad (1)$$

2. Cohen's kappa, which corrects for chance agreement [4]:

$$\text{kappa} = \frac{P_o - P_c}{100 - P_o} \quad (2)$$

-1 < kappa < 1

P_c = percentage agreement on the basis of chance

As a measure of agreement, we use Cohen's kappa, which has the advantage that kappa values for various conditions can be compared with each other. We will also present percentage agreement, because this measure is used more often than Cohen's kappa and is therefore easier to interpret. However, one should bear in mind that percentages of agreement cannot be compared to each other without correcting for chance agreement.

2.3. CSR

We used a standard, off-the-shelf HMM recognizer. The baseline phone models are continuous density HMMs with 32 Gaussians per state. Every 10 msec, 14 cepstral and their deltas are calculated for frames with a width of 16 msec. The HMMs are trained on 25,104 VIOS utterances (81,090 words).

The topology of the baseline HMMs is as follows: each HMM consists of six states or three segments of two identical states, one of which can be skipped [5]. In total, 38 HMMs were trained. For each of the 35 phonemes, context-independent HMMs were trained. In addition, one model was trained for non-speech sounds, one model was used for filled pauses, and a model consisting of only one state was employed to model silence. For more details on the CSR, see [6].

3. Results

In this section, we investigate how the properties of the CSR influence automatic transcription. To this end, various versions of our CSR are used for obtaining different automatic transcriptions. Next, agreement was calculated between the automatic transcriptions and the reference transcription, and the various agreement values were compared with each other.

3.1. Topology of the HMMs

In previous research [2], we found that, in general, our CSR detects the realization of phones less often than the humans do: The CSR decided that 55% of the phones were present, whereas this percentage was 67% for the listeners. This effect was most clearly present for the /@/-deletion rule. Furthermore, the results in [2] showed that agreement between the automatic transcriptions and the reference transcriptions increased if the /@/s which were judged to be short in duration were denoted as "not present". This could be an indication that the intrinsic minimum duration of 30 msec - imposed by the model topology - makes it less probable that /@/s with a duration shorter than 30 msec are detected by the CSR. For this reason, we decided to investigate whether using different phone model topologies with shorter intrinsic minimum duration could improve automatic transcription. To this end, all /@/s in the training material that had a duration equal to or shorter than 30 msec were used to train a short-/@/ HMM. The remaining /@/s were used for training the long-/@/ HMM, consisting of 3 segments. In addition to baseline HMMs with a topology of 3 segments, two different topologies were used for the short-/@/ HMM:

1. A two-segment topology with an intrinsic minimum duration of 20 msec (2seg), and
2. A one-segment topology, with an intrinsic minimum duration of 10 msec (1seg).

Table 1 shows that agreement increases when a phone model topology is used that allows for a duration of /@/ shorter than 30 msec (compare 1seg/2seg to 3seg). Closer inspection of the data reveals that agreement increases for the /@/-deletion rule, whereas it decreases for the /@/-insertion rule. Consequently, the optimal choice is to use the short-/@/ HMM for automatic transcription of the deletion processes only, and to use the long-/@/ HMM for the insertion process (mixed condition). For this mixed condition, the agreement is somewhat higher.

Table 1: Various HMM topologies

HMMs	3 seg	2 seg	1 seg	mixed
% agreement	75.8	77.1	77.7	78.6
Kappa	0.50	0.52	0.52	0.53

3.2. Degree of contamination of the HMMs

The speech material used for training contains much variation in pronunciation, however, the baseline training lexicon contains only one transcription for each word. Therefore, some of the transcriptions used for training the phone models 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 of the training material leads to contaminated HMMs. Subsequently, the contamination can lead to errors in the automatic transcriptions. Therefore, it is important to minimize the mismatch between the acoustic signal and the transcriptions used for training. One of the approaches we used to minimize the mismatch in the training corpus is by modeling pronunciation variation [3]. In addition to the baseline HMMs, two other sets of HMMs were used (see [3]):

1. HMMs trained on a corpus in which pronunciation variants of 5 phonological rules are transcribed (5 rules)
2. HMMs trained on a corpus in which besides the pronunciation variants of the 5 rules cross-word variation is also transcribed (5 rules + cross)

Table 2 shows that agreement increases when HMMs are used that are less contaminated due to modeling of pronunciation variation.

Table 2: Pronunciation variation modeling

HMMs	baseline	5 rules	5 rules + cross
% agreement	75.8	79.2	79.9
Kappa	0.50	0.56	0.58

It is well known that the amount of variation in spontaneous speech is larger than that in read speech. Consequently, in read speech there will probably be fewer mismatches between the speech signal and the transcriptions. Thus, it is to be expected that HMMs trained on read speech will be less contaminated than those trained on spontaneous speech. Since less contaminated HMMs yield better results (see Table 2), we decided to use HMMs trained on read speech for automatic transcription. The HMMs were trained on the

18,000 phonetically balanced, read sentences of the Dutch Polyphone corpus [7].

Table 3 shows that agreement is indeed higher when read speech HMMs (Polyphone) are used instead of spontaneous speech HMMs (VIOS).

