A Frequency Warping Approach to Speaker Normalization

Li Lee, *Student Member, IEEE,* and Richard Rose, *Member, IEEE*

***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 filterbank in mel-frequency cepstrum feature analysis. An experimental study comparing these techniques to other well-known techniques for reducing variability is described. The results have shown that frequency warping is consistently able to reduce word error rate by 20% even for very short utterances.**

***Index Terms*—Continuous speech recognition, frequency warping, hidden Markov modeling, speaker normalization.**

I. Introduction

ONE MAJOR source of interspeaker variability in hidden Markov model-based (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 13 cm for adult females to over 18 cm for adult 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. The contribution of this paper is to describe a set of frequency warping based speaker normalization techniques that are applied in a single utterance based speech recognition paradigm. In these procedures, the parameters of the frequency transformation that is applied to the utterance are estimated using only the samples from the utterance that is input to the speech recognizer.

Techniques that attempted to "normalize" parametric representations of the speech signal for the purpose of reducing the effects of interspeaker differences have been investigated in the context of vowel identification [3], [7], [13]. Normalization was performed using linear and nonlinear frequency warping functions to compensate for variations in formant positions among speakers. These procedures attempted to solve the difficult problem of estimating the formant positions that correspond to the "true" vocal tract shape of each speaker, and then compensating for these differences.

Recently, Andreou *et al.* proposed a set of maximum likelihood based speaker normalization procedures to extract and use acoustic features that are robust to variations in vocal tract length [1]. The procedures reduced speaker dependent variations between formant frequencies through a simple linear warping of the frequency axis, which was implemented by resampling the speech waveform in the time domain. However, despite the simple form of the transformation being considered, over five minutes of speech was used to estimate the warping factor for each speaker in their study. While this and other studies of frequency warping procedures have shown improved speaker independent automatic speech recognition (ASR) performance, the performance improvements were achieved at the cost of highly computationally intensive procedures [11]. The work presented here represents an extension to that performed by Andreou, *et al.* in that several techniques are proposed for making frequency warping methods efficient, and an experimental study is performed to characterize the behavior of these techniques on a telephone based speech recognition task.

Unlike other speaker normalization procedures, the techniques described in this paper make no attempt at uncovering information relating to the underlying vocal tract shape. Instead, the optimization criterion used to estimate the parameters of the frequency warping transformation is directly related to the degree of mismatch between the input utterance and the speech recognition models. It was thought that the most reasonable means for estimating parameters in any speaker normalization procedure should involve an optimization criterion that is consistent with that used in the speech recognizer. These techniques are applied in the context of HMM-based continuous speech recognition over the public switched telephone network.

The effectiveness and efficiency of these procedures are studied from several different perspectives. In addition to speech recognition performance, experiments are performed to evaluate the convergence properties of the proposed procedures. Methods for improving the efficiency of performing model-based speaker normalization and implementing frequency warping are proposed and evaluated. Finally, comparisons of speaker normalization with other techniques to reduce interspeaker variations are made in order to gain insight into

Manuscript received July 27, 1996; revised February 28, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Picone.

L. Lee is with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: llee@mit.edu).

R. Rose is with AT&T Research, Florham Park, NJ 07932 USA.

Publisher Item Identifier S 1063-6676(98)00096-0.

how to most efficiently improve the speaker robustness of ASR systems. The goal of such a study is to better understand the basic properties of speaker normalization so that the technique can become practical for use in existing applications.

The paper is organized as follows. The next section presents a detailed description of procedures for performing speaker normalization in HMM based speech recognition. Procedures which implement frequency warping, warping factor estimation, model training, and recognition are described. A simple mixture based method that estimates the warping function more efficiently during recognition is also presented. Section III presents an experimental study of the effectiveness of the speaker normalization procedure for a telephone based connected digit recognition task. The data base, the task, and the baseline speech recognition system are described. The effectiveness of speaker normalization is examined from several perspectives, including recognition performance and convergence issues. In Section IV, the speaker normalization procedure is compared with other procedures designed to reduce the effects of speaker and channel variability, including gender-dependent modeling and cepstral mean normalization. Discussion and summary are provided in Section V.

II. A Frequency Warping Approach to Speaker Normalization

This section presents detailed descriptions of the procedures used to implement a frequency warping approach to speaker normalization. These procedures attempt to reduce the interspeaker variation of speech sounds by compensating for variations in vocal tract length among speakers. Because distortions caused by vocal tract length differences can be modeled by a simple linear warping in the frequency domain of the speech signal, the normalization procedure scales the signal frequency axis by an appropriately estimated warping factor.

It should also be noted that frequency warping is performed in the context of speaker independent ASR, where speaker independent HMM's are trained using utterances from a large population of speakers. The application of frequency warping to HMM training is investigated so that a speaker independent HMM can be produced that is defined over a frequency-normalized feature set. Frequency warping is applied during recognition in order to reduce the mismatch between the test utterance and the frequency-normalized HMM model.

This section is divided into four parts. First, the warping factor estimation process is presented in Section II-A. Second, Section II-B describes the iterative procedure used to train HMM's using normalized feature vectors from the training data. Section II-C describes procedures for warping factor estimation and frequency warping during HMM speech recognition. The first warping factor estimation procedure involves two recognition passes over the input utterance. The second more efficient procedure treats frequency warping as a set of "class-dependent" transformations. Finally, the implementation of frequency warping as part of the filterbank feature extraction front-end is described in Section II-D.

A. Warping Factor Estimation

Conceptually, the warping factor represents the ratio between a speaker's vocal tract length and some notion of a reference vocal tract length. However, reliably estimating vocal tract length of speakers based on the acoustic data is a difficult problem. In the work described here, the warping factor is chosen to maximize the likelihood of the normalized feature set with respect to a given statistical model, so that the "reference" is taken implicitly from the model parameters. Even though lip movements and other variations change the length of the vocal tract of the speaker according to the sound being produced, it is assumed that these types of variations are similar across speakers, and do not significantly affect the estimated warping factor. Therefore, one warping factor is estimated for each person using all of the available utterances. Evidence supporting the validity of this assumption will be presented in Section III.

The warping factor estimation process is described mathematically as follows. The basic notation is defined here. In the short-time analysis of utterance j from speaker i, the samples in the tth speech frame, obtained by applying an M-point tapered Hamming window to the sampled speech waveform, are denoted with $s_{i,j,t}[m]$, $m = 1 \cdots M$. The discrete-time Fourier transform of $s_{i,j,t}[m]$ is denoted as $S_{i,j,t}(\omega)$, and the cepstral feature vectors obtained from this spectrum is denoted as $\vec{x}_{i,j,t}$. The entire utterance is represented as a sequence of feature vectors $X_{i,j} = \{\vec{x}_{i,j,1}, \vec{x}_{i,j,2}, \cdots, \vec{x}_{i,j,T}\}$.

