Analysis of N-Best Output Hypotheses for Fast Speech in Large Vocabulary Continuous Speech Recognition

Tibor Fábián, Thilo Pfau, Günther Ruske

Institute for Human-Machine-Communication, Technical University of Munich
Arcisstr. 21, D-80290 Munich, Germany
fab@mmk.e-technik.tu-muenchen.de

Abstract

The performance of speech recognition systems often deteriorate considerably with fast speech. Particularly when the recognizer is run in mismatched conditions, e.g. fast speech, the performance can be improved by properly selecting one of the N-best recognition output hypotheses. For the selection of the best hypothesis, different speech rate measures were taken into account. First, to show the potential of the speech rate as a selection criterion, an "ideal" speech rate value is assumed, which is calculated from the known transcription. Phoneme and vowel rate are compared. Second, a phoneme recognizer is used to estimate the speaking rates of unknown sentences. Tests on the spontaneously spoken German Verbmobil material showed a relative decrease of 6.6% in the word error rate for fast speech, when taking the estimated vowel rate which is almost as good as using the "ideal" vowel rate (relative improvement of 7.64%). The most accurate match of N-best output hypotheses shows that the word error rate could ideally be decreased by 26.75%.

1. Introduction

The performance of a speaker-independent automatic speech recognition system is dependent on the speech rate [1]. This is pointed out in the following table, where the word error rate for fast speech is clearly higher than for slow or average speech.

Speech Rate	slow	average	fast
WER [%]	24.2	33.8	44.4

Table 1: Word error rates (WER) for slow, average and fast speech in percent (Verbmobil task)

On the one hand, this dependence is caused by the fact that in many cases not enough speech data exist in the training material for fast speech. On the other hand, the temporal and spectral characteristics of speech, e.g. duration of vowels (see Figure 1) or the position of formant frequencies [2] change with varying speech rate.

There are different techniques to improve the recognition performance for fast speech, for example the training of the acoustic models for fast speech [3][4][11][12], or the modeling of speech rate with additional output probabilities attached to the transitions between states of Hidden Markov Models [5]. Moreover speech rate dependent decision trees can be constructed for improving the HMM discrimination ability [8].

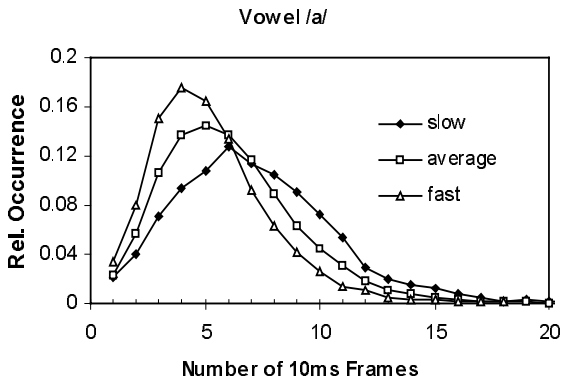

Figure 1. Histogram of vowel duration for vowel /a/ as a function of the number of frames for slow, average and fast speech.

A further new and simple method will be described in this paper, namely the analysis of N-best output hypotheses by taking the speech rate into account. This means, that not only the best hypothesis but a list of the possible N-best hypotheses will be produced by the recognizer. Using an additional knowledge source, not yet integrated in the search process, e.g. the speech rate, the N-best list is resorted and a new hypothesis is then selected from the list.

In this study, we have examined the change of the recognition performance for fast speech with a large vocabulary continuous speech recognition system on the evaluation set 1996 of the German Verbmobil spontaneous scheduling task comprising 53 sentences classified as "fast" [13].

2. Analysis of N-best output hypotheses

The task of a speech recognizer is - generally - to find the best word sequence for unknown speech data [6]. This task

is frequently carried out by using acoustic modeling to determine the best phoneme sequence, while the best hypotheses of words can be determined with the help of a dictionary and a language model.

Using the analysis of N-best hypotheses (N-best rescoring), which is described in this paper, the workload of the recognizer in the selection of the best hypotheses can be "lightened", as the hypothesis which matches best to the spoken utterance has not to be scored as the best one, but must only be found in the N-best list. This approach is especially helpful in "mismatched" conditions, i.e. conditions which are not or not sufficiently represented within the training material. This is normally true for fast speech, as most databases do not contain large amounts of fast speech material. Thus the integration of knowledge about the speaking rate could be helpful for the analysis and selection of N-best hypotheses.