Table 3: Spontaneous vs. read speech

HMMs	spontaneous (VIOS)	read (Polyphone)
% agreement	75.8	84.9
Kappa	0.50	0.57

3.3. Acoustic resolution of the HMMs

Some researchers use acoustic models with a low acoustic resolution for automatic transcription, see for instance [8] and [9]. This is probably done because they expect that HMMs with a low acoustic resolution produce better transcriptions. We investigated whether this assumption is true, by using HMMs with varying acoustic resolutions for automatic transcription. In addition to the baseline HMMs with 32 Gaussians/state, we also used HMMs with lower acoustic resolutions of 16, 8, 4 and 2 Gaussians/state.

Table 4 shows that agreement varies using HMMs with different numbers of Gaussians/state, but the tendency that low resolution HMMs perform better than high resolution HMMs was not be observed.

Table 4: Various acoustic resolutions

HMMs	32	16	8	4	2
% agreement	75.8	75.4	76.9	73.9	75.4
Kappa	0.50	0.49	0.52	0.47	0.49

3.4. Context-independent vs. context-dependent HMMs

We also compared context-independent (CI) with context-dependent (CD) HMMs. CD-HMMs generally yield lower WERs. However, we hypothesize that CD-HMMs do not necessarily generate better transcriptions.

To better illustrate our point, we can look at the following example: In the citation forms of our VIOS training corpus, 41,615 /n/s are present, and 7,227 of these /n/s occur in the context /@n|/ ('|'= word boundary). However, a large part of these latter /n/s is not realized: According to the CSR, the /n/s occurring in the context /@n|/ are not present in 3,337 cases (46%). Consequently, if a CD-HMM is trained for /@n|/, then the /n/ is not present in about half of the training tokens. It is not likely that such a CD-HMM is suitable to detect whether an /n/ is present or not.

In more general terms, if a phone F (Focus) in a specific left (L) and right (R) context is often deleted, i.e. /LFR/ → /L-R/ ('-' = deletion), the CD-HMM for /LFR/ might be less suitable for automatic transcription than a CI-HMM for /F/.

Since previous research has shown that the most frequent deletion process in our VIOS material is: /@n|/ → /@-|/, we focused on this particular context. In our baseline transcriptions all /n/s in this context are deleted, since in the linguistic literature this is generally considered to be the most likely pronunciation. In order to test our hypothesis, we first re-inserted all the /n/s in the /@-|/ context. Table 5 shows that agreement increases if all /n/s are replaced in the baseline transcriptions (compare CI: /@n|/ to CI: /@-|/).

In addition to the two sets of CI-HMMs, two sets of CD-HMMs were trained:

1. Only for the context /@n|/ a CD-HMM is trained, for all other contexts CI-HMMs are trained (CD: 1: /@n|/).
2. CD-HMMs are trained for all contexts that occurr more than 200 times in the training material (CD: many).

Table 5 shows that for the two sets of CD-HMMs, the agreement between the automatic transcriptions and the reference transcriptions is lower than the agreement for the corresponding set of CI-HMMs (CI: /@n|/).

Table 5: CI vs. CD HMMs

HMMs	CI		CD	
	/@-|/	/@n|/	1: /@n|/	many
% agreement	75.8	77.9	75.8	76.9
Kappa	0.50	0.52	0.47	0.45

4. Agreement and WER

In the previous sections, various sets of HMMs were used to obtain automatic transcriptions. For all these sets of HMMs we also calculated WER (= (sub+del+ins)/N). Since one will often select the best CSR that is available, i.e. the CSR with the lowest WER, for generating automatic transcriptions, it is interesting to look at the relation between WER and the agreement between the automatic transcriptions and the reference transcriptions.

WER was calculated for an independent test set of 6,276 VIOS utterances (21,106 words), using the same lexicon that was used for the automatic transcription of the pronunciation variants of the 5 rules. A scatter plot of kappa as a function of WER is shown in Figure 1. In this figure, the following symbols indicate the different HMMs that were used:

- ● Baseline HMMs
- ◇ Short-/@/ HMMs (section 4.1)
- ○ HMMs from pronunciation variation modeling research (section 4.2)
- × Read speech HMMs (Polyphone) (section 4.2)
- □ HMMs with various acoustic resolutions (section 4.3)
- △ Context-dependent HMMs, (section 4.4)

The relation that is to be expected between WER and degree of agreement is; the lower the WER, the higher the kappa value. However, Figure 1 does not show the expected relation between WER and kappa. Moreover, the CD-HMMs with the lowest WER (CD: many) yield the lowest kappa value, and the read speech HMMs with a high kappa value yield a high WER.