In the context of frequency warping, $S^{\alpha}_{i,j,t}(\omega)$ is defined to be $S_{i,j,t}(\alpha\omega)$. The cepstrum feature vectors that are computed from the warped spectrum is denoted as $\vec{x}^{\alpha}_{i,j,t}$, and the warped representation of the utterance is represented as a sequence of the warped feature vectors $X^{\alpha}_{i,j} = \{\vec{x}^{\alpha}_{i,j,1}, \vec{x}^{\alpha}_{i,j,2}, \cdots, \vec{x}^{\alpha}_{i,j,T}\}$.

Additionally, $W_{i,j}$ refers to the word level transcription of utterance j from speaker i. This transcription can either be known in advance or obtained from the speech recognizer.

Finally, we let

- $\mathbf{X}^{\alpha}_i = \{X^{\alpha}_{i,1}, X^{\alpha}_{i,2}, \cdots, X^{\alpha}_{i,N_i}\}$ denote the set of feature space representations for all of the available utterances from speaker i, warped by α,
- $\mathbf{W}_i = \{W_{i,1}, W_{i,2}, \cdots, W_{i,N_i}\}$ denote the set of transcriptions of all of the utterances,
- $\hat{\alpha}_i$ denote the optimal warping factor for speaker i,
- λ denote a given HMM trained from a large population of speakers.

Then, the optimal warping factor for speaker i, $\hat{\alpha}_i$, is obtained by maximizing the likelihood of the warped utterances with respect to the model and the transcriptions

$$\hat{\alpha}_i = \arg\max_{\alpha} \Pr(\mathbf{X}^{\alpha}_i | \lambda, \mathbf{W}_i). \tag{1}$$

However, a closed-form solution for $\hat{\alpha}$ from (1) is difficult to obtain. This is primarily because frequency warping corresponds to a highly nonlinear transformation of the speech recognition features. Therefore, the optimum warping factor is obtained by searching over a grid of 13 factors spaced evenly between 0.88 $\leq \alpha \leq$ 1.12. This range of α is chosen to

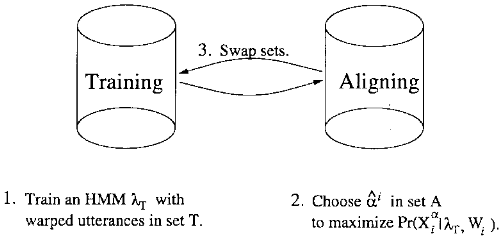
3. Swap sets.
Training
Aligning
1. Train an HMM λ_T with warped utterances in set T.
2. Choose $\hat{\alpha}^i$ in set A to maximize $Pr(X_i^\alpha|\lambda_T, W_i)$.

Fig. 1. HMM training with speaker normalization.

roughly reflect the 25% range in vocal tract lengths found in adults.

B. Training Procedure

The goal of the training procedure is to appropriately warp the frequency scale of the utterances for each speaker in the training set consistently, so that the resulting speaker independent HMM will be defined over a frequency-normalized feature set. It is clear from (1) that the warping factor estimation process requires a preexisting speech model. Therefore, 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 Fig. 1.

First, the speakers in the training data are divided into two sets, training (T) and aligning (A). An HMM, λ_T, is then built using the utterances in set T. Then, the optimal warping factor for each speaker i in set A is chosen to maximize $\Pr(\mathbf{X}_i^\alpha|\lambda_T, \mathbf{W}_i)$. Since we assume the vocal tract length to be a property of the speaker, all of the utterances from the same speaker are used to estimate $\hat{\alpha}$ for that speaker. Sets A and T are then swapped, and we 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.

With a large amount of training data from a large number of speakers, it may not be necessary to divide the data set into half. If the data were not divided into two separate sets, it can be easily shown that the iterative procedure of estimating warping factors and then updating the model always increases the likelihood of the trained model with respect to the warped data. Suppose we use $\hat{\mathbf{X}}_{j-1}$ to denote the set of all warped training vectors from all speakers in iteration $j - 1$, and λ_{j-1} to denote the model trained with this data. Then, in reestimating the warping factors during the jth iteration, the warping factors are chosen to increase the likelihood of the data set, $\hat{\mathbf{X}}_j$, with respect to λ_{j-1}

$$\Pr(\hat{\mathbf{X}}_j|\lambda_{j-1}, \mathbf{W}) \geq \Pr(\hat{\mathbf{X}}_{j-1}|\lambda_{j-1}, \mathbf{W}). \tag{2}$$

In addition, the use of the Baum–Welch algorithm to train λ_j using $\hat{\mathbf{X}}_j$ guarantees the following:

$$\Pr(\hat{\mathbf{X}}_j|\lambda_j, \mathbf{W}) \geq \Pr(\hat{\mathbf{X}}_j|\lambda_{j-1}, \mathbf{W}). \tag{3}$$

By combining (2) and (3), it is seen that the likelihood of the data with respect to the model is increased with each iteration of training

$$\Pr(\hat{\mathbf{X}}_j|\lambda_j, \mathbf{W}) \geq \Pr(\hat{\mathbf{X}}_{j-1}|\lambda_{j-1}, \mathbf{W}). \tag{4}$$

While this informal proof of convergence does not hold when the data is divided in half, empirical evidence is presented in Section III to show that the model likelihood converges even in that case.

C. Recognition Procedure

During recognition, the goal is to warp the frequency scale of each test utterance to "match" that of the normalized HMM model λ_N. Unlike the training scenario, however, only one testing utterance is used to estimate $\hat{\alpha}$, and the transcription is not given. Two procedures are discussed for maximum likelihood estimation of the warping factor. The first, discussed in Section II-C1, is a three-step procedure that requires two recognition passes over the input utterance. The second procedure, described in Section II-C2, chooses the correct warping transformation by classifying the input utterance according to a set of "warp class" models before recognition.

1) Multiple-Pass Strategy: Since no satisfactory solution for direct form estimation of the warping factor in (1) has been obtained, the optimum $\hat{\alpha}$ is found by aligning warped utterances with respect to a hypothesized word string. The following three-step process, as illustrated in Fig. 2, is used.

