SPEAKER NORMALIZATION USING EFFICIENT FREQUENCY WARPING PROCEDURES

Li Lee [†] *and Richard C. Rose*

AT&T Bell Labs, Murray Hill, NJ 07974, U.S.A.

ABSTRACT

In an effort to reduce the degradation in speech recognition performance caused by variation in vocal tract shape among speakers, a frequency warping approach to speaker normalization is investigated. A set of low complexity, maximum likelihood based frequency warping procedures have been applied to speaker normalization for a telephone based connected digit recognition task. This paper presents an efficient means for estimating a linear frequency warping factor and a simple mechanism for implementing frequency warping by modifying the filter–bank in mel–frequency cepstrum feature analysis. An experimental study comparing these techniques to other well–known techniques for reducing variability is described. The results showed that frequency warping was consistently able to reduce word error rate by 20% even for very short utterances.

1. INTRODUCTION

One major source of interspeaker variability in hidden Markov model(HMM) based continuous speech recognition is the variation of vocal tract shape among speakers in a population. The positions of spectral formant peaks for utterances of a given sound are inversely proportional to the length of the vocal tract. Since the vocal tract length can vary from approximately 13cm for females to over 18cm for males, formant center frequencies can vary by as much as 25% between speakers. This source of variability results in a significant degradation from speaker dependent to speaker independent speech recognition performance.

Andreou, *et.al.*, proposed a set of maximum-likelihood speaker normalization procedures to explicitly compensate for variations in vocal tract length [1]. These procedures reduced speaker-dependent variations between formant frequencies through a simple linear warping of the frequency axis. While this and other studies of frequency warping procedures have shown improved speaker-independent ASR performance, the performance improvements were achieved at the cost of highly computationally intensive procedures [2] [3].

This paper extends the work of Andreou, *et.al.*, by presenting an experimental study of some properties of the speaker normalization procedures, and proposing more efficient methods for implementing frequency warping and for incorporating the speaker normalization process into HMM recognition. A mixture-based technique for directly estimating the linear frequency warping factor is described and evaluated. A method of implementing frequency warping by directly modifying the shapes of the component filters in the filter-bank front-end is used. The procedures were evaluated on a telephone-based connected digit recognition task in which the utterances range in length from only a single digit to seven digits long.

This paper consists of three major parts. First, we describe the HMM-based speaker normalization training and recognition procedures which were proposed by Andreou, *et.al.*. More efficient extensions to these procedures, including mixture-based warping factor estimation, are proposed. The second section describes the front-end signal processing, and presents a method of directly incorporating frequency warping in the filter-bank front-end. The third section presents results from an experimental study of several properties of the speaker normalization procedures. An experimental study was performed to compare the effects of speaker normalization to more well-known compensation techniques and to investigate the importance of a number of experimental variables.

2. NORMALIZATION PROCEDURES

This section reviews the HMM-based speaker normalization procedures developed in [1], and presents more efficient alternative procedures which can be more easily applied to speech recognition systems that must be implemented in "real–time". In speaker normalization, the standard procedures are preceded by estimation of an optimal warping factor, $\hat{\alpha}$, during both HMM training and recognition. During HMM training, one $\hat{\alpha}$ is first estimated for each speaker in the training set, and then all of the warped utterances are used to build a "normalized" HMM. Similarly, during recognition, we first estimate an $\hat{\alpha}$ based on the input utterance, and then decode the utterance using the warped feature vectors.

During both testing and training, $\hat{\alpha}$ is estimated by maximizing the likelihood of the utterance with respect to a given Hidden Markov Model. Suppose that λ denotes a set of HMM models, X_i^α denotes the cepstrum domain observation vectors for a set of utterances from speaker i warped by α, and W_i denotes the corresponding transcriptions for the utterances. Then, the optimal warping factor for speaker i, $\hat{\alpha}_i$, is defined as follows:

$$\hat{\alpha}_i = \arg\max_{\alpha} \Pr(X_i^\alpha | \lambda, W_i). \qquad (1)$$

Since it is very difficult to obtain a closed–form solution for $\hat{\alpha}$ from Equation 1, a grid search over a set of 13 factors is used. We choose a range of 0.88 to 1.12 for the possible $\hat{\alpha}$ to roughly reflect the 25% range in vocal tract lengths found in humans.