Furthermore an N-best rescoring is suitable for integrating knowledge sources in the decision process, which are not available before the end of an utterance or for which the integration into the search process is computationally very expensive. The speaking rate can vary considerably between different utterances of the same speaker or even within one utterance. Therefore an analysis of N-best hypotheses for fast speech is a suitable method, because at the end of the utterance it is easier to obtain information about the speaking rate. A post-selection can then be carried out by taking the speech rate into account.

2.1. Most accurate match of N-best output hypotheses

In this section the potential of the N-best list rescoring concerning the possibility of improving the recognition performance is examined. For this purpose the word error rate, which is the ideal criterion, is taken for the selection of one hypothesis of the N-best recognition output hypotheses. In this case the recognition result must be known in advance, of course. This method is thus only suitable for recognition performance analysis. All other methods which select a definite hypothesis from the N-best list by using additional information, can at best achieve the performance of the most accurate match of N-best output hypotheses, i.e. the most accurate match is an upper bound for all other methods.

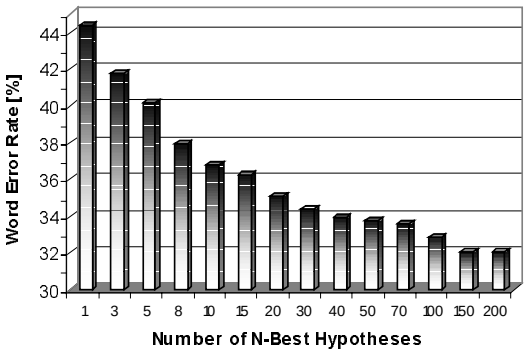

Figure 2. Word error rates for the most accurate match of different numbers of N-best hypotheses.

The results of our experiment for "fast" speech are presented in the Figure 2. It shows that by using this best selection criterion the word error rate can be reduced by 12.33% at N-best list length 200 (relative 27.78%).

In practice, the selection of the best hypothesis can not be carried out using the reference (which is not known), but some other supplementary information must be utilized. For example, if some information about the speaking rate is available, this information can be used as a criterion for N-best list analysis. By comparing this information about the speaking rate with the speaking rates of the different hypotheses of the N-best list, the most suited hypothesis is selected from the N-best list.

2.2. Computing speech rate from transcribed test data

In this section the use of the speech rate as a criterion for rescoring N-best recognition output hypotheses is evaluated. Therefore we first use an "ideal" speech rate estimator, i.e. we extract the information about the speaking rate from the transcription, to be able to assess the potential of this method on fast speech. The procedure is now described in detail.

First, a list with the reference speech rates ROSref,ph (reference Rate of Speech using phonemes) for the "fast" sentences is created. The phoneme rate is taken as the measure for the speech rate, which is calculated from the number of phonemes in the transcription file and the number of frames in the sentence. Second, a further list was created including the N-best recognizer output hypotheses of a determined sentence. The "recognized" speech rate ROSrec,ph (recognized Rate of Speech using phonemes) for each hypothesis is registered in this list. The ROSrec,ph is calculated for a hypothesis using the number of phonemes in the hypothesis and the number of frames of the sentence to be recognized. The number of phonemes of each word can be determined from the recognizer dictionary.

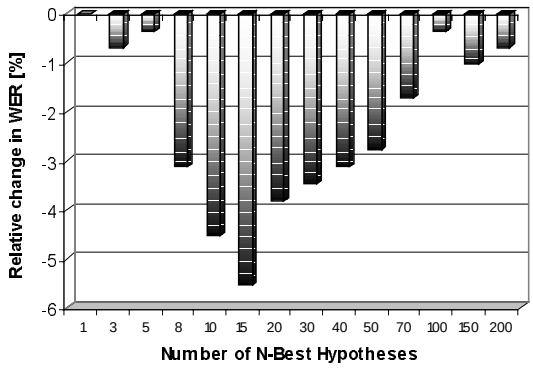

Figure 3. Relative change in word error rate using ROSref,ph as selection criterion for different numbers of N-best output hypotheses.

The difference between the ROSref,ph and the ROSrec,ph is taken as a criterion to select the hypothesis of the N-best hypotheses list which shows the least difference. The selection of the hypothesis was carried out for different N-best list lengths. The results are summarized in Figure 3.

As we can see in Figure 3, it is possible to achieve a relative decrease in word error rate of 5.55% for the N-best list length 15. This means that actually using this method could not completely take advantage of the potential of the most accurate match, however a relatively high decrease of the recognition-error for "fast" speech could be achieved.

For an error analysis the changes in the number of deletions, substitutions and insertions as a function of the N-best list length are of particular interest. A closer look at Figure 3 reveals that the word error rate can be decreased efficiently using ROSref,ph. This fact confirms that the consideration of the speech rate is advantageous for the selection of the best hypothesis. The question is, why then an optimal result (see 2.1) can not be achieved.