1) The unwarped utterance $X_{i,j}$ and the normalized model λ_N are used to obtain a preliminary transcription of the utterance. The transcription obtained from the unwarped features is denoted as $W_{i,j}$.
2) $\hat{\alpha}$ is found using (1) as follows:

$$\hat{\alpha} = \arg\max_{\alpha} \Pr(X_{i,j}^{\alpha}|\lambda_N, W_{i,j}).$$

The probability is evaluated by probabilistic alignment of each warped set of feature vectors with the transcription W.
3) The utterance $X_{i,j}^{\hat{\alpha}}$ is decoded with the model λ_N to obtain the final recognition result.

2) Mixture Based Warping Factor Estimation: In the earlier discussions of warping factor estimation, the warping factor is conceptualized simply as a representation of the ratio between a speaker's vocal tract length and some notion of a reference vocal tract length. However, warping factor estimation can also be considered as a classification problem. During speaker normalization, each speaker is first classified according to an estimate of his/her vocal tract length, and class-dependent transformations are then applied to the speech to yield a final feature set, which is used in recognition. From this point of view, speakers are placed into different classes based on the warping factor estimated using their utterances, and the warping factor can be better described as a class identifier. Intuitively, the feature space distributions of untransformed speech from the different classes of speakers would vary due to the acoustic differences of speech produced by vocal tracts of different lengths. Therefore, if statistical models of the feature space distribution of each class are available, it may be possible to determine the warping factor by finding out which

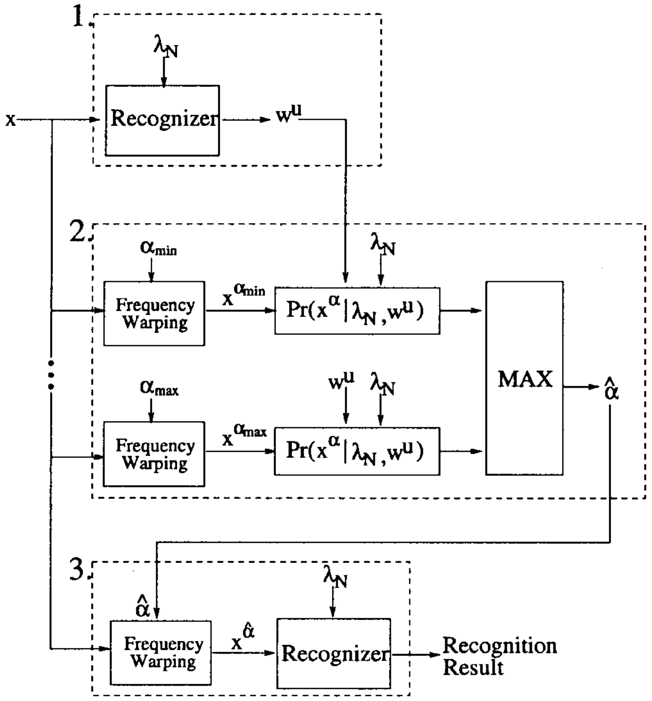

Fig. 2. HMM recognition with speaker normalization.

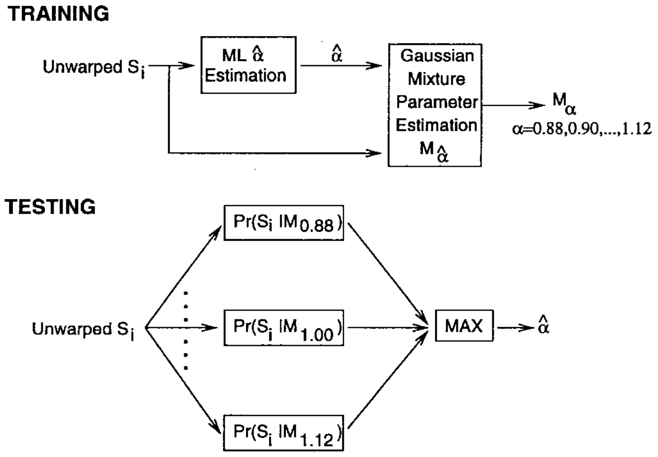

Fig. 3. Mixture based optimal factor estimation.

class distribution is most likely to have generated the given sequence of feature vectors.

The mixture based warping factor estimation technique described here is motivated by this classification perspective of speaker normalization. In training, after warping factors have been determined for all of the speakers using the process shown in Fig. 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 that 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 is chosen for the distribution that yields the highest likelihood over the entire utterance. The speech is warped using this estimated warping factor, and the resulting feature vectors are then used for HMM decoding. A block diagram describing this process is shown in Fig. 3.

This mixture based method results in faster recognition time, because it eliminates the need to obtain a preliminary transcription using the unwarped utterance that is used for performing probabilistic alignment at all of the grid points. However, unlike the method described in Section II-C, it does not take advantage of the temporal information in the signal during warping factor estimation, so that the estimated warping factor may be less accurate.

D. Filterbank Analysis with Frequency Warping

In the previous sections, the processes of HMM training and recognition with speaker normalization were defined independent of the analysis method used to obtain the cepstrum. Here we describe how the Davis–Mermelstein mel-frequency filterbank front-end can be modified to include frequency warping.

The standard Davis–Mermelstein filterbank front-end works by windowing the speech, calculating its magnitude spectrum, passing that through a mel-scaled filterbank, and finally using an inverse cosine transform to arrive at the cepstrum [2]. While it is perhaps most intuitive to perform frequency warping by resampling the speech in the time domain prior to passing the signal through the front-end, it is possible and more efficient to push the warping process into the filterbank front-end itself [5]. Frequency warping can be implemented by simply varying the spacing and width of the component filters of the filterbank without changing the original speech signal For example, to compress the speech signal in the frequency domain, we keep the frequency scale of the signal the same, but stretch the frequency scale of the filters. Similarly, we compress the filterbank frequencies to effectively stretch the signal frequency scale. This process is illustrated in Fig. 4. Since only one single DFT needs to be performed in each frame, there is no need to resample the original signal.

E. Discussion of Bandwidth Differences

When the frequency axis is warped linearly, the bandwidth of the resulting signal differs from that of the original. For the experiments described in this work, the sampling rate is fixed at 8 kHz, imposing a limit on the maximum signal bandwidth of 4 kHz. However, with the warping factors ranging between 0.88 and 1.12, the bandwidths of the warped signals range between 3.52 and 4.48 kHz. Since the search for the “best” warping factor is made using a frequency band from 0 to 4 kHz, the compressed signals do not contain useful information over the entire 4 kHz, and the expanded signals contain information above 4 kHz that is not used. Different bandwidths that result from different warping factors represent a source of mismatch between the warped signal and the model. The filterbank front-end mitigates the mismatch somewhat by blurring the exact location of the band-edge in the warped signals. This results from the wide filters used near the signal band-edge being almost 700 Hz wide.

