A Combination of Speaker Normalization and Speech Rate Normalization for Automatic Speech Recognition

T.Pfau, R. Faltlhauser, and G. Ruske

Inst. for Human-Machine-Communication, Technische Universität München, Arcisstr. 21, D-80290 Munich, Germany
tel.: +49 89 289-28554, fax: +49 89 289-28535, email: {Pfau, Faltlhauser, Ruske}@ei.tum.de

ABSTRACT

In this contribution a normalization procedure for automatic speech recognition is introduced which aims at reducing speaking rate specific variations of the features of the phonetic classes. A "spurtwise" calculation of normalization factors allows to capture changes of the speaking rate within one utterance. The cost-saving implementation using linear interpolation of the original features and a word graph rescoring procedure leads to a moderate increase in computational load compared to the baseline system without speech rate normalization.

In addition a two-step procedure which combines vocal tract length normalization (VTLN) and speech rate normalization (SRN) has been developed. Experiments showed, that applying SRN to a VTLN-based recognition system leads to relative reduction in word error rate of 4.2%. This is comparable to the decrease observed when using SRN on a system without VTLN. All in all the combination of VTLN and SRN results in a 15% reduction of word error rate compared to the baseline system.

1. INTRODUCTION

The speaker independent processing of spontaneously spoken human-to-human dialogues is a special challenge to present automatic speech recognition systems. Several factors contribute to additional variations of speech signals in comparison to read speech where the speaking mode is generally well defined. These additional variations result in higher confusions between the phonetic classes of the pattern matching process consequently to an increase in word error rates of large vocabulary continuous speech recognition (LVCSR) systems.

Among other factors like e.g. the use of pronunciation variants, the speaking rate is a source of variation which can lead to a considerable increase in error rate. Especially when people talk faster than normal the performance of state of the art LVCSR systems is often poor with error rates for fast speech being twice to four times higher than normal (e.g. [1], [2], [3], [4], [5], [6]). However, we observed a 30 to 40% increase in error rate on both the Verbmobil evaluation and crossvalidation set 1996 ([7], [8]).

Several approaches have been suggested to improve recognition of fast speech. Changes of hidden Markov model (HMM) state transition probabilities ([1], [9] and [3]) as well as an adaptation of a neural net phonetic probability estimator ([1], [9]) proved to be suitable methods. In earlier studies we showed the usefulness of maximum aposteriori (MAP) estimation to adapt the acoustic models (HMMs) to fast speech ([10]). In addition it was demonstrated that the use of pronunciation variants and a maximum likelihood based approach to vocal tract length normalization (VTLN) were helpful for this purpose ([8]).

The basic idea of normalization procedures in general −in contrast to adaptation– is to reduce the variations which have to be captured within the acoustic model of each phonetic class of the recognizer. By reducing the variations of the feature vectors of the phonetic classes, the distributions of the acoustic models in the feature space will get "sharper". Thus the normalization procedure will reduce the overlap between the different "normalized" phonetic classes and thus the confusion of the recognizer. As a consequence the normalization procedure has to be applied during parameter estimation to create "normalized" models as well as during testing.

In previous studies ([8]) we were able to show that the reduction of speaker specific variations as well as the reduction of phonetic variations (use of pronunciation variants) can be helpful to reduce the error rates for fast speech. We assume that the combination of speaker specific, phonetic and speech rate specific variations is especially adverse for fast speech. A normalization with respect to one of the different sources of variation is advantageous for extreme conditions (e.g. fast speech). In our experiments we showed that VTLN is especially effective for fast speech, whereas the use of pronunciation variants proved to have potential to reduce error rates for both slow and fast speech.

In this study we will present a method of combining two different approaches to normalize the feature vectors extracted from the speech signal. Apart from the well known vocal tract length normalization (see section 3.2) to reduce speaker specific variations, a modified approach to speech rate normalization (SRN) is introduced (see section 3.1). Whereas the first approach aims at reducing "secondary" (=not related to the speaking rate) variations the second approach results in a reduction of "primary" (=speech rate specific) variations of the feature vectors. A special emphasis is put on the combination of normalization (see section 3.3) with respect to both primary and secondary variations.

2. METHODS

2.1 Speech Rate Normalization (SRN)

Principle: The idea of speech rate normalization is introduced in [11] and is called cepstrum length normalization (CLN). The pronciple is to normalize the phone duration by stretching the length of the utterance in order to match the "standard" lengths of the phonetic units which are obtained from the training corpus. It is based on the observations of [5] that the dynamic features (first and second order derivatives of the spectral features) are most affected by changes of the speaking rate.

