CREATING HIDDEN MARKOV MODELS FOR FAST SPEECH

T.Pfau, G.Ruske

Institute for Human-Machine-Communication, Technical University of Munich,
Arcisstr. 21, 80290 München
tel.: +49 89 289-28554, fax: +49 89 289-28535, e-mail: Thilo.Pfau@ei.tum.de

ABSTRACT

This paper deals with the problem of building HMMs suitable for fast speech. Fast speech leads to increased error rates on various tasks. In the first part of the paper an automatic procedure is presented to split speech material into different categories according to the speaking rate, which is fundamental for all investigations on the speaking rate.

In the second part the problem of sparse data available for the estimation of HMMs for fast speech is discussed. A comparison of different methods to overcome this problem follows. The main emphasis here is set on robust reestimation techniques like maximum aposteriori estimation (MAP) as well as on methods to reduce the variability of the speech signal and therefore to be able to reduce the number of HMM parameters. Vocaltract length normalization (VTLN) is chosen for that purpose. In the last part a comparison of various combinations of the methods discussed is presented basing on error rates for continuous speech recognition on fast speech. The best method (VTLN followed by MAP reestimation) results in an overall decrease of the error rate of 10% relative to the baseline system.

1. INTRODUCTION

The speaking rate (also called speech rate or rate of speech) is a main source of variability of speech signals. Not only interspeaker variability but also intraspeaker variability is the result, because in different situations speakers tend to vary the speaking rate. This variability results in a bad recognition performance especially for fast speech, when standard speech recognition systems are used (see figure 1).

There are two basic approaches to adapt speech recognition systems to the speaking rate which will be referenced here as explicit and implicit adaptation. Explicit adaptation means the process of measuring the rate of speech and then taking actions to adapt the recognizer. Implicit adaptation on the other hand is an adaptation method where the speaking rate must not be known. For example the overall likelihood of a processed utterance can be used to decide which of different available knowledge sources (e.g. speech rate specific HMMs or a speech rate specific pronunciation lexicon, which both have to be created in advance) to use. However, for both approaches it is necessary to create speech rate specific models in a first step to be able to study the effects of the speaking rate.

Various more or less successful attempts have been made to improve recognition performance on fast speech. A retraining of the acoustic models (MLP) on fast speech in [2,3,4] as well as changes in the transition probabilities of hidden Markov models in [1,2,3,4] led to decreased error rates on fast speech. Improvements of up to 25% relative to the baseline performance could be achieved. These investigations were made on TIMIT and WSJ tasks. In first experiments, using a maximum likelihood reestimation of the HMM parameters of our recognizer on the fast speech material of the german spontaneous scheduling task (Verbmobil), we were not able to reproduce these improvements. Therefore some approaches suitable to improve the recognition performance on this task will be discussed here.

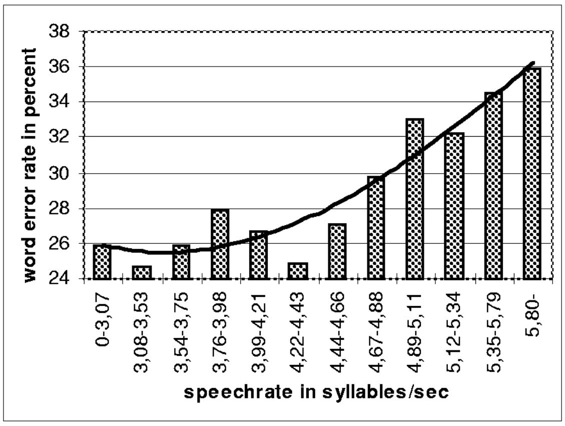

Figure 1: influence of the speaking rate on the word error rate of a large vocabulary continuous speech recognition system

2. AUTOMATIC SPLITTING OF THE SPEECH MATERIAL

For investigations on the speaking rate test and training data must be labelled according to the speaking rate. Thus the speaking rate for each sentence of the material must be determined. Basing on these rates, thresholds for splitting the material into different categories (e.g. slow, average and fast) can be fixed. As for most speech databases like for the german spontaneous scheduling task, which is used in our experiments, such labels are not available, we propose an automatic procedure to determine the speaking rate of spontaneous speech.

Our speech rate estimator [6] which is designed for online estimation of the speaking rate is based on detecting vowels (syllable nuclei) using both the modified loudness and the zerocrossing rate of each frame of a speech signal. To ensure compatibility with this measure we chose to split the material according to the syllable rate. However, using this estimator for splitting the training material is dangerous, as the estimator shows an error rate of about 22% on spontaneous speech. To be independent of the quality of the speech rate estimator when splitting the material according to the speaking rate, the actual speaking rate is better suited. It can be determined very easily when exact (manually created) transcriptions of the material are available. But manually transcribed data is only available for a small part of the database and is expensive to produce. Therefore an alternative automatic procedure is used to find suitable transcriptions for the material. This procedure is described now.

