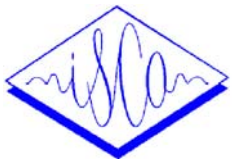

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ON COMBINING VOCAL TRACT LENGTH NORMALISATION AND SPEAKER ADAPATION FOR NOISE ROBUST SPEECH RECOGNITION

Ramalingam Hariharan & Olli Viikki
Nokia Research Center, Speech and Audio Systems Laboratory, Tampere, Finland
Email: {ramalingam.hariharan,olli.viikki}@research.nokia.com

ABSTRACT

This paper investigates the combination of vocal tract length normalisation and speaker adaptation in connected digit recognition. In particular, we focus on performing this task under a continuously varying car noise environment. Continuous supervised speaker and environment adaptation is carried out on the test data according to the Bayesian framework. The paper also evaluates various approaches to implement vocal tract length normalisation. The best performance was obtained when the normalisation was performed during both initial speaker-independent training and testing. It was also noticed that, during testing, speaker specific normalisation produced better results than utterance specific normalisation. Our experimental results on the connected digit database show that the joint approach outperforms the system in which on-line Bayesian speaker adaptation is performed on HMM mean parameters. The performance gain was particularly high with so called outlier speakers for whom adaptation is truly needed.

10.21437/Eurospeech.1999-57

1. INTRODUCTION

Vocal Tract Length Normalisation (VTLN) and speaker adaptation have recently attracted a great deal of interest in speech recognition community. Both methods are typically undertaken to reduce the speaker-specific error rate of practical speech recognition systems. While the classical speaker adaptation methods usually aim at altering the parameters of Continuous Density HMMs (CDHMM) [1,2], VTLN techniques [4-6] attempt to modify the feature space in such a way that the speaker-specific pronunciation details are better captured. To guarantee a fast conversion from a speaker-independent to a speaker-dependent system, both classes of methods are able to cope with a small amount of speech data.

Robustness to ambient background noise is one of the most fundamental requirements of any practical speech recognition system. Despite the intensive research done in the field of noise immune feature extraction, current speech recognition systems rely on the features which are not particularly robust to real-world distortions. Even a small mismatch between the training and test conditions, e.g. a change in a microphone or the presence of low-level background noise, can substantially degrade the recognition performance. Therefore, noise robustness in practical speech recognition systems is usually achieved by collecting data from a test environment and adjusting the CDHMM parameters to better match the test conditions. Environment adaptation is usually based on the same CDHMM adaptation techniques developed for speaker adaptation.

In the present work, we attempt to combine VTLN [4] and Bayesian on-line speaker adaptation [9] in a connected digit recognition task under a continuously varying car noise environment. The objective of VTLN is to minimise large inter-speaker variability during the speaker-independent training process. On-line Bayesian adaptation is first needed to characterise the speaker-specific details in HMMs and secondly to update the CDHMM parameters to better match the present noise conditions, i.e. we perform joint speaker and environment adaptation. Thus, our approach is somewhat similar to speaker adaptive training [10]. Experimental results show the viability of the presented approach. The combined VTLN and Bayesian on-line adaptation was found to provide 14.4% decrease in error rate than that obtained with adaptation alone.

2. VOCAL TRACT LENGTH NORMALISATION

It is very well known [3] that one of the major sources of inter-speaker variation is the vocal tract length of the speaker. The vocal tract length varies from approximately 13 cm for females to 18 cm for males. Since the formant frequency peak positions are inversely proportional to vocal tract length, this variation causes a shift of the formant centre frequencies. This effectively results in a scaling of the frequency axis. VTLN attempts to normalise the speech signal to an average vocal tract length by finding a frequency-scaling factor for each speaker (or utterance), so that the parameterised speech signal is independent of inter-speaker differences.

The normalisation or warping can be implemented in the front end by either re-sampling in time domain [5] or by modifying the width of the mel-spaced filter bank [4]. In this paper, we chose to implement the second approach by modifying the centre frequencies of the mel filter bank.

The two major issues involved in the implementation of VTLN are (1) estimation of the scaling factor, given the speech data and (2) implementation of the normalisation function given the normalisation factor. The method described in [6] estimates the scaling factor based on the measurement of formant frequencies. Another popular approach is to search from a discrete set of possible val-

ues that maximises the likelihood of the warped utterance with respect to the given HMM and the transcription. In this work, we use the latter approach to implement VTLN. The normalisation is typically performed only for the speech portion of the utterance.

