6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEAKER NORMALIZATION AND PRONUNCIATION VARIANT MODELING: HELPFUL METHODS FOR IMPROVING RECOGNITION OF FAST SPEECH

Thilo Pfau, Robert Faltlhauser, Günther Ruske

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

ABSTRACT

The presented paper addresses the problem of creating hidden Markov models for fast speech. The major issues discussed are robust parameter estimation and reducing within-model variations. Regarding the first issue, the use of the maximum a posteriori parameter estimation is discussed. To reduce within-model variations, a maximum likelihood based vocal tract length normalization procedure and a statistical approach to model pronunciation variants are applied.

Experiments with a large vocabulary continuous speech recognition system were carried out on the German spontaneous scheduling task (Verbmobil) to prove the effectiveness of the investigated methods. The results show that a combination of pronunciation variant modeling and vocal tract length normalization is most effective. On fast speech, a relative improvement of 16.3% compared to the baseline models was achieved. Pronunciation variant modeling combined with the maximum a posteriori reestimation proved to be the second best method resulting in a 14.9% relative improvement. In addition, this combination does not cause any additional computational load during recognition.

10.21437/Eurospeech.1999-78

1. INTRODUCTION

A major problem in automatic speech recognition is the variability of speech signals. Identical phonemes uttered in different situations and/or by different speakers result in variable speech signals. These acoustic variations have to be represented by the parameters of the acoustic models (hidden Markov models, HMMs) used by the recognition system. A large amount of training material has to be used for estimating the parameters of the acoustic models in order to ensure the recognition process to be speaker and situation independent. However, the recognition performance for certain speakers and situations can still be much worse than the average. To solve this problem, a lot of research was done on speaker adaptation and speaker normalization methods in the last few years.

Apart from speaker specific variations, the speaking rate is another source of variability which leads to increased error rates on fast speech using standard speech recognition systems. This effect was observed on TIMIT and WSJ tasks ([3], [4]). To counteract the observed degradation in performance several approaches were investigated: a retraining of the acoustic models on fast speech ([2], [3] and [4]) and changes in the transition probabilities of hidden Markov models ([1], [2], [3] and [4]) proved to be suitable methods to decrease error rates on fast speech. The reported methods resulted in improvements of up to 25% relative to the baseline performance. Previous experiments performed by the authors on the German spontaneous scheduling task (Verbmobil) showed robust reestimation techniques and speaker normalization to increase recognition accuracy on fast speech [5]. In these experiments the error rate was decreased by 4.3% using robust reestimation methods. A combination of both, robust reestimation and speaker normalization, resulted in a 10% improvement. The objectives of the presented project were to further improve the accuracy of automatic speech recognition systems on fast spontaneous speech.

The main focus of this paper is the use of pronunciation variants for reducing within-model variations especially for fast speech. Additionally the combination of robust reestimation with the reduction of within-model variations is evaluated.

2. HIDDEN MARKOV MODELS FOR FAST SPEECH

A major problem for the estimation of acoustic models (HMMs) for specific speaking rates is the limited amount of training data available. In order to retain a sufficient amount of training data for each category, the number of "speaking rate categories" has to be limited. Since the utterances of the different categories show a reduced variability of speaking rates (the speaking rate is the criterion for splitting the material!), the speaking rate related variations of the speech signals are reduced within these categories. A limited number of parameters should hence be sufficient to model the HMM parameters for each category adequately.

However, the reduced amount of training data per category results in a reduced variation with respect to other sources of variability. Although the criterion for the split is the speaking rate and not the speaker's identity, the different categories implicitly contain utterances of different speakers just because the speaking rate ranges of the speakers differ. Since the speakers have different characteristics these characteristics are not sufficiently covered by the reduced training material for each category. If this effect is not taken into consideration while estimating speech rate specific HMMs, the resulting models are "less speaker independent".

Another source of variability is the use of pronunciation variants. In general, the reduced speech rate specific amount of training material contains a restricted subset of words with a restricted subset of pronunciation variants. In standard speech recognition systems the acoustic variations caused by different pronunciation variants have to be modeled by the parameters of the subword HMMs of the canonical pronunciation (during training and recognition only the canonical transcription of the words are considered). On the reduced training material, not all possible variants can be learned adequately by the

parameters of the corresponding subword HMMs. Therefore the resulting models are "less pronunciation variant independent".

2.1 Robust Parameter Estimation

2.1.1 Maximum A Posteriori Estimation (MAP)

The first approach discussed is the use of a parameter estimation algorithm which is more robust than the conventional maximum likelihood (ML) estimation. Since not all sources of variability are sufficiently represented within the speech rate specific categories, the use of ML training on the speech rate specific material alone will result in an overadaptation to the training data, which increases error rates during recognition.