One possible solution to this problem is to consider warping functions that are piecewise linear or even nonlinear, such that

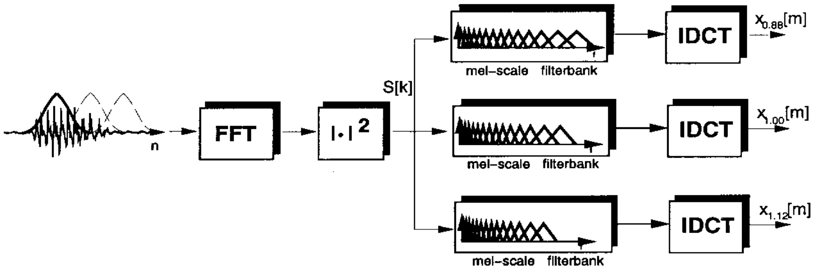

Fig. 4. Mel filterbank analysis with warping.

the bandwidth of the warped signal is the same as that of the original. For example, a piecewise linear warping function like the following may be considered:

$$G(f) = \begin{cases} \alpha f, & 0 \le f \le f_0 \\ \dfrac{f_{\max} - \alpha f_0}{f_{\max} - f_0}(f - f_0) + \alpha f_0, & f_0 \le f \le f_{\max} \end{cases}. \tag{5}$$

In (5), $f_{\max}$ denotes the maximum signal bandwidth, and f_0 can be an empirically chosen frequency that falls above the highest significant formant in speech. The effect of classes of functions like the above should be to reduce the effects of discontinuities at the band-edge. Preliminary experiments using such a piecewise linear warping function for speaker normalization suggested that they may indeed be more robust than a simple linear warping [12]. In addition, Oppenheim and Johnson described a set of nonlinear frequency warping functions that are implementable by a series of allpass filters and map the frequency range $0 \le \omega \le 2\pi$ onto itself [8]. However, because such warping functions have no simple correlation to physical sources of variations, only the linear warping function is used in this paper, and exploration of other functions is left for future work.

III. Baseline Experiments

This section presents an experimental study of the effectiveness of the speaker normalization procedures described in Section II. The principle measure of effectiveness is speech recognition performance obtained on a connected digit speech recognition task over the telephone network. In addition to characterizing the effect on speech recognition performance, a number of additional issues are investigated and discussed. Experiments were performed to understand the ability of the speaker normalization procedures to decrease interspeaker variability, and to produce normalized HMM's that describe the data more efficiently.

The section is divided into five parts. After the task, data base, and speech recognizer are described in Sections III-A and III-B, ASR performance before and after speaker normalization is presented in Section III-C. Section III-D presents an analysis of the distribution of the chosen warping factors among the speakers in the training set to verify the effectiveness of the maximum likelihood warping factor estimation procedure. Section III-E presents statistics on the ability of the warping factor estimation procedure to generate reliable estimates on very short utterances of only one or two digits in length. Finally, Section III-F provides empirical study of the convergence properties of the iterative procedure for estimating the warping factor during HMM training.

A. Task and Data Bases

Two telephone based connected digit data bases were used in this study. The first, DB1, was used in all of the speech recognition experiments. It was recorded in shopping malls across 15 dialect-distinct regions of the United States, using two carbon and two electret handsets that were tested and found to be in good working condition. The size of the vocabulary was eleven words: "one" to "nine," as well as "zero" and "oh." The speakers read digit strings between one and seven digits long in a continuous manner and the utterances were recorded over a long-distance telephone connection. Each utterance ranged from about 0.5 to 4 s in duration. The training utterances were endpointed, whereas the testing utterances were not. All of the data was sampled at 8 kHz. Table I describes the training and testing sets in more detail.

A second connected digit data base, DB2, was used to evaluate properties of the speaker normalization procedures which required more data per speaker than available in DB1. DB2 was taken from one of the dialect regions used for DB1, but contains a larger number of utterances per speaker. In DB2, approximately 100 digit strings were recorded for each speaker. A total of 2239 utterances, or 6793 digits, were available from 22 speakers (ten males, 12 females).

Throughout this paper, word error rate is used to evaluate the performance of various techniques. The error rate is computed as follows:

$$\%\,\text{Error} = 100 \cdot \frac{\text{Sub} + \text{Del} + \text{Ins}}{\text{Total Number of Words}} \tag{6}$$

where "Sub" is the number of substitutions, "Del" is the number of deletions, and "Ins" is the number insertions. These quantities are found using a dynamic programming algorithm to obtain the highest scoring alignment between the recognized word string and the correct word string.

B. Baseline Speech Recognizer

The experiments described in this paper have been conducted using an HMM speech recognition system built at AT&T. Each digit was modeled by eight to ten state continuous density left-to-right HMM's. In addition, silence was explicitly modeled by a single state HMM. The observation densities were mixtures of eight multivariate Gaussian distributions with diagonal covariance matrices. Thirty-nine dimensional feature

TABLE I
DATA BASE DB1 DESCRIPTION. THE ROWS LABELED "CARBON UTTERANCES" AND "ELECTRET UTTERANCES" CORRESPOND TO THE NUMBER OF UTTERANCES THAT WERE RECORDED USING CARBON AND ELECTRET HANDSET TRANSDUCERS

	Training Set	Testing set
# digits	26717	13185
# total utterances	8802	4304
# carbon utterances	4426	2158
# electret utterances	4376	2146
# male speakers	372	289
# female speakers	341	307

vectors were used: normalized energy, $c[1]$–$c[12]$ derived from a mel-spaced filterbank of 22 filters, and their first and second derivatives. The performance metric used was word error rate. This configuration is used for all of the experiments described in this section unless otherwise noted.

C. Speech Recognition Performance

Table II shows the recognition word error rate on DB1 using only the baseline recognizer, and using the baseline recognizer with the speaker normalization procedures. 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. The error rates for utterances through the carbon and electret handsets are shown separately in the second and third columns, and averaged in the last column.

There are several observations that can be made from Table II. 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. 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 by [9], and a possible explanation for the performance discrepancy was provided there. Finally, this performance difference between carbon and electret transducers is reduced after speaker normalization.

While it is important that speech recognition performance be the final criterion for judging the performance of any speaker normalization procedure, it is also important to understand the behavior of the procedure at a more fundamental level. In the remaining experiments presented in the section, the frequency warping procedure is investigated in terms of its effect on the distribution of the estimated warping factors and its effect on the characteristics of the HMM.