Let θ denote the parameters of the given HMM model, W the decoded model transcription from an initial recognition pass and $X(\alpha) = f_\alpha(X)$ corresponds to the cepstral domain observation vectors warped according to the function f_α. The optimal warping factor can then be determined as

$$\hat{\alpha} = \arg\max_{\alpha} P\left(X(\alpha)\middle|\alpha, \theta, W\right). \quad (1)$$

The normalisation can be applied during both the training and the recognition phases. During training, a single warping factor α_i is determined for each speaker i. The normalised speech feature vectors are then used to build a set of HMMs. The resulting normalised models, which are independent of inter-speaker variations, are usually faster to be adapted to a particular speaker. In the recognition phase, an initial recognition pass with the unwarped observation vectors is done to obtain the model transcription. The warping factor that gives the maximum likelihood is then used in the second recognition pass to get the final recognition output.

3. CDHMM PARAMETER ADAPTATION

To date, HMM adaptation techniques can broadly be categorised into two different groups:

- Local adaptation methods, e.g. Bayesian approach [2], where parameter statistics are estimated at the mixture-density level and each Gaussian mixture is updated independently.
- Model-transformation based adaptation methods, e.g. MLLR [1], where one attempts to find a global transform, or a set of transforms, which is applied to all Gaussian mixture densities (or a group of mixture densities).

Both approaches are usually based on the Maximum Likelihood (ML) criterion. Furthermore, adaptation is typically restricted only to Gaussian mean vectors. In this paper, we favour the Bayesian adaptation approach as it has proved its superiority with respect to MLLR in noise robust small vocabulary speech recognition systems [12]. Let $\mathbf{m}_{jk}$ be the mean vector of the kth Gaussian mixture in the jth state of an HMM, and $\mathbf{o}_t$ is a feature vector at time t. Based on the Bayesian principle, the new mean estimate for the Gaussian mean vector can now be expressed in the form

$$\hat{\mathbf{m}}_{jk} = \frac{\tau \cdot \mathbf{m}_{jk} + \sum_{t=1}^{T} d_{jk}(t)\mathbf{o}_t}{\tau + \sum_{t=1}^{T} d_{jk}(t)} \quad (2)$$

where $\mathbf{m}_{jk}$ is the original mean vector, $\hat{\mathbf{m}}_{jk}$ is its updated version, τ is adaptation learning rate, and $d_{jk}(t)$ denotes the probability of observing mixture component k in state j at time t. This probability is defined as

$$d_{jk}(t) = \gamma_j(t)\frac{c_{jk}\mathrm{N}\left(\mathbf{o}_t\middle|\mathbf{m}_{jk}, \mathbf{U}_{jk}\right)}{\sum_{l=1}^{K} c_{jl}\mathrm{N}\left(\mathbf{o}_t\middle|\mathbf{m}_{jk}, \mathbf{U}_{jk}\right)}, \quad (3)$$

where $\mathbf{U}$ denotes the diagonal covariance matrix, c is the mixture weight and $\gamma_j(t)$ is the state occupation probability in state j at time t.

4. COMBINING VTL NORMALISATION AND HMM ADAPTATION

One main objective of this work was to investigate how vocal tract length normalisation and HMM adaptation should be combined in real-world speech recognition systems. The attainable recognition rate and implementability are the two key factors that has been considered in this study. In [9], we observed that supervised[1] on-line adaptation is clearly superior to off-line adaptation in continuously varying noise conditions. Supervised on-line Bayesian adaptation was therefore chosen to be done during recognition. VTLN can be implemented in the recognition system in different ways. We investigated the performance of the following modes of applying VTLN with Bayesian adaptation:

A. VTLN applied only during the training phase. This arrangement ensures that there is no extra run-time computational burden during recognition.
B. Speaker specific VTLN during training (same warping factor for each speaker) and utterance-wise VTLN in testing. VTLN during testing ensures that the normalised feature vectors match well with the lower variance normalised Gaussian mixture densities.
C. Speaker specific VTLN both during training and testing phases. This approach reduces the computational requirements with respect to B which is essential if VTLN is applied in practical systems.