Using a criterion like the one in Equation 1 to estimate a warping function is extremely important. Estimating a warping function that provides a better match to the HMM model, instead of trying to solve the difficult problem of obtaining an estimate of the "true" vocal tract shape for a particular speaker, is much more likely to have an impact on speech recognition performance. While there have been many examples of more interesting frequency warping transformations applied to speaker nor-

[†] Li Lee is currently with Dept. of EECS, Massachusetts Institute of Technology, Cambridge, MA

0-7803-3192-3/96 $5.00©1996 IEEE

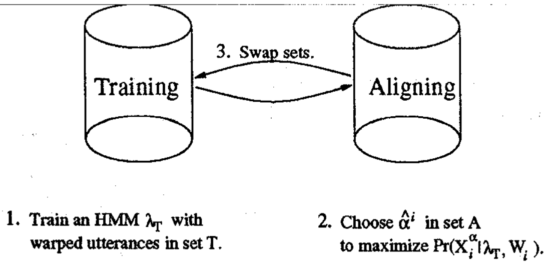
3. Swap sets.
Training
Aligning
1. Train an HMM λ_T with warped utterances in set T.
2. Choose $\hat{\alpha}^i$ in set A to maximize $\Pr(X_i^\alpha|\lambda_T, W_i)$.

Figure 1: HMM training with speaker normalization

malization in speech recognition, none have used an optimization criterion which is consistent with that used in the recognizer to estimate the parameters of the transformation.

2.1. Training Procedure

The goal of the training procedure is to warp the frequency scale of the utterances for each speaker in the training set so that the resulting speaker independent HMM will be defined over a frequency normalized feature set. An iterative procedure is used to alternately choose the best warping factor for each speaker, and then build a model using the warped training utterances. A diagram of the procedure is shown in Figure 1. After dividing the training speakers into two sets, training(T) and aligning(A), we first train an HMM, λ_T, using the utterances in set T. Then, the optimal warping factor for each speaker i in set A is chosen to maximize $\Pr(X_i^\alpha|\lambda_T, W_i)$. All of the utterances from the same speaker are used to estimate $\hat{\alpha}$ for that speaker. We then swap the sets, and iterate this process of training an HMM with half of the data, and then finding the best warping factor for the second half. A final frequency normalized model, λ_N, is built with all of the frequency warped utterances when there is no significant change in the estimated $\hat{\alpha}$'s between iterations.

2.2. Testing Procedure

During recognition, the goal is to warp the frequency scale of the test utterance to "match" that of the normalized HMM model λ_N. Unlike the training scenario, however, we have only one testing utterance with which to estimate $\hat{\alpha}$, and the transcription is not given. A three-step process is used:

1. Decode the unwarped utterance x using the λ_N model. Denote the transcription as w^u.
2. Set $\hat{\alpha} = \arg\max_\alpha \Pr(x^\alpha|\lambda_N, w^u)$, where x^α is the frequency warped utterance. This probability is evaluated by probabilistic alignment of each warped set of feature vectors with the transcription w^u.
3. Decode the utterance $x^{\hat{\alpha}}$ with the model λ_N.

2.3. More Efficient Testing Procedures

The testing procedure is computationally intensive because the inefficiency of performing an exhaustive grid search is compounded by the need to use probabilistic alignment at each possible α. As a simplification, we have tried to use more coarsely sampled search grids during recognition. We have also devised a mixture-based method to choose $\hat{\alpha}$ during recognition. This method is as shown in Figure 2.

During training, after warping factors have been determined for all of the speakers using the process shown in Figure 1, mixtures of multivariate Gaussians are trained to represent the feature space distributions of each of the possible classes. That is, for each warping factor, mixtures are trained using the *unwarped* feature vectors from utterances which were assigned to that warping factor. Then, during recognition, the probability of the incoming utterance before frequency warping is evaluated against each of these distributions, and the warping factor $\hat{\alpha}$ is chosen for the distribution which yields the highest likelihood over the entire utterance. The speech is warped using $\hat{\alpha}$ and the resulting feature vectors are then used for HMM decoding.

