Proc. of Intl. Conf. Acoustic, Speech and Signal Proc., Munich, Germany, April 1997.

ON COMBINING FREQUENCY WARPING AND SPECTRAL SHAPING IN HMM BASED SPEECH RECOGNITION

Alexandros Potamianos and Richard C. Rose

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

ABSTRACT

Frequency warping approaches to speaker normalization have been proposed and evaluated on various speech recognition tasks [1, 2, 3]. These techniques have been found to significantly improve performance even for speaker independent recognition from short utterances over the telephone network. In maximum likelihood (ML) based model adaptation a linear transformation is estimated and applied to the model parameters in order to increase the likelihood of the input utterance. The purpose of this paper is to demonstrate that significant advantage can be gained by performing frequency warping and ML speaker adaptation in a unified framework. A procedure is described which compensates utterances by simultaneously scaling the frequency axis and reshaping the spectral energy contour. This procedure is shown to reduce the error rate in a telephone based connected digit recognition task by 30-40%.

1. INTRODUCTION

A major hurdle in building successful automatic speech recognition applications is non–uniformity in performance across a variety of conditions. Many successful compensation and normalization algorithms have been proposed in the literature dealing with different sources of variability. Typical examples in telecommunications applications of speech recognition include inter–speaker, channel, environmental, and transducer variability. In practice, a speech utterance may be simultaneously affected by many sources of variability, and there may be many acoustic correlates associated with a given source of variability. As a result, it is important that different procedures for compensating for acoustic distortions be tightly coupled with one another. This paper attempts to address how linear model transformation and speaker normalization by frequency warping can be implemented as a single procedure to compensate for these sources of variability.

Model adaptation techniques have been used for improving the match between a set of adaptation utterances and the hidden Markov model (HMM) used during recognition. The parameters of a linear transformation are estimated using a maximum likelihood criterion and the transformation is applied to the HMM parameters [4]. A common problem among these techniques is the existence of speakers in a population whose speech recognition performance does not improve after adaptation. This can be especially true for unsupervised, single utterance based adaptation scenarios. It is generally thought that only those distributions in the model that are likely to have generated the adaptation observations have a chance to be mapped to the target speaker. Therefore, if the "match" between the model and the adaptation utterance is not reasonably "good" to begin with and the number of adaptation utterances is limited, then the utterance cannot "pull" the model to better match the target speaker.

Speaker normalization by frequency warping has been used for estimating a frequency warping function that is applied to the input utterance so that the warped utterance is better matched to the given HMM model. As is the case for model adaptation, there exists a subset of utterances for which frequency warping does not improve performance. The ineffectiveness of speaker normalization for these utterances is thought to be due to the interaction of other sources of variability in the process of estimating the "best" warping function. If both the model adaptation and speaker normalization procedures are limited by the initial relationship between the HMM model and the input utterance, then perhaps a solution to this problem is to search for an optimum warping function and an optimum model transformation in the same procedure. This is the principle focus of this paper.

The paper is organized as follows. First, the frequency warping based speaker normalization procedure is described in Section 2. In Section 3, a combined procedure for frequency warping and model adaptation is described and applied to a single utterance based adaptation paradigm. A discussion of the application of frequency warping as applied to childrens' speech recognition using HMM models trained from adult speakers is given in Section 4. Finally, discussion and summary is provided in Sections 5 and 6.

2. SPEAKER NORMALIZATION USING FREQUENCY WARPING

In [3], an efficient frequency warping algorithm for speaker normalization was proposed and applied to telephone based speech recognition. The frequency warping approach to speaker normalization compensates mainly for inter-speaker vocal tract length variability by linear warping of the frequency axis by a factor α. By applying frequency warping during both training and recognition it was shown that word error rate can be reduced by approximately 20%. The frequency warping algorithm described in [3] is briefly presented next.