In an initial step „sharp" phoneme HMMs are created from a limited part of the database for which manually created transcriptions are available. After an initial LBG clustering procedure, continuous density phoneme HMMs are trained using maximum likelihood estimation (ML) on these transcriptions.

In a second step statistical pronunciation graphs are built for the training material. These graphs are built from the transliterations using a lexicon with several pronunciations per word. The resulting graphs are assigned with probabilities [7].

Then the „sharp" HMMs are used to perform a Viterbi segmentation on the pronunciation graphs for the whole material (test and training material). To improve the quality of the HMMs, the pronunciation graphs of the training material can be first used to estimate larger sized HMMs doing a ML estimation applying Viterbi training along these graphs. Then segmentations for the whole material can be found using these improved models.

Finally the results of the segmentation process can be used to determine the speaking rate of the sentences and to fix the thresholds to divide the material into the different bins.

3. THE PROBLEM OF SPARSE DATA

When acoustic models (here: HMMs) for specific speech rates have to be estimated, the main problem is the reduced amount of training data available. Thus the number of speaking rate categories has to be limited in order to guarantee a robust estimation of the model parameters. The two following aspects have to be considered when HMMs for specific speech rates shall be estimated.

First, as the utterances of the different categories show a reduced variability of speaking rates (this is the criterion for splitting!) the variation of the speech signal caused by different speaking rates is reduced within these categories. A limited number of parameters should hence be sufficient to model the properties of these specific acouctic models for each category adequately.

On the other hand the reduced amount of training data per category also results in a reduced number of training speakers for each category. Although the criterion for splitting is the speaking rate and not the speaker, the different categories contain different speakers just because the range of speaking rates shown by different speakers is not the same. This results in a reduced variation of the speech signals and therefore the resulting HMMs are not able to model the speaker specific variations adequately. The speech rate specific models are „less speakerindependent" than the speech rate independent models.

4. HIDDEN MARKOV MODELS FOR FAST SPEECH

In this section some basic approaches for robust parameter estimation on fast speech material are presented.

4.1. Maximum Likelihood Estimation

Standard maximum likelihood (ML) estimation is generally not well suited for a robust estimation of stochastic parameters on a reduced amount of data, as the new estimates only base on the reduced training material available.

4.2. Maximum Aposteriori Estimation

Maximum aposteriori (MAP) estimation has proven to be an efficient method to estimate parameters on sparse data [9]. Especially in speaker adaptation tasks this method has been used successfully in the past. Methods like MAP-VFS (vector field smoothing) and MAP-VFC (vector field correlation) are powerful methods for speaker adaptation with very few adaptation data [10,11]. The advantage of MAP based approaches is that some general models can be combined with specific training material. And as in MAP theory there is no difference between speaker specific and speech rate specific material, MAP reestimation is a promising approach for maintaining the general information of the speech rate independent HMMs and simultaneously capturing the speech rate specific effects on the speech rate specific training material.

4.3. Vocaltract Length Normalization

Vocaltract length normalization (VTLN) is a powerful method for freeing speech signals from influences of varying vocaltract lengths of different speakers. It is used successfully to improve recognition accuracy of LVCSR systems [12,13,14]. As described in section 3 the speech rate specific training material contains a smaller number of speakers. Speaker normalization techniques like VTLN can thus be helpful to make the speech rate specific models more speaker independent by freeing the spectrum of the speech signal from speaker specific influences. A reduced number of parameters should hence be sufficient for these models. Thus they should be a good basis for robust reestimation on a limited amount of training data.

4.4. VTLN and ML/MAP Estimation

ML reestimation on a limited amount of data should provide better results when performed on VTLN models for which a reduced number of model parameters is sufficient (MLVTLN). As the amount of speech rate specific data is strongly limited a combination of a robust estimation in combination with a reduction of the number of the HMM parameters seems to be advantageous for our purpose. Therefore a MAP-retraining (i.e. a more robust estimation) starting from speaker normalized models (i.e. reduced number of parameters) should be very well suited (MAPVTLN).