An alternative procedure is the combination of general models (which are representative with respect to several sources of variability) with speech rate specific models, which are reestimated on speech rate specific speech material. This approach is comparable to speaker adaptation, where "large" speaker independent models are reestimated using the limited training material available for a certain speaker. In speaker adaptation maximum a posteriori (MAP) related approaches are frequently used ([7] and [8]).

The following MAP equations ([9]) are used to reestimate the means, the diagonal covariance matrices and the mixture weights of the context independent phoneme HMMs on the category "very fast" of the training material:

mean: $$\tilde{m}_{ik} = \frac{\tau_{ik} m_{ik} + \sum_{t=1}^{T} c_{ikt} x_t}{\tau_{ik} + \sum_{t=1}^{T} c_{ikt}}$$

variance: $$\tilde{r}_{ik}^{-1} = \frac{\beta_{ik} + \sum_{t=1}^{T} c_{ikt} (x_t - \tilde{m}_{ik}) \cdot (x_t - \tilde{m}_{ik})^t}{\alpha_{ik} - p + \sum_{t=1}^{T} c_{ikt}}$$

mixture weight: $$\tilde{w}_{ik} = \frac{\nu_{ik} - 1 + \sum_{t=1}^{T} c_{ikt}}{\sum_{k=1}^{K} \nu_{ik} - K + \sum_{k=1}^{K} \sum_{t=1}^{T} c_{ikt}}$$

The indices i and k refer to the k-th mixture of the i-th state and p equals the dimension of the feature vector. Similar to other applications of the MAP approach ([7], [9]) a fixed common prior parameter τ is used to make the reestimation of the HMM-parameters more robust. The other prior parameters (α, β and ν) can be computed from the τ values [9]. The HMM transition probabilities, which are not critical to estimate, are reestimated using standard ML equations.

2.2 Reducing Within-Model Variations

The use of the restricted MAP approach (one common prior parameter τ) for reestimating the HMM parameters on the speech rate specific material is suboptimal, as the variable sensitivity of the distributions with respect to speech rate is not considered. On the other hand the unrestricted MAP approach requires individual prior parameters for each distribution of the HMMs which are difficult to determine.

If the HMMs are freed from modeling variations caused by other sources of variability, as presented in an alternative approach, it is no longer important whether or not these variations are sufficiently represented in the speech rate specific training material. As a result the model parameters can fully "concentrate" on modeling the speech rate relevant variations which can be optimized in an additional reestimation step.

Two sources of variability are taken into consideration:

- the length of the vocal tract,
- the use of pronunciation variants.

2.2.1 Vocal Tract Length Normalization (VTLN)

One major speaker specific source of variability of speech signals is the length of the vocal tract. The main effect is the shifting of formant frequencies in vowel regions which are shifted down for longer vocal tract lengths and shifted up for shorter ones. Several approaches for vocal tract length normalization for continuous speech recognition have been published ([10], [11], [12] and [13]).

In this study a ML based vocal tract length normalization procedure (comparable to [11]) was used during training and recognition.

2.2.2 Pronunciation Variant Modeling (PROVAR)

A second important source of variability is the use of different pronunciation variants for one word in different "situations". In the following, the approach applied for pronunciation variant modeling is described in detail.

Transcriptions and hidden Markov models:

For the creation of pronunciation variant based phonemic transcriptions a training dictionary with an average number of 2.35 pronunciation variants per word was used. This dictionary contains up to 70 variants per word and is described in detail in [14].

Based on the dictionary and the transliterations of the utterances, the pronunciation variants of each word were included as parallel paths in pronunciation graphs. After the creation of preliminary "sharp" HMMs on a limited amount of manually transcribed data, a Viterbi segmentation procedure was performed on the whole training material using these pronunciation graphs. In a second step the resulting phonetic transcriptions were used to estimate new HMMs with an increased number of parameters. The training material was divided into two equal parts with the average log probability of the Viterbi segmentation process as a splitting criterion. Only those utterances which resemble the preliminary models closely (i.e. show a high log probability) were used for the creation of the new models in the second step. With the second step models, another Viterbi segmentation process was performed along the graphs. With these resulting final transcriptions several iterations of Viterbi training were carried out to estimate the final HMM parameters on the whole training set.

Finally these HMMs were reestimated on the speech rate specific training material applying MAP reestimation to derive speech rate specific pronunciation variant HMMs.

Test dictionary:

In general, multiple pronunciation variants per word result in increased confusions in the search space, as the pronunciation variants of different words become more similar. Several approaches for the creation of pronunciation variant test dictionaries of reasonable size can be found in literature ([15], [16]). The strategy presented in this paper is simple and similar to [14]: only those variants were integrated in the test dictionary which occur at least once in the final transcriptions of the training set. Therefore the average number of pronunciation variants per word was reduced from 2.35 to 1.8 (9608 variants for 5329 words) and the maximum number of variants per word was 38.