Figure 1 illustrates the use of the same 10 ms frame grid for three different speaking rates (slow, medium and fast). Especially the calculation of the first and second order derivatives (delta- and

From: Proc. ICSLP 2000, IEEE

deltadelta-features) is affected. Due to the lengthening or shortening of the phonetic units for the different speaking rates these dynamic features can belong to different phonetic units. The "phonetic distance" between the static features and the dynamic features changes. Whereas all kinds of features of the central frame of the phoneme "u:" (see figure 1) are assigned to this phonetic unit ("u:") for slow speech, the dynamic features are assigned to neighboring units ("g", "t" and "@") for fast speech. This results in essential differences between speech rate specific versions of the phonetic units.

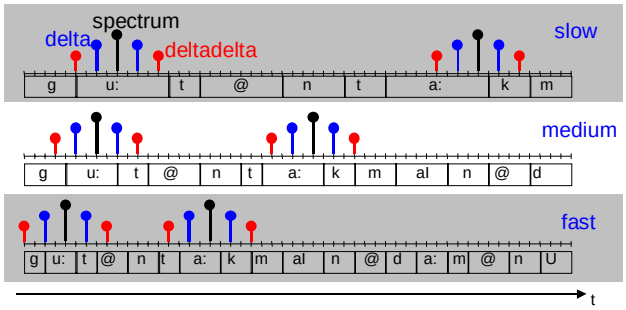

Figure 1: Illustration of the calculation of static (spectrum) and dynamic (delta and deltadelta) features for different speaking rates for the first part of the utterance "guten Tag meine Damen und Herren" (engl. "good afternoon ladies and gentlemen") using a constant frame grid. The transcription is given in SAMPA.

To overcome this problem the frame grid for calculating the dynamic features has to be adjusted to the actual speaking rate. This is illustrated in figure 2. A broader grid is used for slow speech and a narrower one for fast speech. Thus the dynamic features are calculated using similar parts of the speech signal and the "phonetic distance" should remain constant for different speaking rates.

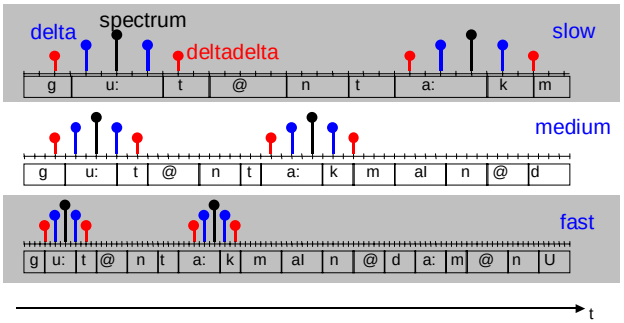

Figure 2: Illustration of the calculation of static (spectrum) and dynamic (delta and deltadelta) features for different speaking rates for the first part of the utterance "guten Tag meine Damen und Herren" (engl. "good afternoon ladies and gentlemen") using a frame grid which is adjusted to the actual speaking rate. The transcription is given in SAMPA.

Finding an optimal frame grid: Different algorithms for the determination of a new frame grid are evaluated in [11]. Whereas a phone-by-phone stretching results in high improvements in performance in supervised mode, this method completely fails in unsupervised mode. Therefore the authors suggest to apply a sentence-by-sentence procedure using an average factor ρ_{mid}:

$$\rho_{mid} = \frac{1}{N}\sum_{i=1}^{N}\rho_i \qquad (1)$$

This mean factor is obtained by averaging all N phone-by-phone normalization factors ρ_i which are calculated from a gamma distribution estimated on the training material.

speech rate categories per utterance	number of utterances on the training material
"fast" only	1194
"medium" only	5485
"slow" only	140
"medium" and "fast"	1576
"medium" and "slow"	1953
"slow" and "fast"	64
all categories	877
no valid spurt	66
sum	11355

Table 1: Intra-utterance variation of the speaking rate on the training material of the Verbmobil evaluation 1996.

This sentence-by-sentence procedure cannot be used to model changes of the speaking rate within one sentence. However, we found that a considerable number of utterances within the training material contains speech of different speech rate categories. Table 1 shows the distribution of spurts (spurt=between pause region, see [12]) of different speech rate categories for the utterances of the training material of the Verbmobil evaluation 1996. A separate value of the speaking rate was determined for each spurt. For 60% (6820 sentences) of the training material the speaking rate remains constant, whereas 40% of the sentences contain spurts of different speech rate categories. For these sentences, it is important to calculate different normalization factors for each spurt. Therefore we decided to use a spurtwise average of the normalization factor ρ_i. In addition preliminary experiments showed that it is favorable to calculate this spurtwise factor by summing up the expected and the actual lengths of the phones instead of calculating an average quotient using equation 1.

$$\rho_{sum} = \frac{\sum_{i=1}^{N} d_{exp,i}}{\sum_{i=1}^{N} d_{act,i}} \qquad (2)$$

Calculating normalized features: After having determined the new optimal frame grid, speech rate normalized features are calculated. First, the static components on the new grid are interpolated using the static features of the original 10 ms grid. A simple linear interpolation proved to be sufficient. Second, new dynamic features are calculated using the static features of the new grid.