Frequency warping is implemented in the mel-frequency filterbank front-end by linear scaling of the spacing and bandwidth of the filters. Scaling the front-end filterbank is equivalent to resampling the spectral envelope using a compressed or expanded frequency range. The speaker normalization algorithm works as follows. For each utterance, the optimal warping factor $\hat{\alpha}$ is selected from a discrete ensemble of possible values so that the likelihood of the warped utterance is maximized with respect to a given HMM and a given transcription. The values of the warping factors in the ensemble typically vary over a range corresponding to frequency compression or expansion of approximately ten percent. The size of the ensemble is typically ten to fifteen discrete values. Let $X^\alpha = g_\alpha(X)$ denote the sequence of cepstrum observation vectors where each observation vector is warped by the function $g_\alpha()$, and the warping is assumed to be linear. If λ denotes the parameters of the HMM model, then the optimal warping factor is defined as

$$\hat{\alpha} = \arg\max_{\alpha} P(X^\alpha | \alpha, \lambda, H) \quad (1)$$

where H is a decoded string obtained from an initial recognition pass. Finally, the frequency warped observation vector $X^{\hat{\alpha}}$ is used in a second recognition pass to obtain the final recognized string. Note that the procedure is computationally efficient since maximizing the likelihood in Eq. 1 involves only the probabilistic alignment of the warped observation vectors X^α to a single string H.

3. SPEAKER NORMALIZATION AND SPEAKER ADAPTATION

3.1. Providing a Larger Ensemble of Alternatives

This section describes a simple method for implementing a parametric linear transformation on the HMM model and a parametric frequency warping of the input utterance under a single statistical framework. The method can be interpreted as a means for expanding the ensemble of alternatives that are being evaluated during adaptation thus obtaining a better match between the input utterance and the model.

In Section 2, frequency warping was described as a method of transforming an utterance according to a parametric transformation $g_\alpha()$ in order to maximize the likelihood criterion given in Eq. 1. There is a large class of maximum likelihood based model adaptation procedures that can be described as parametric transformations of the HMM model. For these procedures, we let $\lambda_\gamma = h_\gamma(\lambda)$ denote the model obtained by a parametric linear transformation $h_\gamma()$. The form of the transformation depends on a number of issues including both the nature of the sources of variability and the number of observations available for estimating the parameters of the transformation. However, the same maximum likelihood criterion is used for estimating γ as was used for estimating α:

$$\hat{\gamma} = \arg\max_{\gamma} P(X | \gamma, \lambda_\gamma, H) \ . \quad (2)$$

Our goal is to combine the frequency warping and model adaptation methods in a maximum likelihood framework. The optimal parameters of the model transformation $\hat{\gamma}$ and the frequency warping $\hat{\alpha}$ can be simultaneously estimated so that

$$\{\hat{\alpha}, \hat{\gamma}\} = \arg\max_{\{\alpha,\gamma\}} P(X^\alpha | \alpha, \gamma, \lambda_\gamma, H) \ . \quad (3)$$

The potential of this class of procedures was investigated in the context of speaker adaptation from single utterances. In this case, $h_\gamma()$ is a set of transformations applied to the means of the model distributions or the observation sequence. Two procedures are considered for implementing the combined optimization implied by Eq. 3 and are discussed next.

The first of these implementations is illustrated by the block diagram in Fig. 1. An ensemble of HMM models is generated λ^α, where each model is trained from observation vectors warped according to $g_\alpha()$. The optimum model transformation is obtained by searching over an ensemble of "warp class" models, λ_k^α, $k = 1, \ldots, K$, and also over a set of N-best string candidates H_n, $n = 1, \ldots, N$. In a separate study, Matsui and Furui estimated the transformation, $h_\gamma(\lambda)$, by searching over the set of N-best candidates alone [5]. For each model and each string candidate, we solve for the $\hat{\gamma}_{\alpha,H_n}$ which maximizes $P(X|\lambda^{\hat{\alpha}}, H_n, \gamma)$. Of course, this procedure results in an increase in computational complexity by a factor of KxN, the size of the combined ensemble of warp class models and N-best candidates respectively.

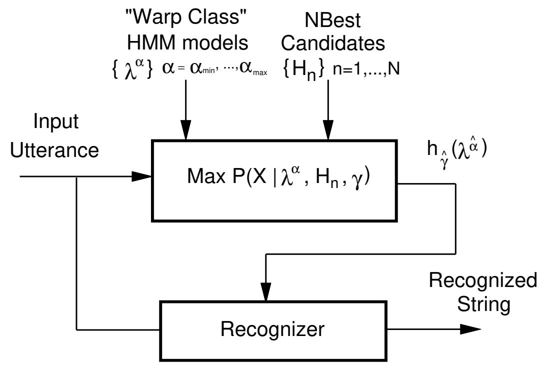

Figure 1: Single utterance based speaker adaptation where optimum model transformation is computed with respect to an ensemble of models and an ensemble of word transcriptions.