The confusion during recognition was further reduced by weighting the variants in the test dictionary. Therefore the a posteriori probabilities of each variant given the word were needed. The variant weights were normalized to sum up to 1.0 for all variant weights of a word. During recognition the logarithm of this weight was multiplied with a predefined pronunciation factor and then added to the score of the Viterbi path whenever the end of a variant was reached. This strategy resembles the integration of the language model score into the score of the Viterbi path.

3. RESULTS

3.1 Experimental Conditions

The experiments described were performed on the evaluation set 1996 of the German spontaneous scheduling task with a total number of 343 sentences of different speech rates including 53 "very fast" sentences. The adaptation strategy used for different speaking rates is "explicit adaptation", i.e. a switching between speech rate specific models, which were estimated on the speech rate specific training material in advance. Test and training material were split into different speech rate specific categories under the assumption of an ideal speech rate detector. The speech rates of the utterances were determined as described in [5]. The five speech rate categories "very slow", "slow", "average", "fast" and "very fast" were defined.

Baseline System (BASELINE)

The baseline recognition system was a HMM-based Viterbi recognizer using 52 context-independent continuous density HMMs for 44 single phonemes (including noise and silence models) and 8 phoneme combinations. The probability density functions of the HMM states were modeled by Gaussian mixture densities with diagonal covariance matrices and were estimated with standard Viterbi ML estimation. The pre-processing unit provided loudness based values for each of 20 acoustic channels on the Bark scale up to 8 kHz. These values were computed every 10 ms over a Hamming window of 16 ms length. This first data set was completed by energy and zerocrossing rate and contained 22 features. First and second order derivatives of this first data set were added to form the 66-dimension final feature vector. The search engine used a tree shaped canonical dictionary with 5329 words and a bigram language model.

Pronunciation Variant Based System (PROVAR)

For the pronunciation variant based recognizer the phoneme set was reduced to 44 phonemes by removing the phoneme combinations. The acoustic preprocessing, the methods for modeling and estimating the probability density functions and the language model were not altered, but the pronunciation variant based test dictionary of section 1.1.2 was used.

Normalized Systems (VTLN and PROVAR-VTLN)

A VTLN version of the BASELINE and of the PROVAR system are created using the ML based VTLN approach.

Speech Rate Specific Systems (MAP and PROVAR-MAP)

The HMM-parameters of the BASELINE system and the PROVAR system are reestimated on the "very fast" training material with MAP estimation.

3.2 Word Error Rates

To be able to compare the proposed methods, word error rates (WER), substitutions (sub.), insertions (ins.) and deletions (del.) on the complete test set and on the speech rate specific categories are presented. The relative improvement in percent compared to the BASELINE system is given in the last column of Table 2 to Table 6. For the PROVAR derived systems (PROVAR-MAP and PROVAR-VTLN) the performance is compared to the original PROVAR system additionally (Table 4 and Table 6).

test set	WER	sub.	del.	ins.
total	36.5	25.6	5.6	5.3
very slow	28.4	18.3	2.3	7.8
slow	36.6	25.4	4.6	6.6
middle	35.2	24.8	5.3	5.1
fast	37.4	26.7	6.7	4.0
very fast	47.0	33.8	10.1	3.1

Table 1: Recognition performance (in percent) on the complete test set and on the different speech rate specific categories using the BASELINE system.

test set	WER	sub.	del.	ins.	rel. imp. to BASELINE
total	33.8	23.1	4.5	6.2	7.4
very slow	26.2	16.8	1.0	8.4	7.7
slow	34.1	21.9	4.8	7.4	6.8
middle	32.5	22.2	3.8	6.5	7.7
fast	36.0	26.3	5.2	4.5	3.9
very fast	42.4	30.4	8.7	3.3	9.8

Table 2: Recognition performance (in percent) on the complete test set and on the different speech rate specific categories using the VTLN system.

test set	WER	sub.	del.	ins.	rel. imp. to BASELINE
total	35.2	23.6	8.1	3.5	3.6
very slow	23.1	13.9	2.7	6.5	18.7
slow	35.1	24.0	5.4	5.7	4.1
middle	34.0	24.0	5.6	4.4	3.4
fast	38.5	27.7	5.5	5.3	-2.9
very fast	41.8	29.8	9.8	2.2	11.1

Table 3: Recognition performance (in percent) on the complete test set and on the different speech rate specific categories using the PROVAR system.

test set	WER	sub.	del.	ins.	rel. imp. to PROVAR	rel. imp. to BASELINE
total	33.6	22.5	6.3	4.8	4.5	7.9
very slow	22.6	15.2	2.3	5.1	2.2	20.4
slow	34.6	22.5	6.6	5.5	1.4	5.5
middle	32.7	22.2	5.2	5.3	3.8	7.1
fast	36.6	25.0	7.7	3.9	4.9	2.1
very fast	39.3	26.3	11.3	1.7	6.0	16.3

Table 4: Recognition performance (in percent) on the complete test set and on the different speech rate specific categories using the PROVAR-VTLN system.