Using this method, there is no need to obtain a preliminary decoding using the unwarped utterances, nor is there a need to perform probabilistic alignment at all of the grid points. However, unlike the method described in section 2.2, this mixture-based method does not take advantage of the temporal information in the signal.

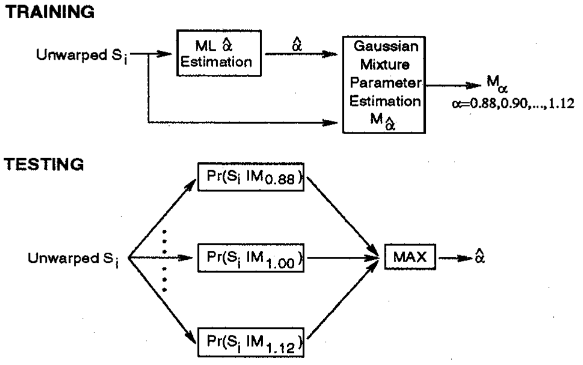

Figure 2: Mixture-based warping factor estimation

3. FRONT-END SIGNAL PROCESSING

Frequency warping can be incorporated directly into the mel-spaced filter-bank front-end by varying the spacing and width of the component filters and keeping the speech spectrum unchanged. This is illustrated by the block diagram in Figure 3. Instead of resampling the sampled speech waveform, we can push the warping process into the filter-bank stage. For example, to compress the speech signal in the frequency domain, we can keep the frequency scale of the speech signal the same, but stretch the frequency scale of the filters. Similarly, we can compress the filter-bank frequencies to effectively stretch the signal frequency scale.

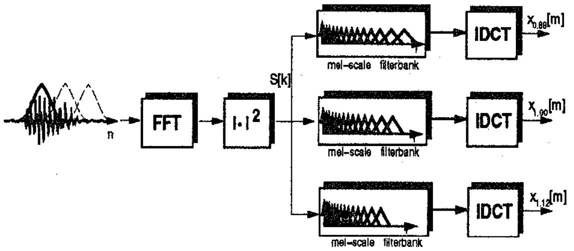

Figure 3: Frequency warping in the filter-bank domain

When the frequency axis is warped, the bandwidth of the resulting signal differs from that of the original. For the experiments described in this work, the original sampling rate band-limits the unwarped signal to 4kHz bandwidth. Consequently, with the warping factors ranging between 0.88 and 1.12, the bandwidths of the warped signals range between 3.52 kHz and 4.48 kHz. Because comparisons for the "best" warping factor are made over a constant range between 0 and 4kHz, the compressed signals do not contain useful information over the entire 4 kHz, and the stretched signals contain information above 4 kHz that are not used. Different bandwidths at different warping factors represent a source of mismatch between the warped signal and the model. In the mel-spaced filterbank front-end, this problem is mitigated somewhat by the fact that filters near the band-edge have bandwidths approaching 800 Hz. Thes wider filters serve to blur the exact location of the band-edge.

4. EXPERIMENTAL STUDY

This section describes a set of experiments using the speaker normalization techniques described earlier. We first describe the recognition task and database. Next, recognition results comparing speaker normalization to gender-dependent models and cepstral mean normalization are shown. We then present experiments using the more efficient recognition procedures described earlier. The next subsection presents results on the effect of HMM parametrization on speaker normalization procedures. Finally, we show data on how the number of iterations taken during training changes the recognition performance.

4.1. Task and Recognition System

The experiments were performed on a telephone-based connected digit recognition task. The database was recorded in shopping malls across 15 dialectally distinct regions in the US, using two carbon and two electret handsets. The speakers read digit strings between 1 and 7 digits over a telephone. The training set contained 713 speakers, with 8802 utterances totaling 26717 digits. The testing set contained 596 speakers, with 4304 utterances totaling 13185 digits. The training utterances were endpointed, whereas the testing utterances were not. We use word error rate as the performance metric.

Continuous-density left-to-right HMMs with 8-10 states were used to model each digit in the recognizer. In addition, silence was explicitly modeled by a single-state HMM. The observation densities were mixtures of 8 multi-variate Gaussian distributions with diagonal covariance matrices. 39-dimensional feature vectors were used: normalized energy, c[1]–c[12] derived from a mel-spaced filter-bank of 22 filters, and their first and second derivatives.