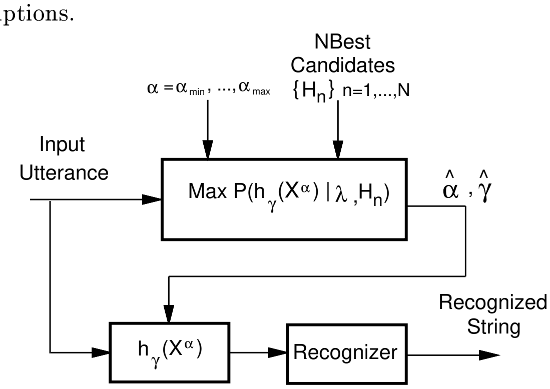

Figure 2: Single utterance based speaker adaptation where optimum transformation is applied on the observation sequence.

A second more computationally efficient implementation of the combined optimization procedure can be used for simple definitions of the model transformation, $h_\gamma()$. If the model transformation corresponds to a single fixed transformation applied to all HMM means, it can be applied to the observation sequence instead of the HMM.[1] Similarly, instead of building "warp class" models, frequency warping can be applied directly on the observation sequence during testing This simplifies significantly both the computational load and the memory requirements of the speaker normalization and adaptation procedure. As before, we attempt to simultaneously optimize the transformation with respect to α and γ by maximizing the likelihood $P(h_\gamma(X^\alpha)|\alpha, \gamma, H, \lambda)$.

The procedure is described in Fig. 2. For each warping index α and each string candidate H_n, we solve for the $\hat{\gamma}_{\alpha,H_n}$ which maximizes $P(h_\gamma(X^\alpha)|\alpha, \gamma, H_n, \lambda)$. Next, the warping index $\hat{\alpha}$ is selected so that $P(h_{\hat{\gamma}}(X^\alpha)|\alpha, \hat{\gamma}, H_n, \lambda)$. is maximized. Finally, the transformed observation vector $h_{\hat{\gamma}}(X^{\hat{\alpha}})$ is used in a second recognition pass to obtain the

[1] The inverse transformation has to be applied to the observations. For simplicity we use the same notation for transformations applied to either the observations or to the HMMs.

Adaptation Method	Digit Error
Baseline	3.4 %
Baseline+Warp Trained	2.9 %
Warp	2.5 %
Bias	2.5 %
Warp+Bias	2.2 %
N-best+Warp+Bias	2.1 %

Table 1: Results of speaker adaptation experiments.

final recognized string.

3.2. Adaptation Experiments

Single utterance based adaptation experiments were performed on a connected digit speech corpus that was collected over the public switched telephone network. The training corpus consisted of 8802 single to seven digit utterances (a total of 26717 digits), and the test corpus contained 4304 utterances (13185 digits) from 242 male and 354 female speakers. For these experiments, the form of the transformation is linear frequency warping followed by a single linear bias applied to the warped observation sequence

$$h_\gamma(X^\alpha(t)) = X^\alpha(t) - \gamma, \tag{4}$$

where X^α(t) is the cepstrum observation vector at time t warped by $g_\alpha()$. To estimate the optimum γ it was assumed that only the highest scoring Gaussian in the mixture contributes to the likelihood computation thus simplifying the estimate

$$\hat{\gamma} = (\sum_t \frac{X^\alpha(t) - \mu_{j(t)}}{\sigma_{j(t)}})/(\sum_t \frac{1}{\sigma_{j(t)}}) \tag{5}$$

where $\mu_{j(t)}, \sigma_{j(t)}$ are the mean and variance of the most active Gaussian j in the mixture at time instant t.

Speaker adaptation results are reported in Table 1 in terms of the percentage of digits that were erroneously recognized. Context independent continuous Hidden Markov digit models with mixtures of eight Gaussians per state were used for recognition. The "Baseline" digit accuracy is shown in the first row of Table 1. The second baseline experiment, labeled "Baseline+Warp Trained", refers to the improved acoustic models obtained by applying the frequency warping algorithm during training. The training procedure was as follows. First, the optimum linear frequency warping factor $\hat{\alpha}$ was estimated for each speaker in the training set so that $P(X^\alpha|\alpha, \lambda, H_c)$ was maximized, where H_c is the known transcription corresponding to X. Then, improved state alignment was obtained using the warped observation vectors $X^{\hat{\alpha}}$. Finally, HMM models were trained from the original (unwarped) utterances X using the segmentation information obtained from the warped utterances. A 15% reduction of word error rate in our test set was achieved by using warping during training. The "Warp Trained" HMMs are used for the adaptation experiments in the remainder of this section.