For the estimation of speech rate normalized HMMs during training the phonetic transcription of each spurt is used to determine the optimal normalization factor.

For normalizing the feature vectors during recognition both a supervised and an unsupervised procedure were implemented. Whereas in supervised mode the transcription is assumed to be known, in unsupervised mode a phone recognizer is used to determine an estimate $sr_{est,sp}$ of the speaking rate of each spurt sp. In addition the mean $\overline{sr}_{est}$ of the spurtwise estimates on the training material is calculated to be able to determine approximate values for the normalization factors of each spurt:

$$\rho_{sum,approx,sp} = \frac{sr_{est,sp}}{\overline{sr}_{est}} \qquad (3)$$

Recognition: Using both supervised or unsupervised SRN, the speech rate normalized feature vectors are applied to perform a second recognition pass using a lattice rescoring algorithm.

1. In the first recognition pass a word graph is produced on the basis of unnormalized feature vectors and unnormalized acoustic models. In unsupervised mode, a phone recognizer is run simultaneously.
2. In a second pass, the speech rate normalized feature vectors and the speech rate normalized HMMs are used to rescore the word graph.

2.2 Vocal Tract Length Normalization (VTLN)

Principle: The VTLN is a well known approach to speaker normalization, which aims at reducing speaker specific variations of the speech signal caused by different lengths of the vocal tract. Different approaches can be found in the literature which differ both in the warping-functions used to transform the original spectrum into the normalized spectrum as well as in the way of finding an optimal value of the warping factor α which is commonly used to define the amount of warping (e.g. [13], [14], [15], and [16]).

Warping procedure: For the training of speaker normalized acoustic models we used a ML-based approach similar to [14]. The warping is performed using a piecewise linear function for transforming the speaker specific into the normalized spectrum.

During recognition either supervised or unsupervised VTLN can be performed. Whereas in supervised mode the transcription is assumed to be known when choosing the optimal warping factor, in unsupervised mode a Gaussian mixture model (GMM) is used to choose the warping factor [15].

2.3 Combination of VTLN and SRN

The combination of VTLN and SRN is implemented in a two-step procedure. First, an optimal warping factor is determined to calculate speaker normalized feature vectors, these feature vectors are then used to perform the speech rate normalization.

As described in section 3.1 the SRN procedure is realized in supervised mode during training and as a second recognition pass applying a word graph rescoring algorithm during recognition. Instead of using unnormalized acoustic models and unnormalized feature vectors, vocal tract length normalized models and feature vectors are applied.

3. EXPERIMENTS AND RESULTS

Recognition experiments were conducted on the speech material of the Verbmobil evaluation 1996. Decision tree based context dependent HMMs (triphones) were estimated on the speech material. Two different baseline systems were examined. Table 2 shows the word error rates (WER) in percent achieved with these baseline systems. The main difference between the two systems is the "acoustic resolution" of the models, which means that model BASELINE1 has a lower resolution using a maximum of 20 mixtures per triphone state, whereas for model BASELINE2 a higher resolution with a maximum of 68 mixtures per triphone state is used. According to the higher resolution of model BASELINE2 the word error rate is considerably lower compared to model baseline1.

	WER
BASELINE1	32.1
BASELINE2	25.7

Table 2: Recognition performance of two different baseline models without any normalization procedure.

In table 3 the results of supervised and unsupervised SRN with the acoustic models BASELINE2 are shown. A relative improvement (column "rel. imp." in table 3) of 4.3% respectively 2.7% can be achieved with the supervised or the unsupervised procedure.

	WER	rel. imp.
supervised SRN	24.6	4.3
unsupervised SRN	25.0	2.7

Table 3: Recognition performance of supervised and unsupervised SRN and relative improvement compared to BASELINE2.

In table 4 the effects of applying the VTLN procedure to models "baseline1" is shown. A relative improvement of 11.2% can be observed. It includes the results of supervised VTLN only, as in preliminary experiments no significant differences were found between supervised and unsupervised VTLN. In agreement with previous experiments ([8]) an increase in relative improvement is observable with higher speaking rates.

	WER	rel. imp.
complete set	28.5	11.2
slow	28.4	8.3
medium	26.6	11.9
fast	32.0	12.6

Table 4: Recognition performance of supervised VTLN and relative improvement compared to baseline 1 (BASELINE1-VTLN). The performance on the complete test set is given as well as the performance for three different speech rate categories.