4.2. Baseline Recognition Results

Table 1 compares the recognition performance results using the baseline system, speaker normalization, gender-dependent models, and cepstral mean subtraction. The first row reports the word error rate observed when testing unwarped feature vectors using models trained on unwarped feature vectors. The second row reports the error rate observed using speaker normalization. The models were trained using frequency-normalized feature vectors obtained after the first iteration of the iterative HMM training procedure. In implementing cepstral mean subtraction, the mean of the cepstral vectors over the non-silence portions of each utterance is computed and subtracted from the entire utterance; the results are shown in the third row. Finally, gender-dependent models were implemented using 2 sets of HMMs with 8 Gaussians per mixture were trained, one for the female speakers and one for the male speakers, and the results are shown in the fourth row. The error rates for utterances spoken through the carbon and electret handsets are shown separately in the second and third columns, and averaged in the last column.

Condition	Carbon	Electret	All
Baseline	2.8 %	4.1 %	3.4 %
Speaker Normalization	2.4 %	3.1 %	2.7 %
Cepstral Mean Norm.	2.5 %	3.7 %	3.1 %
Gender-Dep. Models	2.3 %	3.4 %	2.9 %

Table 1: Performance of speaker normalization procedures as compared to using no warping, to using gender-dependent models, and to cepstral mean normalization.

There are several observations that can be made from Table 1. First, it is clear from the table that the overall word error rate is reduced by approximately 20% through the use of frequency warping during both HMM training and recognition. Speaker normalization performed better than both cepstral mean normalization and gender-dependent models, even though the GD models used twice as many model parameters as the others. The second observation concerns the relative error rate obtained using carbon and electret transducers. For both conditions, the error rate for the carbon transducers is significantly lower than that for the electret. These results are consistent with those observed in [4], and a possible explanation for the performance discrepancy was provided there. Finally, it is interesting to note that this performance difference between carbon and electret transducers is reduced after speaker normalization.

4.3. Distribution of Warping Factors

We also examined the distribution of the chosen warping factors over the speaker set to verify our intuition about distortions caused by vocal tract length variations, and the distribution of vocal tract length over the human population. Histograms of the chosen warping factors for the speakers in the training set are shown in Figure 4. On average, about 15 utterances are used to estimate the warping factor for each speaker. The value of the estimated warping factor is displayed along the horizontal axis, and the number of speakers who were assigned to each given warping factor is plotted on the vertical axis. Warping factors below 1.00 correspond to frequency compression, and those above 1.00 correspond to frequency expansion. The mean of warping factors is 1.00 for males, 0.94 for females, and 0.975 for all of the speakers.

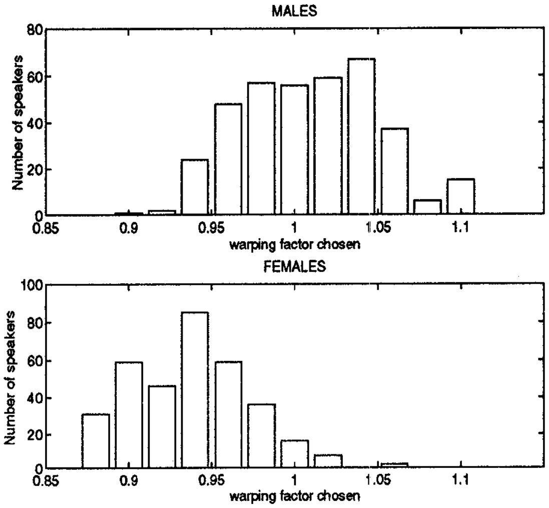

Figure 4: Histogram of warping factors chosen for speakers in the training set

Clearly, the average warping factor among males is higher than that among females. This satisfies our intuition because females tend to have shorter vocal tract lengths, and higher formant frequencies. As a result, it is reasonable that the normalization procedure chooses to compress the frequency axis more often for female speech than for male speech. At the same time, however, the fact that the mean of the estimated warping factors over all speakers is not 1.00 is somewhat surprising, because the iterative training process was initiated with a model built with unwarped utterances. One explanation for this result lies in the difference in the effective bandwidth between utterances whose frequency axes have been compressed or expanded to different degrees. One side-effect of frequency compression is the inclusion of portions of the frequency spectrum which may have originally been out-of-band. If parts of the discarded spectra carry information useful for recognition, the ML warping factor estimation is likely to be biased toward frequency compression.