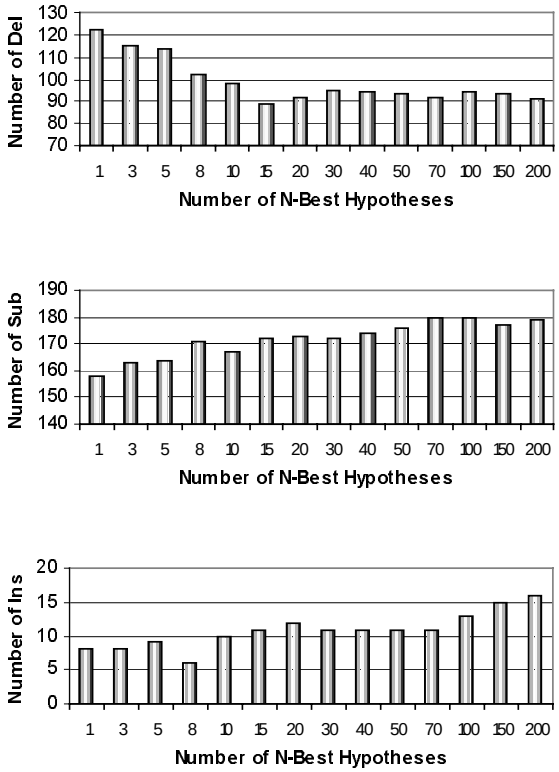

Figure 4. Number of deletion errors (Del), substitution errors (Sub) and insertion errors (Ins) for different numbers of N-best output hypotheses.

Figure 4 shows that, in general, the number of deletions is decreased from 122 to 89, while the number of insertions only increases from 8 to 16 with increasing N-best list length (from 1 to 200). The number of substitutions is actually increased because the hypotheses are more suited to the reference rate which is why they are preferred in the selection. However they do not always contain the correct words.

Additionally we want to consider a second speech rate measure, the vowel rate. Equivalent to the phoneme rates the speech rate ROSref,vow (reference Rate of Speech using vowels) and ROSrec,vow (recognized Rate of Speech using vowels) were calculated using the number of vowels instead of the number of phonemes. The determined results can be seen in Figure 5.

A relative decrease in word error rate of 7.64% with the N-best list length 10 can be achieved. This is a slightly better result compared to the use of the phoneme rate (5.55% relative change, see Figure 3). Additionally this result is more robust to variations of the N-best list length, since comparable improvements can be achieved for N-best list lengths from 8 to 20. For this reason the speech rate will be calculated with the vowel method in the following procedure (see 2.3).

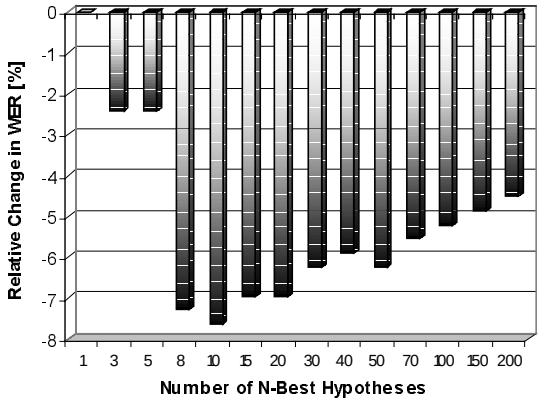

Figure 5. Relative change in word error rate using ROSref,vow as selection criterion for different numbers of N-best output hypotheses.

In practice, it is not possible to extract information from a known transcription of the speech data file, as unknown sentences are recognized. Nevertheless, there are different procedures to measure the speech rate of speech data with unknown content: e.g. the use of a phoneme recognizer, vowel detector [7][9] or the use of energy based measures [10].

2.3. Computing speech rate using a phoneme recognizer

To achieve an estimate of the speaking rate ROSest,vow (estimated Rate of Speech using vowels) a phoneme recognizer is used to produce a possible phoneme sequence of the unknown utterance. The speech rate can then be determined from the number of the vowels contained in this phoneme sequence and from the number of the frames, as described in section 2.2. This estimated speech rate is used as the selection criterion for the N-best list rescoring procedure, which therefore can be performed for unknown speech data. The recognition of the phoneme sequence does, however, contain mistakes and this will be also reflected in the test results. The tests were again carried out with the 53 "fast" speech sentences, as in the previous evaluations.