D. Distribution of Chosen Warping Factors

In evaluating the effectiveness of the warping factor estimation procedure, two issues are of concern. First, while there is no absolute measure of the "correct" warping factor for each speaker, the chosen warping factors over the entire speaker population should satisfy our intuition about the distortions caused by vocal tract length variations. Secondly, the normalization procedures should result in speech utterances and model representations that exhibit reduced interspeaker variation. These two issues are addressed in this and the next sections.

TABLE II
WORD ERROR RATE BEFORE AND AFTER USING SPEAKER NORMALIZATION

Condition	Carbon	Electret	All
Baseline	2.8 %	4.1 %	3.4 %
Speaker Normalization	2.4 %	3.1 %	2.7 %

Histograms of the chosen warping factors for the speakers in the training set are shown in Fig. 5. On average, about 15 utterances are used to estimate the warping factor for each speaker. The warping factors chosen for the males are shown on top, and those for the females shown on the bottom. 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.

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 that may have originally been out-of-band. If parts of the discarded spectra carry information useful for recognition, the maximum likelihood warping factor estimation is likely to be biased toward frequency compression.

We note here that the mean of estimated warping factors is not required to be 1.0 under model-based warping factor estimation because any notion of a "reference" vocal tract length must be considered in reference to the model parameters. It is the relative differences in warping factors chosen for different speakers which is most significant to the ability of the procedure to generate a consistently frequency-normalized feature set.

E. Warping Factor Estimation with Short Utterances

A major assumption made in the paper is that the vocal tract length of the speaker is a long-term speaker characteristic. Therefore, it is assumed that the variations in effective vocal tract length due to the production of different sounds do not significantly affect the warping factor estimation process. Under this assumption, with "sufficient" amounts of data

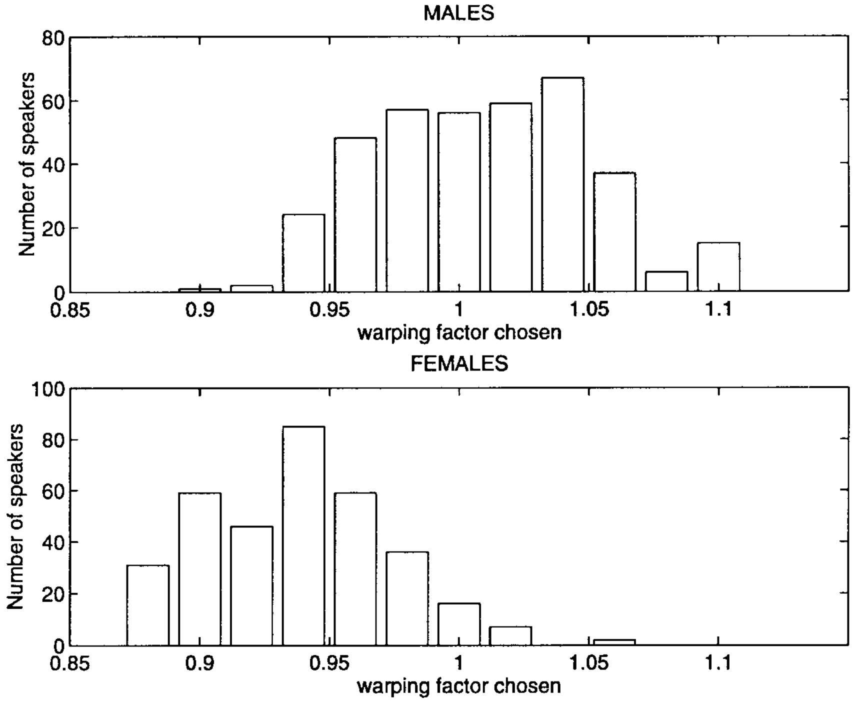

Fig. 5. Histogram of warping factors chosen for speakers in the training set.

for each utterance, the warping factor estimates should not vary significantly among different utterances by the same speaker. This section presents an experiment that attempted to test and better understand this assumption by gathering and examining statistics reflecting how the warping factor estimates change across utterances of different durations for the same speaker. These statistics also reflect the ability of the maximum likelihood based warping factor estimation method to generate reliable estimates even when the utterances are very short.

In this experiment, the three-step speaker normalization recognition procedure depicted in Fig. 2 was used on the data in DB2, where approximately 100 utterances are available for each of 22 speakers. The set of all utterances $\mathbf{X}_i$ from speaker i is divided roughly evenly into two sets based on the number of digits in each utterance. The set of utterances containing one or two digits is denoted by S_i, and the set of utterances containing three to seven digits is denoted by L_i. For each speaker i, the means and standard deviations of the warping factor estimates for utterances within each of S_i and L_i are computed. The differences between the means computed for S_i and L_i are examined to observe any significant differences in the warping factor estimates as the amount of available data increases. The standard deviations are also compared to see if the variance of warping factor estimates over different utterances decreases with longer utterances.

Fig. 6 shows two plots in which the mean and standard deviation of warping factor estimates for utterances in S_i are plotted against those statistics computed over L_i, for all of the speakers in DB2. In the top plot, the x-axis denotes the mean of the warping factor estimates among utterances in set S_i, and the y-axis denotes the mean of the warping factor estimates among utterances in set L_i. Points marked by "*" correspond to the female speakers, and those marked by "+" correspond to the male speakers. In the bottom plot, the x-axis denotes the standard deviation of the warping factor estimates among utterances in set S_i, and the y-axis denotes the standard deviation of the warping factor estimates among utterances in set L_i. "X"'s are used to marked the data points. In both plots, the line $y = x$ is drawn as a reference to aid in discussing the trends in the plotted points.

Two important observations can be made based on the top plot of Fig. 6. First, the means of the warping factor estimates of the male speakers are always higher than those of the female speakers regardless of the length of the utterance. Second, the mean of the warping factor estimates over the longer utterances is significantly higher than the mean over the shorter utterances among the male speakers. This difference ranged from only 1% to almost 7.5%. While the cause of this trend is not clear, one possible explanation may be that for the shorter utterances, a larger portion of the available data consists of silences and other nonvoiced sounds for which the frequency warping compensation model is not appropriate. Since the test utterances are not endpointed, a large portion of the single-digit utterances is not speech. The computed likelihood over nonspeech frames may be higher for feature vectors corresponding to frequency compression, because frequency compression results in the inclusion of portions of the frequency spectrum which would have been discarded otherwise.