Table 5 contains the results after applying both VTLN and SRN. Relative improvements are given compared to the use of BASELINE1 models and compared to the application of VTLN alone (BASELINE1-VTLN). Word error rates as well as relative improvements are given for the complete evaluation set (row "complete set") and for each of three speech rate categories. For details on splitting the speech material into different speech rate categories see [10].

	WER	rel. imp. to VTLN	rel. imp. to baseline1
complete set	27.3	4.2	15.0
slow	26.5	7.0	14.6
medium	25.9	2.6	14.2
fast	30.9	3.5	15.7

Table 5: Recognition performance using a combination of VTLN and SRN. The performance on the complete test set is given as well as the performance for three different speech rate categories.

All in all a relative reduction in word error rate of 15% can be achieved using a combination of VTLN and SRN (see last column of row "complete set" of table 5). Higher improvements are achieved for both high and low speaking rates.

4. DISCUSSION

In this contribution a new version of SRN was presented, which is based on adjusting the inter frame distance to the actual speaking rate. The normalization factor is determined for each spurt of an utterance and thus changes in the speaking rate can be captured. Using linear interpolation of the original features and a word graph rescoring procedure leads to a moderate increase in the computational load, compared to a full second recognition pass. An overall reduction of the word error rate of 2.7% (4.3%) was achieved in unsupervised (supervised) mode. However, experiments using an ideal criterion (minimum of WER) for optimizing the normalization factor showed, that a 17% reduction in WER is possible on this task. Thus more investigations in finding a better way of optimizing the normalization factor should be initiated.

In agreement with our monophone-based experiments [8], recent triphone-based investigations show higher improvements in recognition performance for increasing speaking rate.

In additional experiments a combination of ML-based VTLN and SRN was evaluated. Those experiments have shown that, applying SRN to VTLN-based features leads to an overall decrease in error rate of 4.2% which is comparable to the gain observed when applying SRN to unnormalized features. Higher improvements can be achieved for "slow" and "fast" speech in comparison to "medium" speech, for which the lowest error rate reductions occurred.

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] N. Mirghafori, E. Fosler, and N. Morgan, "Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes", 4th European Conference on Speech Communication and Technology, Madrid, Spain, Vol. 1, pp. 491-494, 1995.

[2] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate", 5th European Conference on Speech Communication and Technology, Rhodes, Greece, Vol. 4, pp. 2079-2082, 1997.

[3] M. A. Siegler and R. M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems", IEEE International Conference on Acoustics, Speech and Signal Processing, Detroit, Michigan, Vol. 1, pp. 612-615, 1995.

[4] T. Brøndsted and J. P. Madsen, "Analysis of speaking rate variations in stress-timed languages", 5th European Conference on Speech Communication and Technology, Rhodes, Greece, Vol. 1, pp. 481-484, 1997.

[5] F. Martínez, D. Tapias, and J. Álvarez, "Towards speech rate independence in large vocabulary continuous speech recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle, Washington, Vol. 2, pp. 725-728, 1998.

[6] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on conversational pronunciations", Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Netherlands, Vol. 1, pp. 35-40, 1998.

[7] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection", IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle, Washington, Vol. 2, pp. 945-948, 1998.

[8] T. Pfau, R. Faltlhauser, and G. Ruske, "Speaker normalization and pronunciation variant modeling: helpful methods for improving recognition of fast speech", 6th European Conference on Speech Communication and Technology, Budapest, Hungary, Vol. 1, pp. 299-302, 1999.

[9] N. Mirghafori, E. Fosler, and N. Morgan, "Towards robustness to fast speech in ASR", IEEE International Conference on Acoustics, Speech and Signal Processing, Atlanta, Georgia, Vol. 1, pp. 335-338, 1996.

[10] T. Pfau and G. Ruske, "Creating hidden Markov models for fast speech", 5th International Conference on Spoken Language Processing, Sydney, Australia, Vol. 2, pp. 205-208, 1998.

[11] M. Richardson, M. Hwang, A. Acero, and X. D. Huang, "Improvements on speech recognition for fast talkers", 6th European Conference on Speech Communication and Technology, Budapest, Hungary, Vol. 1, pp. 411-414, 1999.

[12] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle, Washington, Vol. 2, pp. 729-732, 1998.

[13] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization", IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, Georgia, Vol. 1, pp. 346-349, 1996.

[14] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping", IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, Vol. 2, pp. 1039-1042, 1997.

[15] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, and H. Ney, "Recent improvements of the RWTH large vocabulary speech recognition system on spontaneous speech", IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, Vol. 3, pp. 1671-1674, 2000.

[16] J. McDonough, W. Byrne, and X. Luo, "Speaker normalization with all-pass transforms", 5th International Conference on Spoken Language Processing, Sydney, Australia, Vol. 6, pp. 2307-2310, 1998.