5. EXPERIMENTAL EVALUATION

5.1 Test Database and Settings

The database used for the evaluation was a Finnish connected digit database recorded in a car environment. This database consisted of 9 speakers (4 males and 5 females) with each speaker speaking at least 1,000 utterances. Recordings were carried out during four recording days over a time-span of one month. Each recording session took approximately one hour during

[1] In this and our previous work, supervised adaptation was realised so that parameter statistics accumulation was only carried out with correctly recognised utterances.

which a test speaker spoke connected digit triples in continuously changing noise conditions depending on the speed of the car, road, and traffic conditions. In the off-line recognition experiments, the order of the test utterances was further randomised so that consecutive utterances were *not* necessarily spoken in similar noise conditions. This arrangement enabled us to simulate a practical usage pattern. In addition to these test utterances, each speaker also uttered 30 connected digit triples in a noise-free car environment. These utterances were used in the VTLN experiments to find the best warping factor for each speaker in the speaker specific VTLN experiments.

A feature vector set consisting of 39 coefficients - 12 FFT based MFCCs, log-energy, and their first and second-order time derivatives - were used for all the experiments in this paper. These feature vectors were further normalised to have similar parameter statistics in all noise conditions as described in [8].

5.2 Speaker-Independent Training

A separate Finnish language connected digit database was used for training the initial and vocal tract length normalised whole-word speaker-independent HMMs. Training utterances of around 375 speakers, spoken in a car environment, under various noise conditions, were used for the purpose. For multi-environment type of speaker-independent training, all utterances were pooled together and a set of state duration constrained [11] CDHMMs (2 Gaussian mixtures per state) were estimated according to the ML criterion. The number of male and female speakers, the number of digits, and the transitions between the digits were balanced in the training database.

The estimation of initial VTLN SI HMMs converged after two training iterations. Hence, all the VTLN results presented here are based on two ML training iterations.

5.3 VTLN Experiments

Following the observations made in [4], a discrete set of warping factors from 0.88 to 1.12 in steps of 0.02 were used in our experiments to find the best warping factor. It should be noted that the bandwidth of the warped signal stretches beyond the 4 kHz for some warping factors and this information is not used. However our experiments with piece-wise linear warping of the mel filter banks (which tries to include all the information from 0-4 kHz) produced similar recognition results as that obtained with simple linear warping, thereby verifying the theory [4] that the higher frequency regions contain little or no useful information.

The results of applying VTLN without on-line HMM adaptation are summarised in Table 1. The baseline results (digit string recognition accuracy) with speaker *unnormalised* SI HMMs are given in the first row. The results of applying VTLN during training and testing separately are given in the next two rows. The final row presents the results when VTLN is applied during both training and testing. As shown in Table 1, there is an improvement in the recognition accuracy (error rate decrease of more than 7%) over the baseline when VTLN is applied during testing alone. If the normalisation was applied during both training and testing (utterance wise warping), over 21% string error rate decrease was obtained. But, interestingly enough, applying normalisation during training alone decreased the recognition rate. This could be due to the fact that the Gaussian mixture densities obtained by VTLN are sharper with lower variance values. These models are not able to handle the relatively larger variations in the speaker un-normalised feature vectors of the test utterances. On the other hand, when VTLN is applied only during the testing phase, the coarser models trained from un-warped utterances can still be used to recognise the warped utterances. The best results, not surprisingly, are obtained when VTLN is applied both during training and testing.

The recognition rates are also presented separately for males and females. The tests produced relatively higher recognition accuracy for females than for males, when VTLN is used both for training and testing.

Table 1: Results using VTLN.

VTLN Train	VTLN Test	Digit String Recognition Accuracy (in %)			Error Rate Decrease (%)
		Male	Female	Aver.	
No	No	88.41	71.06	77.51	-
Yes	No	85.92	64.15	72.15	-22.42
No	Yes	90.18	72.56	79.14	7.25
Yes	Yes	89.68	77.91	82.29	21.25

5.4 Combining VTLN & HMM Adaptation

Next, we combined VTLN and Bayesian adaptation aiming at optimising the speaker-specific recognition rate. The results obtained in these experiments are given in Table 2. The error rates in the last column are computed with respect to the average value in the second row.

Table 2: Results using a combination of on-line HMM adaptation and VTLN.

VTLN Train	VTLN Test	HMM Adapt	Digit String Recognition Accuracy. (in %)			Error Rate Decrease (%)
			Male	Female	Aver.	
No	No	No	88.41	71.06	77.51	-
No	No	Yes	98.97	93.60	95.61	-
Yes	No	Yes	99.06	93.72	95.71	2.28
Yes	Yes	Yes	98.92	94.02	95.85	5.47
Yes	Yes[2]	Yes	99.07	94.55	96.24	14.35

[2] Speaker specific warping factor in testing.