Two observations can be made from the second plot of Fig. 6. First, it is clear that the standard deviation of the warping factor estimates generally decreases for the set of longer utterances. This implies that the warping factor estimation process does become more "stable" as the amount of available data increases. Second, the standard deviation

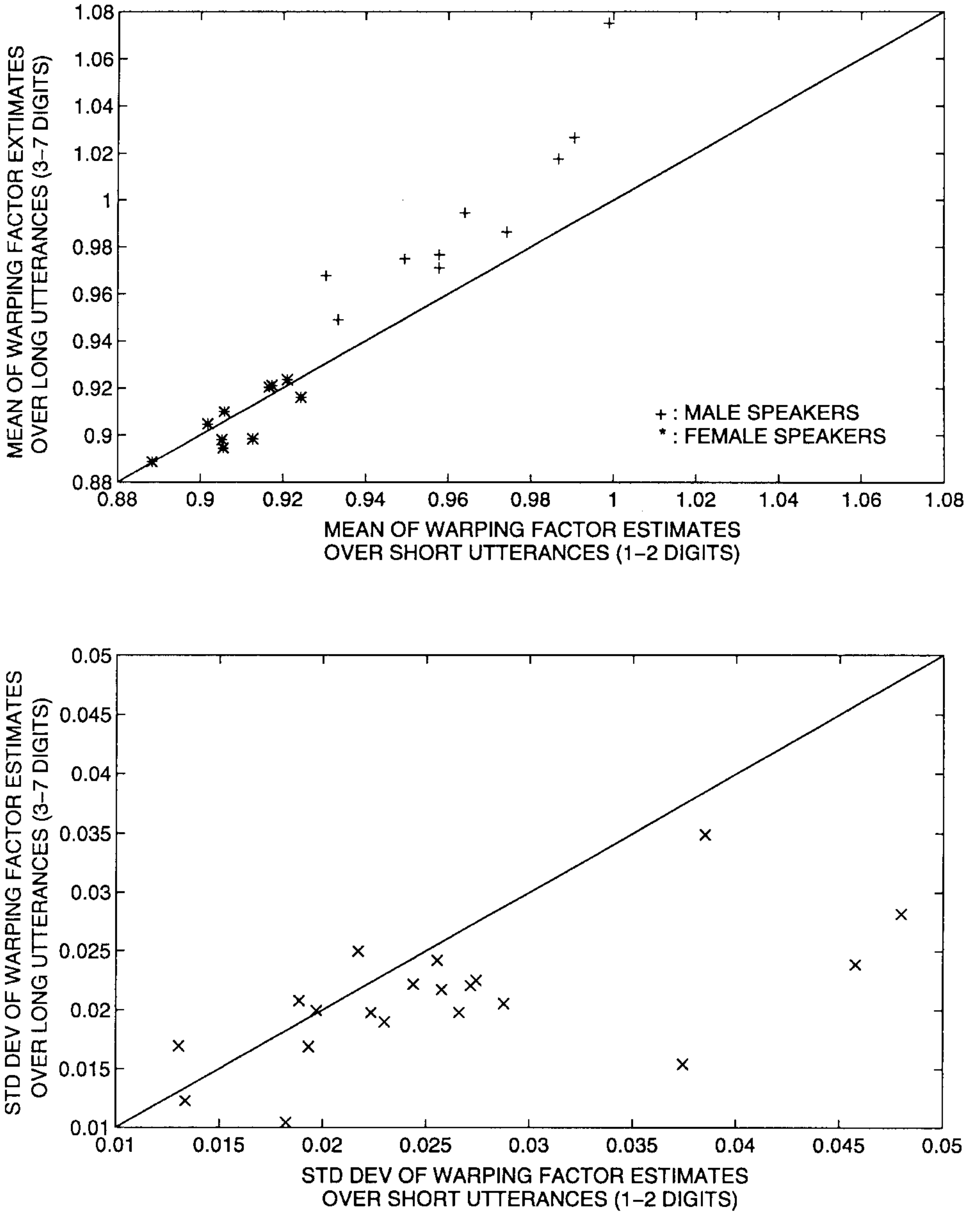

Fig. 6. Comparisons of means and standard deviations among utterances of different lengths.

of the warping factor estimates over the shorter utterances is less than 0.04 for a majority of the speakers. Taking into account that the possible warping factors are spaced 0.02 apart in the grid search process, we see that the warping factor estimation process produces estimates that do not vary greatly from utterance to utterance, depending on the particular phonetic content of the utterance. Hence, these observations are consistent with our assumption that the vocal tract length of the speaker does not change significantly with the sound being produced.

F. Convergence of Model Training Procedure

This section presents an experiment performed to understand the convergence properties of the iterative training procedure. In the standard Baum–Welch HMM training algorithm, the likelihood of the training data with respect to the models is mathematically guaranteed to increase at the end of each iteration. While the iterative normalized-HMM training procedure is not guaranteed to converge mathematically, we study changes in recognition error rate on the training and testing data as the number of training iterations is increased. This experiment also serves to further test whether the frequency warping procedures are indeed reducing the speaker variability (at least in the training set), and that the normalized HMM's are becoming more efficient over the iterations.

Table III shows how the model likelihood and recognition word error rate on the training and testing data changes as the number of training iteration increases. In the table, the second column shows the average log-likelihood of the warped training data with respect to the frequency-normalized model. The third column shows recognition performance when the frequency-normalized models were used to decode the same data that was used to train them. The fourth column shows

TABLE III
AVERAGE MODEL LOG-LIKELIHOOD AND WORD ERROR RATE ON TRAINING AND TESTING DATA AFTER 0–3 TRAINING ITERATIONS WHERE SPEAKER NORMALIZATION WITH FREQUENCY WARPING IS APPLIED TO THE TRAINING DATA

No. of Iter.	Model Log-Likelihood	Train Set	Test Set
0	-32.08	2.4 %	2.9 %
1	-31.35	1.7 %	2.7 %
2	-31.13	1.3 %	2.9 %
3	-31.09	1.3 %	2.9 %

TABLE IV
PERFORMANCE OF MORE EFFICIENT SPEAKER NORMALIZATION RECOGNITION PROCEDURES

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%

recognition results on the testing set using the three-step process described in Section II-C. The model used for the results shown in the first row, 0 iterations, was built with unwarped data.