5. RESULTS

The training conditions used for the experiments described in this section are chosen according to the Verbmobil evaluation 1996. As test material a combination of the Verbmobil cross-validation set 1996 and the evaluation set 1996 is used as the evaluation set alone only contains about 50 sentences of fast speech, which is not significant to show effects on the error rates. Both test and training material were splitted into three categories (slow, average and fast) according to the procedure described in section 2. The thresholds were set to $\mu+\sigma$ for fast speech and to $\mu-\sigma$ for slow speech as calculated on the training set. Recognition was performed using our standard HMM speech recognition system [8]. For the VTLN based training and recognition a ML based frequency warping is performed directly on the linear spectrum (FFT) in our standard preprocessing unit [5].

The whole training material consists of about 10 mio frames of spontaneous scheduling dialogues of about 600 speakers whereas the reduced training material for fast speech only consists of 800.000 frames of about 400 speakers, which is a reduction of about 90% in the number of frames and a reduction of about 33% in the number of speakers.

In the first experiment three iterations of ML and MAP estimation were performed on the fast training material. Figure 3 shows the average training scores per frame when trained on fast speech. As a reference the line indicates the average score obtained on the whole material with the baseline models. After three iterations the average score for ML estimation outperforms this value whereas using MAP estimation this value is not yet reached. This is a comprehensible result as the MAP reestimated parameters do not change as fast as the ML reestimated parameters because of the combination of the baseline models with the specific training material. The bad scores during the first training iteration show the mismatch between the unspecific models (which are used in the first training iteration) and the fast speech material.

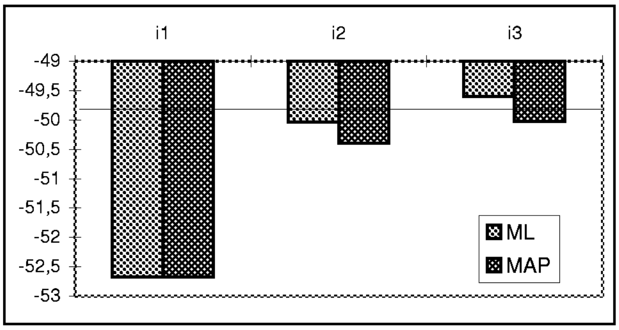

Figure 3: average training scores per frame using ML and MAP reestimation on fast speech

Table 1 shows the recognition rates of speech rate independent continuous density phoneme HMMs with about 1.5 mio acoustic parameters. These models are referenced as standard or baseline models in the following.

WER	substitutions	deletions	insertions
45,3%	31,1%	10,2%	3,9%

Table 1: recognition results using the standard HMMs.

Table 2 shows the recognition performance of the ML reestimated models of the first two iterations. Like in all other reestimation procedures described here, all acoustic parameters are retrained, i.e. means and diagonal covariance matrices of the gaussian densities as well as mixture weights and transition probabilities of the states of the phoneme HMMs. However, using the ML reestimated models of the first training iteration the recognition performance is only improved slightly. Using the models of the second iteration even results in an increased error rate. This is clearly due to the lack of robustness using ML estimation, the parameters are overadapted to the limited training material.

iter	WER	substitutions	deletions	insertions
1	44,4%	30,6%	9,8%	4,0%
2	44,8%	30,9%	9,6%	4,3%

Table 2: recognition results using ML reestimated HMMs

In table 3 the recognition performance on the fast sentences using MAP reestimation is shown. Using the first iteration models, the error rate can be reduced by 4,9% relative to the standard models. Using the second iteration models again reduces the error rates slightly. An overall improvement of 6,6% relative to the baseline models can be achieved. This effect can be explained by the influence of the parameters of the standard models which weakens from iteration to iteration during training. After two MAP training iterations the models have still preserved enough general information and they have already collected enough speech rate specific information.

iter	WER	substitutions	deletions	insertions
1	43,1%	29,9%	9,9%	3,3%
2	42,3%	30,0%	8,9%	3,4%

Table 3: recognition results using MAP reestimated HMMs

In a second experiment the standard models were normalized performing five iterations of VTLN training. Then the normalized models were retrained using ML and MAP reestimation.

In table 4 the results achieved with speaker normalized models trained on the whole material are presented. These standard VTLN models are 8,8% relative better than the standard models on the fast material, whereas they are only 4,5% relative better on the whole test set. Thus, VTLN reduces the confusability in the feature space especially for fast speech.