Figure 6 shows, that using ROSest,vow a relative decrease in word error rate of 6.6% can be achieved with an N-best list length of 10. This decrease is slightly higher compared to the use of ROSref,ph and comparable to the decrease achieved by using ROSref,vow, although ROSref,ph

and ROSref,vow are calculated from the known transcription while ROSest,vow is only an estimated speech rate.

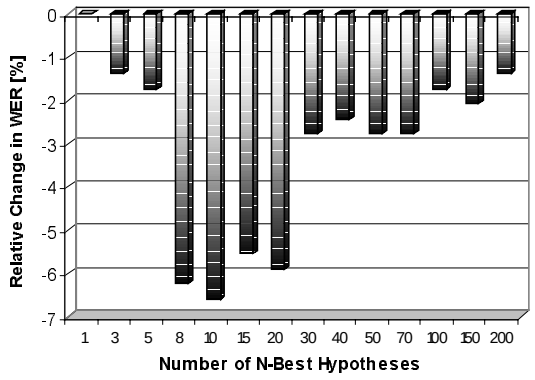

Figure 6. Relative change in word error rate using ROSest,vow as selection criterion for different numbers of N-best output hypotheses.

3. Summary

The analysis of N-best output recognition hypotheses of a large vocabulary continuous speech recognition system has a lot of potential, particularly in mismatched conditions like e.g. "fast" speech. A possible selection criterion for a post evaluation of the recognized output hypotheses is the speech rate.

Different measures of the speaking rate were used as selection criteria. First, using "ideal" speech rate estimates, determined from the transcription relative improvements of 5.55% respectively 7.64% were achieved with phoneme and vowel rate. Second, the use of an estimated vowel rate basing on phoneme recognition was performed. This method reduced the word error rate of the speech recognizer for fast speech by relatively 6.6%. However, to be able to achieve the ideal most accurate match of N-best output hypotheses (27.78% relative change), further methods have to be investigated.

4. References

[1] Martínez F., Tapias D., Álvarez J. and León P. "Characteristics of Slow, Average and Fast Speech and Their Effects in Large Vocabulary Continuous Speech Recognition", *Proceedings of Eurospeech Conference*, Rhodes, Greece, 1997, pages 649-672.

[2] Kuwabara H., "Acoustic and Perceptual Properties of Phonemes in Continuous Speech as a Function of Speaking Rate", *Proceedings of Eurospeech Conference*, Rhodes, Greece, 1997, pages 1003-1006.

[3] Faltlhauser R., Pfau T., Ruske G. "Creating Hidden Markov Models for Fast Speech by Optimized Clustering", *Proceedings of Eurospeech Conference*, Budapest, Hungary, 1999, pages 407-410.

[4] Pfau T., Faltlhauser R., Ruske G. "Speaker Normalization and Pronunciation Variant Modeling: Helpful Methods for Improving Recognition of Fast Speech", *Proceedings of Eurospeech Conference*, Budapest, Hungary, 1999, pages 299-302.

[5] Tuerk A., Young S. "Modeling of Speaking Rate Using a Between Frame Distance Metric", *Proceedings of Eurospeech Conference*, Budapest, Hungary, 1999, pages 419-422.

[6] Rabiner L., Juang B.H. "Fundamentals of Speech Recognition", Prentice Hall PTR, Englewood Cliffs, New Jersey, 1993.

[7] Pfau T., Ruske G. "Estimating the Speaking Rate Using Vowel Detection", *Proceedings of ICASSP,* Seattle, Washington, 1998, pages 945-948.

[8] Faltlhauser R., Pfau T., Ruske G. "On the Use of Speaking Rate as a Generalized Feature to Improve Decision Trees", *Proceedings of ICSLP*, Peking, China, 2000, paper no. 748. vol. I, pages 317-320.

[9] Faltlhauser R., Pfau T., Ruske G. "On-line Speaking Rate Estimation Using Gaussian Mixture Models", *Proceedings of ICASSP*, Istanbul, Turkey, 2000, vol. 3, pages 1355-1358.

[10] Morgan N., Fosler E. "Combining Multiple Estimators of Speaking Rate", *Proceedings of ICASSP* Seattle, Washington, 1998, pages 729-732.

[11] Mirghafori N., Fosler E., Morgan N. "Towards Robustness to Fast Speech in ASR", *Proceedings of ICASSP*, Atlanta, Georgia, 1996, Vol.1, pages 335-338.

[12] Siegler M.A., Stern R.M. "On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems", *Proceedings of ICASSP*, Detroit, Michigan, 1995, pages 612-615.

[13] Pfau T., Ruske G. "Creating Hidden Markov Models for Fast Speech", *Proceedings of ICSLP*, Sydney, Australia, 1998, paper no. 255.