WER	sub.	del.	ins.	rel. imp. to BASELINE
44.3	31.3	10.4	2.6	5.7

Table 5: Recognition performance (in percent) on the "very fast" subset using the MAP system.

WER	sub.	del.	ins.	rel. imp. to PROVAR	rel. imp. to BASELINE
40.0	26.9	11.7	1.4	4.5	14.9

Table 6: Recognition performance (in percent) on the "very fast" subset using the PROVAR-MAP system.

4. DISCUSSION

Comparing the BASELINE and the PROVAR system (Table 1 and Table 3) it is evident that the use of pronunciation variants, which reduces the within-model variability of the HMM parameters, is suitable to reduce error rates for extreme speech rates. The highest relative improvements can be observed for the extreme categories "very slow" and "very fast".

For the VTLN system (Table 2) the highest relative improvement of 9.8% is obtained in the category "very fast". When comparing the word error rates of the PROVAR and the PROVAR-VTLN system (Table 3 and Table 4) again the highest relative improvement of 6.0% can be found in the category "very fast" resulting in an overall improvement of 16.3% compared to the BASELINE system.

Regarding the MAP retraining on the category "very fast", a slightly worse relative improvement of 4.5% (Table 6) is achieved when using the PROVAR models instead of the canonical models for which a 5.7% relative improvement was obtained (Table 5). When evaluating this result the very low error rate of 41.8% for the baseline PROVAR system must be considered. The best result with the MAP approach achieved an overall improvement of 14.9% using the PROVAR-MAP system which presents the second highest improvement of the proposed methods. An advantage of this method compared to the best method (PROVAR-VTLN) is, that no additional computational load is caused during recognition.

No further improvements are observed after application of a MAP retraining to the VTLN based models (VTLN and PROVAR-VTLN). Additional investigations have to be made to clarify the cause for this effect.

With regard to the pronunciation variant modeling individual speech rate specific variant weights instead of the speech rate independent weights are expected to result in further improvements. These individual weights would allow to take speech rate specific pronunciation variants into consideration. However, the problem of this approach is to ensure a robust estimation of the weights on the limited amount of speech rate specific training material.

5. ACKNOWLEDGEMENTS

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

6. 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] Pfau T., Ruske G., *Creating Hidden Markov Models for Fast Speech*, Proc. ICSLP '98, paper no. 255, Sydney, Australia, November/December 1998.

[6] Pfau T., Ruske G., *Estimating the Speaking Rate by Vowel Detection*, Proc. ICASSP '98, pp. 945-948, Seattle, Washington, May 1998.

[7] Takahashi J., Sagayama S., *Vector-Field-Smoothed Bayesian Learning for Incremental Speaker Adaptation*, Proc. ICASSP '95, pp. 696-699, Detroit, Michigan, May 1995.

[8] 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.

[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] Eide E., Gish H., *A Parametric Approach to Vocal Tract Length Normalization*, Proc. ICASSP '96, pp. 346-348, Atlanta, Georgia, May 1996.

[11] Zhan P., Westphal M., *Speaker Normalization Based on Frequency Warping*, Proc. ICASSP '97, pp. 1039-1042, Munich, Germany, April 1997.

[12] Lee L., Rose R.C., *Speaker Normalization Using Efficient Frequency Warping Procedures*, Proc. ICASSP '96, pp. 353-356, Atlanta, Georgia, May 1996.

[13] Westphal M., Schultz T., Waibel A., *Linear Discriminant – A New Criterion For Speaker Normalization*, Proc. ICSLP '98, paper no. 755, Sydney, Australia, November/ December 1998.

[14] Schiel F., Kipp A., Tillmann H.G., *Statistical Modeling of Pronunciation: it's not the Model, it's the Data*, Proc. Modeling Pronunciation Variation for Automatic Speech Recognition, pp.131-136, Rolduc, the Netherlands, May 1998.

[15] Finke M., Waibel A., *Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition*, Proc. Eurospeech '97, pp. 2379-2382, Rhodes, Greece, September 1997.

[16] Sloboda T., *Dictionary Learning: Performance through Consistency*, Proc. ICASSP '95, pp. 453-456, Detroit, Michigan, May 1995.