From the table, it is clear that multiple iterations increased the likelihood of the data with respect to the model. The improved performance on the training data shows that a significant amount of variance among the speakers in the training set has been reduced. However, while multiple training iterations improved the recognition performance on the training data dramatically, recognition performance on the test data did not improve. Additionally, it is interesting that using the speaker normalization procedure during recognition with an unnormalized HMM (first row of table) still offers a significant improvement over the baseline. This is due to the fact that the speaker normalization procedure used during recognition is, on its own, reducing the amount of mismatch between the testing speakers and the models of the training speakers.

IV. EFFICIENT APPROACHES TO SPEAKER ROBUST SYSTEMS

This section considers the frequency warping approach to speaker normalization in terms of its computational requirements. It is also considered in relation to existing methods designed to reduce the effects of speaker and channel variability on speech recognition performance. In comparing the frequency warping approach to speaker normalization with these other techniques, we gain additional insight into the advantages and disadvantages of using this physiologically motivated procedure over other “statistically based” compensation and modeling procedures.

This section presents several sets of experimental results in three parts. Section IV-A compares the performance of the mixture based warping factor estimation procedure described in Section II-C-2 with that achieved using variations of the multiple-pass method. Section IV-B studies how speaker normalization procedures compare with gender-dependent models and cepstral mean normalization. Finally, Section IV-C studies whether the effects of speaker normalization can be achieved simply by using more parameters in the HMM. A closely associated question is whether the complexity of the HMM’s affects the amount of performance gain achieved by speaker normalization.

A. Performance of Mixture Based Warping Factor Estimation

Table IV shows the results of applying the mixture based warping factor estimation procedure described in Section II-C2. The first row of the table gives the error rate for the baseline speech recognizer described in Section III-B without frequency warping. The search method referred to as “HMM-based” in the first column refers to the multiple-pass procedure described in Section II-C1, which involves performing probabilistic alignment with respect to a hypothesized transcription at each possible warping factor. The second through fifth rows of the table show the recognition performance when the number of possible warping factor values is decreased from 13 to 3 points. The last row of the table shows the recognition error rate when the mixture based warping factor estimation method is used. Each of the mixtures used 32 multivariate Gaussians. This experiment was performed on speech data base DB1.

A comparison among rows two through five in Table IV 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 three. Compared with the baseline system with no frequency warping, allowing only three possible warping factors still offers a 15% reduction in error rate.

Comparing the second and last rows of the Table IV, 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.

B. Comparison with Other Approaches

There has been a large body of work on characterizing and compensating for speaker variability in speech recognition. In this section, speaker normalization is compared with two other approaches to improve an ASR system’s robustness to speaker variability. First, gender-dependent modeling, an example of an approach to speaker class-dependent modeling, is implemented and tested. Second, we investigate cepstral mean normalization (CMN), an example of a technique that uses long-term spectral averages to characterize fixed speaker and channel characteristics. These techniques are described, and the recognition results are presented below.

1) Gender-Dependent Models: In gender-dependent modeling, two sets of HMM’s are trained: one using speech from

TABLE V
PERFORMANCE OF SPEAKER NORMALIZATION PROCEDURES AS COMPARED TO USING NO WARPING, TO USING GENDER-DEPENDENT MODELS, AND TO CEPSTRAL MEAN NORMALIZATION

Condition	Carbon	Electret	Both
Baseline(no warping)	2.8%	4.1%	3.4%
Speaker Normalization	2.4%	3.1%	2.7%
GD Models	2.3%	3.4%	2.9%
CMN	2.5%	3.7%	3.1%

males and another using speech from females. During the Viterbi search for the most likely state sequence in recognition, these HMM's are used to create two separate gender-specific networks. Again, the maximum likelihood criterion is used to find the best state sequence. Because the average vocal tract length differs significantly between males and females and GD modeling can capture such differences, GD models can be considered to "approximate" the speaker normalization process. For this reason, it is important to understand whether the extra computational requirements of speaker normalization results in higher performance.

2) Cepstral Mean Normalization: Long-term spectral averages have been used to characterize both speaker and channel characteristics [5]. CMN is an example of one of these techniques that has been successfully used in ASR to compensate for both types of distortions. In our implementation of CMN, the mean of the cepstral vectors in the nonsilence portions of each utterance is assumed to characterize long-term characteristics of the speaker and channel. Therefore, the cepstral mean is computed and subtracted from the entire utterance. Two processing steps are taken. First, an energy-based speech activity detector is used over the entire utterance, and the cepstral mean is computed over those frames that are marked as speech. Then, new feature vectors are obtained by subtracting this mean from each cepstral vector in the utterance. In cases where long delays cannot be tolerated, the estimate of the mean vector can be updated sequentially by applying a sliding window to the utterance. The use of a speech activity detector is also very important to the successful application of this technique. Recognition performance has been found to degrade when the mean vector is computed over a large number of silence frames. By forcing the cepstral mean to be zero for all utterances in training and in testing, CMN compensates for differences in convolutional distortions that may arise from either speaker and channel differences between training and testing.

3) Experimental Results: Table V shows recognition word error rates on DB1 using the baseline models, speaker normalization, gender-dependent models, and CMN. The errors are shown separately for utterances spoken through the carbon and electret handsets in the first and second columns. The third column shows the overall error rate. The baseline and speaker normalization results are the same as those shown in Table I. All models used eight to ten states per digit, and mixtures of eight multivariate Gaussians as observation densities. We note here that since two sets of models are used in gender-dependent models, these models used twice the number of model parameters as the other methods.

The overall results show that the error rates were reduced by 20% with speaker normalization, by 15% with gender-dependent models, and by 10% with CMN. For all of the conditions in the experiment, recognition performance on the test data spoken through the carbon transducers is better than that for the electret transducers, even though the model was trained from data spoken through both carbon and electret tranducers. This result is consistent with those presented in [9], and some possible explanations are presented there.

4) Speaker Normalization versus Class-Dependent Models: Gender-dependent modeling is one example of a large class of techniques where class-dependent models are trained for different speaker groups according to gender, dialect, or by automatic clustering of speakers [6], [10]. Using this set of procedures, the separate HMM's are used in parallel during recognition to simultaneously determine the class that the speaker belongs to, as well as the string transcription of the utterance. It is important to realize that, with enough data, a similar approach could be taken for the speaker normalization procedures. One could train different sets of HMM's using training speakers assigned to each warping factor, and decode using all of the HMM's. However, one common problem in training class-dependent models is that as the number of classes increases, the models may become undertrained.