The baseline results with SI HMMs are given in the first row. In the second row, results are provided when Bayesian on-line adaptation is applied to the utterances using initial VTLN models. There is a moderate decrease in word error rate (2.3%) with respect to the baseline adaptation results, when the speaker normalised models were used for adaptation. A much higher decrease of 5.5% was observed when we applied VTLN during both training and testing. Since the performance improvements were still quite marginal, an additional experiment was done, where the warping factors were found for each speaker from a separate set of utterances. The warping factor of each speaker was computed as the median value of the utterance-wise warping factors obtained from the utterances of that speaker. This estimated value was then fixed and it was further used to warp all the test utterances spoken by this speaker. This kind of speaker specific warping together with adaptation produced the highest decrease in error rate of 14.4% over plain HMM adaptation alone. These results show that, even though the gains are not essentially additive (the baseline adaptation results are already very high), there is a substantial gain combining both VTLN and adaptation. In particular, the performance improvement was higher with outlier speakers. For one of these outlier speakers who had a SI baseline recognition rate of 29.4%, the rate increased to 81.3% with pure on-line adaptation. Combined VTLN and adaptation increased the recognition rate to 85.6%.

It is essential from the implementation point of view that the on-line computational demands of any new algorithm are as small as possible. The main computation involved in VTLN is the estimation of the warping factor. To search for the optimal warping factor, we needed to test 13 different values for α. In practice, it is obvious that this extensive search for each test utterance is not always feasible. In speaker specific VTLN, the warping factors can be computed off-line using a small number of initial utterances of the speaker. Once this is determined, there are no further extra computational requirements for VTLN and is hence very suited for practical systems with strict real-time requirements.

6. CONCLUSIONS

We investigated the performance of combining vocal tract length normalisation and speaker adaptation in this paper. In the proposed approach, vocal tract length normalisation was first applied to the input speech signal, and next, CDHMM mean vectors were continuously updated to better match the new speaker's pronunciation characteristics and current noise conditions. The best performance was obtained for the implementation where vocal tract length normalisation was done during both during training and testing. Moreover, speaker specific normalisation produced better results compared to computationally expensive utterance specific normalisation for testing. Recognition results show that VTLN should always be combined with some other compensation approach since the performance gains provided by VTLN alone are quite marginal. Experimental results verify the efficiency of the proposed combined approach. In particular, the proposed approach improved the recognition accuracy for such speakers who obtained a poor recognition accuracy when using the baseline system.

REFERENCES

[1] Leggetter C.J. & Woodland P.C., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech & Language*, Vol. 9, pp. 171-185.

[2] Gauvain, J. L., Lee, C.-H. "Maximum a Posteriori Estimation of Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, April 1994.

[3] Fant. G, *Speech Sounds and Features*, Cambridge, MA: M.I.T. Press, 1973.

[4] Lee L. & Rose R.C., "Speaker Normalisation Using Efficient Frequency Warping Procedures", *Proc. ICASSP'96*, Vol. 1, pp. 353-356, Atlanta.

[5] Andreou A., Kamm T. & Cohen J., "Experiments in Vocal Tract Normalisation", *Proc. CAIP Workshop : Frontiers in Speech Recognition II*,1994.

[6] Eide E. & Gish H., "A Parametric Approach to Vocal Tract Length Normalisation", *Proc. ICASSP'96*, Vol. 1, pp. 346-348, Atlanta.

[7] Pye, D., Woodland P.C., "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition", *Proc. ICASSP'97*, Vol. 2, pp: 1047 –1050, Munich.

[8] Viikki, O., Bye, D., Laurila, K. "A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise", *Proc. ICASSP'98* , pp. 733-736, Seattle, USA.

[9] Viikki, O., Laurila, K., "Incremental On-line Speaker Adaptation in Adverse Conditions", *Proc. ICSLP98*, pp. 1779-1782, Sydney, Australia.

[10] Anastasakos, T., McDonough J., Schwartz R. & Makhoul J, "A Compact Model for Speaker Adaptive Training", *Proc. ICSLP'96*, pp. 1137-1140, Philadelphia.

[11] Laurila, K. "Noise Robust Speech Recognition with State Duration Constraints", *Proc. ICASSP'97* , pp. 871-874, Munich, 1997.

[12] Laurila, K., Vasilache, M., Viikki, O. "A Combination of Discriminative and Maximum Likelihood Techniques for Noise Robust Speech Recognition", *Proc. ICASSP'98*, pp. 85-88, Seattle, USA, May, 1998.