4.4. Efficiency of Recognition Procedure

Experiments which examined the necessity of a large finely-sampled search grid during recognition and the importance of time-alignment in likelihood computation were also performed. These results are as shown in Table 2. The baseline result is taken from the no warping case. HMM-based search method refers to using probabilistic alignment at each possible warping factor during recognition. The mixture-based search method refers to the more efficient search method we proposed earlier and shown in Figure 2. The error rate is computed over all of the test data.

Search method	# Search pts.	Error Rate
Baseline(No warping)	0	3.4%
HMM-based	13	2.7%
HMM-based	7	2.8%
HMM-based	5	2.8%
HMM-based	3	2.9%
Mixture-based	13	2.9%

Table 2: Performance of more efficient speaker normalization recognition procedures

A comparison among rows 2–5 in Table 2 shows that using a successively smaller number of possible warping factors results in a graceful degradation in performance. The recognition error rate increased by only about 7.5% when the number of warping factors decreased from 13 to 3. Compared with the baseline system with no frequency warping, allowing only 3 possible warping factors still offers a 15% reduction in error rate.

Comparing the second and last rows of Table 2, we see that using the mixture-based search method also results in about a 7.5% increase in error rate. This suggests that the temporal information in the speech signal is indeed useful for determining the warping factor. Despite the slightly higher error rate, however, the computational complexity of the warping factor estimation stage during recognition is significantly reduced using the mixture-based method.

4.5. HMM Parametrization

We also investigated how the complexity of the HMMs used affects the amount of performance gain achieved by speaker normalization. The results of the experiment are shown in Table 3. The rows of the table show the recognition results as the number of Gaussians used in each observation density mixture is increased.

Size of Mix.	Baseline	Warping	% Improvement
8	3.4 %	2.7 %	20 %
16	3.2 %	2.4 %	25 %
24	2.5 %	2.0 %	20 %
32	2.6 %	–	–

Table 3: Performance of speaker normalization over different complexity HMMs

It is especially notable from Table 3 that in every case, using frequency warping with a simpler HMM performs better than using no warping with more complex HMMs.

4.6. Effect of Iterative Training Procedure

Finally, we performed experiments to see how the number of iterations change the recognition rate on both the training and the testing data. These experiments verified that the frequency warping procedures do indeed reduce the speaker variability on the training set, and that the "normalized" HMMs become more efficient over the iterations. Further, they allow us to examine whether multiple iterations result in improved recognition performance on the test set. The results are as shown in Table 4. While multiple training iterations improved the recognition performance on the training data dramatically, it did not help recognition during testing. One explanation for this behavior is overfitting to the training data. Another explanation is a "drift" in the average warping factor estimates that can occur in training. It is interesting that using the speaker normalization procedure during recognition with an unnormalized HMM (first row of the table) still offers a significant improvement over the baseline.

No. of Iter.	Train Set	Test Set
0	2.4 %	2.9 %
1	1.7 %	2.7 %
2	1.3 %	2.9 %
3	1.3 %	2.9 %

Table 4: Recognition performance of HMM model on training and testing data after different number of training iterations

5. CONCLUSIONS

This paper has presented an experimental study of the effectiveness of speaker normalization procedures on a telephone-based digit recognition task with very short utterances. The results show that these procedures reduce the digit recognition error rate by approximately 20 %. Additionally, more efficient extensions to these procedures can reduce their computational cost with little degradation in performance.

6. REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen, "Experiments in Vocal Tract Normalization," *Proc. the CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[2] R. Roth, et. al., "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer" *Proc. of the Spoken Language Systems Technology Workshop*, 1995.

[3] L. Lee, A. Potamianos, and R.C. Rose, "Efficient Frequency Warping Procedures for Telephone Based Speech Recognition," *Speech Research Symposium*, 1995.

[4] A. Potamianos, L. Lee, and R.C. Rose, "A Feature-Space Transformation for Telephone Based Speech Recognition," *Eurospeech*, 1995.