In class-dependent modeling techniques like gender-dependent models, no attempt is made to explicitly characterize and compensate for the defining aspects of different classes in feature space so that the spaces modeled by the class-dependent HMM's can become more similar. As a result, there is a need to build complete models carrying both phonetic and classification information for each class. The amount of available training data, therefore, limits the number of speaker classes. In the speaker normalization approach, however, the interclass differences are modeled using a relatively simple parametrization and transformation. It is possible to transform the data from different classes into the same class, and build a model using all of the data, without the occurrence of undertrained models even with a large number of classes. The additional "resolution" in speaker class divisions allows for better recognition performance with speaker normalization. This is clear from the second and third rows in Table V, where the gender-dependent models actually used double the number of model parameters than speaker normalization. The possibility of dividing the training speaker set into 13 different classes is a direct consequence of the physical model and simple parameterization of the transformation process.

C. HMM Parameterization

This section attempts to determine whether the performance improvements given by speaker normalization can be observed by simply increasing the complexity of the HMM's used. When more Gaussians per mixture are used to represent the observation density in each HMM state, the feature space distribution can be more accurately described. However, more complex HMM's use more parameters, incurring greater storage and computational requirements. Moreover, with a limited

TABLE VI
PERFORMANCE OF SPEAKER NORMALIZATION OVER DIFFERENT COMPLEXITY HMM'S

# Gaussians/mix.	Baseline	Warping	% Improvement
8	3.4 %	2.7 %	20 %
16	2.8 %	2.4 %	14 %
24	2.3 %	2.0 %	13 %
32	2.7 %	–	–

amount of training data, there may not be enough data to reliably estimate all of the parameters of highly complex HMM's, resulting in undertrained models.

In this experiment, the size of the Gaussian mixtures used in the observation densities is increased incrementally, and the performance of using the baseline recognizer alone and speaker normalization on DB1 is observed. The results are shown in Table VI. The rows of the table show the recognition results as the number of Gaussians used in each observation density mixture is increased. The second and third columns show the error rates of the baseline and speaker normalization methods. The last column show the amount of error reduction offered by frequency warping in percent.

From the baseline case, it is clear that as the number of Gaussians per mixture increases to 32, the models become undertrained, and no further performance improvements can be observed. The table shows that when the baseline models are not undertrained, using frequency warping with simpler models results in error rates similar to those obtainable using more complex models. The trade-off here is between the computational requirements associated with the normalization procedure and the memory storage requirements associated with higher complexity models. When baseline models are undertrained, however, it is clear that frequency warping is better than simply increasing the complexity of the HMM parameterization.

V. SUMMARY

In this paper, we developed and evaluated a set of speaker normalization procedures that explicitly model and compensate for the effects of variations in vocal tract length by linearly warping the frequency axis of speech signals. The degree of warping applied to each speaker's speech was estimated using the speaker's utterance(s) within a model-based maximum likelihood framework. Using a model-based criterion for estimating 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 normalization in speech recognition, none have used an optimization criterion that is consistent with that used in the recognizer to estimate the parameters of the transformation.

The effectiveness of this set of speaker normalization procedures was examined in an experimental study performed using a telephone based digit recognition data base in which the utterances are between one and seven digits in length. Recognition results showed that using the frequency warping approach to speaker normalization reduces the word error rate by about 20% on this task. The best performance obtained was a word error rate of 2.0%.

The frequency warping approach to speaker normalization was compared to other simple methods for reducing the effects of speaker and channel variability on speech recognition performance. These methods included cepstral mean normalization, gender-dependent modeling, and higher complexity HMM parameterizations. Experimental results showed that the physiologically based speaker normalization procedures investigated in this paper perform significantly better than these statistically motivated methods, which do not explicitly model the effects of known physical sources of variation.

Several unresolved issues remain. The first is the parameterization of the warping function used in speaker normalization. A second issue is whether the procedure should be applied at the segmental level as opposed to applying it to an entire utterance. Finally, a last issue concerns the development of a more consistent criterion for combining HMM parameter estimation with speaker normalization during training. The overall advantage of the procedure as it is currently implemented is that it represents a very efficient physiologically motivated procedure for reducing the mismatch between an input utterance and a speech recognition model.

REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II,* 1994.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-28, pp. 357–366, Aug. 1980.

[3] G. Fant, "Non-uniform vowel normalization," Speech Transmiss. Lab. Rep., Royal Inst. Technol., Stockholm, Sweden, 1975, vols. 2/3, pp. 1–19.

[4] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing,* vol. 2, pp. 291–298, Apr. 1994.

[5] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP'96,* pp. 353–356.

[6] J. D. Markel, B. T. Oshika, and A. H. Gray, Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-25, pp. 330–337, Aug. 1977.

[7] L. Mathan and L. Miclet, "Speaker hierarchical clustering for improving speaker independent HMM word recognition," in *Proc. ICASSP 90,* pp. 149–152.

[8] Y. Ono, H. Wakita, and Y. Zhao, "Speaker normalization using constrained spectra shifts in auditory filter domain," in *Proc. EUROSPEECH'93,* pp. 355–358.

[9] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," *Proc. IEEE,* vol. 60, pp. 681–691.

[10] A. Potamianos, L. Lee, and R. C. Rose, "A feature-space transformation for telephone based speech recognition," in *Proc. EUROSPEECH'95.*

[11] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[12] R. Roth *et al.,* "Dragon systems' 1994 large vocabulary continuous speech recognizer," in *Proc. Spoken Language Systems Technology Workshop,* 1995.

[13] W. Torres, personal communication.

[14] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-25, pp. 183–192, Apr. 1977.

Li Lee (S'92) received the B.S. and M.Eng. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1995. She is currently working toward the Ph.D. degree at the Digital Signal Processing Group, MIT.

From 1991 through the present, she has held internship positions in IBM, AT&T Bell Laboratories, AT&T Laboratories-Research, and Xerox Palo Alto Research Center. Her research activities have included computer networking, speech recognition, digital hardware design, and document display interfaces. Her current research interests include signal processing in distributed computing environments and communications.

Ms. Lee is a member of Tau Beta Pi and Eta Kappa Nu, and holds an AT&T Graduate Fellowship.

Richard Rose (M'88) received the B.Sc. degree (summa cum laude) and the M.Sc. degree (magna cum laude) in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1983 and 1987, respectively, and the Ph.D. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1990.

From July 1983 to July 1988, he was employed by Tadiran Ltd, Israel, where he carried out research in the areas of image coding, image transmission through noisy channels, and general image processing. After completing the Ph.D. degree in 1990, he joined the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently an Associate Professor. His research interests are in information theory, source and channel coding, pattern recognition, image coding and processing, and nonconvex optimization in general.

Dr. Rose was co-recipient of the William R. Bennett Prize Paper Award of the IEEE Communications Society (1990).