WER	substitutions	deletions	insertions
41,3%	28,9%	8,3%	4,1%

Table 4: recognition results using standard VTLN HMMs

Tables 5 and 6 contain the recognition results for ML and MAP reestimated VTLN models respectively. The performance of the ML trained models is not significantly better than the performance of the VTLN standard models. A slight improvement in recognition performance is achieved using one iteration of MAP reestimation.

iter	WER	substitutions	deletions	insertions
1	41,2%	29,1%	8,3%	3,8%
2	41,6%	29,3%	8,2%	4,1%

Table 5: recognition results using ML reestimated VTLN HMMs

iter	WER	substitutions	deletions	insertions
1	40,7%	28,1%	8,4%	4,2%
2	41,2%	28,6%	8,4%	4,1%

Table 6: recognition results using MAP reestimated VTLN HMMs

6. CONCLUSION

In this paper the problem of training HMMs for fast speech is discussed. First an automatic method for splitting the sentences of test and training material into different categories according to the speaking rate is described.

In the second part the problem of sparse data is discussed and some methods are described to robustly estimate the parameters of HMMs suitable for fast speech.

In the last part recognition results using the methods described are presented. It is shown that MAP reestimation is better suited to estimate parameters on a limited amount of data than ML estimation. The MAP reestimated models result in a 6,6% relative decreased error rate compared to the standard models. Furthermore the VTLN is shown to be a proper way of reducing the confusability of hidden Markov models especially on fast speech. The error rates on the fast speech are reduced by 8,8% relative to the standard models when VTLN is used. A MAP reestimation of the VTLN HMMs results in a further reduction of the error rate to about 10,2% relative to the standard models.

Although the improvements presented here are not as high as results on other tasks [3] they are encouraging to take further steps to adapt speech recognition systems to fast speech, having in mind that only the acoustic models were changed. Other important knowledge sources like the pronunciation models were not touched so far. Training was performed on the canonical transcriptions which are also used in the test lexicon during evaluation. Especially for fast speech a modification of the pronunciation models seems to be promising and can be combined easily with the methods described here.

7. ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the Verbmobil Project.

8. REFERENCES

1. Morgan N., Fosler E., Mirghafori N., *Speech Recognition Using On-line Estimation of Speaking Rate*, Proc. EUROSPEECH ’97, pp. 2079-2082, Rhodes, Greece, September 1997.
2. Mirghafori N., Fosler E., Morgan N., *Towards Robustness to Fast Speech in ASR*, Proc. ICASSP ’96, Vol.1, pp. 335-338, Atlanta, Georgia, May 1996.
3. Mirghafori N., Fosler E., Morgan N., *Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes*, Proc. EUROSPEECH ’95, pp. 491-494, Madrid, Spain, September 1995.
4. Siegler M.A., Stern R.M., *On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems*, Proc. ICASSP ’95, pp. 612-615, Detroit, Michigan, May 1995.
5. Ruske G., Beham M., *Gehörbezogene automatische Spracherkennung*. In: „Sprachliche Mensch-Maschine-Kommunikation“, (H. Mangold, Hrsg.), Oldenbourg-Verlag, pp. 33-47, München Wien, 1992.
6. Pfau T., Ruske G., *Estimating the Speaking Rate by Vowel Detection*, Proc. ICASSP ’98, pp. 945-948, Seattle, Washington, May 1998.
7. Kipp A., Wesenick M.-B., Schiel F., *Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech*, Proc. EUROSPEECH ’97, pp. 1023-1026, Rhodes, Greece, September 1997.
8. Plannerer B., Einsele T., Beham M., Ruske G., *A continuous speech recognition system integrating additional acoustic knowledge sources in a data-driven beam search algorithm*, Proc. ICSLP ’94, pp. 17-20, Yokohama, Japan, September 1994.
9. Lee C.-H., Gauvain J.-L., *Speaker Adaptation Based on MAP Estimation of HMM Parameters*, Proc. ICASSP ‘93, pp. 558-561, Minneapolis, Minnesota, April 1993.
10. Takahashi J., Sagayama S., *Vector-Field-Smoothed Bayesian Learning for Incremental Speaker Adaptation*, Proc. ICASSP ’95, pp. 696-699, Detroit, Michigan, May 1995.
11. Takahashi S., Sagayama S., *Tied-Structure HMM Based on Parameter Correlation for Efficient Model Training*, Proc. ICASSP ’96, pp. 467-470, Atlanta, Georgia, May 1996.
12. Eide E., Gish H., *A Parametric Approach to Vocal Tract Length Normalization*, Proc. ICASSP ’96, pp. 346-348, Atlanta, Georgia, May 1996.
13. Zhan P., Westphal M., *Speaker Normalization Based on Frequency Warping*, Proc. ICASSP ’97, pp. 1039-1042, Munich, Germany, April 1997.
14. Lee L., Rose R.C., *Speaker Normalization Using Efficient Frequency Warping Procedures*, Proc. ICASSP ’96, pp. 353-356, Atlanta, Georgia, May 1996.