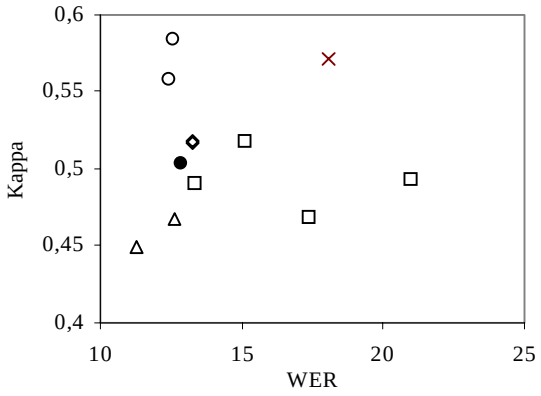

Figure 1: Scatter plot of kappa and WER

5. Discussion

The results we presented in this paper show that a CSR that is better in recognizing words (i.e. with a lower WER) does not always produce better transcriptions (i.e. with a higher agreement). Apparently, automatic transcription is a different task than recognizing words. The results of the read speech HMMs clearly show this point; Due to the mismatch between training and test material, the WER is high, whereas the quality of the automatic transcriptions is reasonably good. In order to obtain high quality automatic transcriptions, one could start with using HMMs that are trained on read speech. However, if the automatic transcriptions need to be further improved, specialized CSRs should be developed which are optimized for the transcription task.

In order to optimize a CSR for the task of making automatic transcriptions, evaluation of the transcriptions is necessary. However, evaluation remains problematic since there is no completely error free reference transcription with which the automatic transcriptions can be compared (see [10], pp. 11-13).

To circumvent the latter problem (at least partly), the following two strategies have been devised for obtaining a reference transcription:

1) A consensus transcription is used, which is a transcription made by several transcribers after they have agreed on each individual symbol.
2) A majority vote principle is used, which means that only that part of the material is used for which the majority of the listeners agreed.

In the current study, we have adopted the latter approach. Our reference transcription was based on a majority of 5 out of 9. It would also have been possible to report the results for a stricter reference transcription, i.e. based on agreement of 6, 7, 8 or 9 out of 9 listeners. For these stricter reference transcriptions, agreement is higher. To give an idea: For a reference transcription of 5 out of 9, kappa varies from 0.45 to 0.58, whereas kappa varies from 0.67 to 0.74 for a reference transcription of 9 out of 9. In other words, for the cases in which the agreement between humans is higher, the agreement between CSR and humans is also higher. Apparently, the other cases are more difficult for both man and machine.

Furthermore, by using a reference transcription based on a majority vote of 5 out of 9, we obtained scores for all 467 cases. If we had used stricter reference transcriptions, the number of cases would have been smaller. For instance, all nine human transcribers agreed on only 246 of the 467 cases (53%).

6. Concluding remarks

We have shown that changing the properties of a CSR does influence the degree of agreement between the automatic transcriptions and the reference transcription: Kappa varies between 0.45 and 0.58, and percentage agreement varies between 73.9 and 84.9%. In short, the quality of the automatic transcriptions can be increased by using 'short' HMMs, 'less contaminated' HMMs, and CI-HMMs. Regarding the acoustic resolution of the HMMs, no clear trends were observed. Furthermore, there is no clear relation between the WER of a CSR and the kappa values.

On the basis of our results we can therefore conclude that for obtaining automatic transcriptions taking the CSR with the lowest WER is not always the optimal solution. Indeed, it appears that for this specific purpose, CSRs should be used that have been specially optimized for automatic transcription.

7. Acknowledgements

The research by Judith M. Kessens was carried out within the framework of the Priority Programme Language and Speech Technology, sponsored by NWO (Dutch Organization for Scientific Research). The authors would like to thank several members of the research group *A²RT* for their useful comments on earlier versions of this paper.

8. References

[1] Schiel, F., "automatic phonetic transcription of non-promted speech", *Proc. ICPhS*, 607-610, 1999.

[2] M. Wester, J.M. Kessens, C. Cucchiarini & H. Strik, "Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer", *accepted for publication in Language & Speech.*

[3] Wester, M., Kessens, J.M, and Strik, H. "Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation", *Proc. of the workshop on pronunciation variation modeling for ASR, Rolduc*, 145-150, 1998.

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

[5] 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., "The Philips Research System for Large-Vocabulary Continuous-Speech Recognition", *Proc. of Eurospeech*, 2125-2128, 1997.

[6] Strik, H., Russel, A.J.M., van den Heuvel, H. Cucchiarini, C. & Boves, L. "A spoken dialog system for the Dutch public transport information service", *Int. Journal of Speech Technology, Vol. 2, No. 2, 1997*, p 119-129.

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

[8] Beulen , K., Ortmanns, S., Eiden, A., Martin, S., Welling, L., Overmann, J., Ney, H., "Pronunciation modelling in the RWTH large vocabulary speech recognizer", *Proc. of the workshop on pronunciation variation modeling for ASR, Rolduc*, 13-16, 1998.

[9] Schiel, F., Kipp, A., Tillmann, H.G., "Statistical Modelling of pronunciation: it's not the model, it's the data", Proc. *of the workshop on pronunciation variation modeling for ASR, Rolduc*, 131-136, 1998.

[10] Cucchiarini, C., *Phonetic transcription: a methodological and empirical study*, Ph.D. thesis, University of Nijmegen, 1993.