Next, we compare the performance of the speaker adaptation algorithms outlined in the previous sections when a single utterance is used to estimate the transformation parameters. The third row of the Table, "Warp", refers to the warping algorithm of Section 2. The amount of linear frequency scaling ranges from 12% compression to 12% expansion and a total of 13 warping factors are allowed in this range. The fourth row of the Table, "Bias", displays the recognition rate when a single linear bias is estimated for the whole utterance without the use of warping. The optimal bias vector $\hat{\gamma}$ maximizes $P(h_\gamma(X)|\gamma, \lambda, H)$, where H is the corresponding transcription obtained from a preliminary decoding pass.

The fifth row of Table 1, labeled "Warp+Bias", refers to warping and bias estimation applied in cascade. Note that a separate bias vector $\hat{\gamma}_\alpha$ was computed and subtracted from each warped observation sequence X^α before the optimal warping index $\hat{\alpha}$ was selected. We have observed that joint optimization of the bias vector and the warping index provides additional performance improvement over separately optimizing the bias and the warping index. This is in agreement with our claim in Section 3.1 that the combined optimization of both model transformation and frequency warping is important for obtaining a better match between the utterance and the model.

The last row in Table 1 labeled "N-best+Warp+Bias", shows the performance of the complete procedure described in Fig. 2, i.e., warping and bias estimation applied to the top four scoring transcriptions. It is interesting to note that by including a larger ensemble of models as "starting points" for adaptation, the word error rate was reduced by approximately 30%. Most of the improvement is due to the combination of the warping and bias adaptation algorithms, while a minimal improvement is due to using N-best alternate hypotheses for estimating the transformation parameters. Note that the reduction in error rate obtained by combining the warping and spectral shaping algorithms is approximately equal to the sum of the reduction in error rates when applying each of the adaptation procedures separately.

Recognition Task	Baseline	Warping
Training: Male Adult Spks Testing: Female Adult Spks	21.4%	5.9%
Training: Male Adult Spks Testing: Children Spks	52.2%	19.3%

Table 2: Word error rate for the connected digit recognition task using frequency warping normalization for mismatched training and testing conditions.

4. SPEAKER NORMALIZATION EXPERIMENTS WITH MISMATCHED DATA

The performance of the speaker adaptation algorithms was investigated for cases where there exists significant acoustic mismatch between the speaker population used during training HMMs and during recognition. Mismatch between several populations of speakers was investigated including children, adult male, and adult female speakers.

In Table 2, we display the word error rate for the connected digit recognition task over the public switched telephone network before and after frequency warping adaptation using a single utterance. In both experiments, the HMMs used for recognition are trained from a population of adult male speakers. The test set for the first row of Table 2 consists of 2800 digit strings of length one to ten (8645 digits) spoken by adult female speakers. In the second experiment, the test set consists of utterances spoken by children speakers ages six to seventeen (2500 digit strings of length one to ten, total of 9466 digits). Note that for children speakers, up to 30% expansion of the frequency scale is allowed during frequency warping. Similar children speaker adaptation experiments were reported in [6]. Despite the small amount of data used to estimate the optimal warping factor (one to ten words) and the simplicity of the

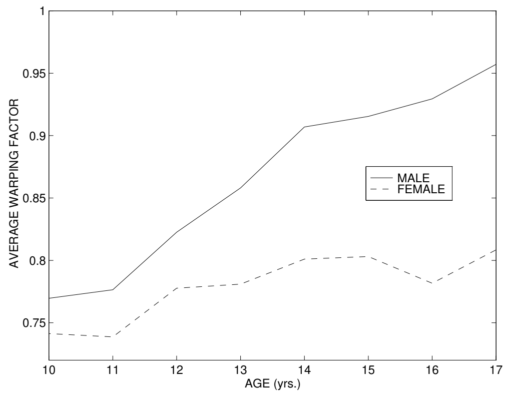

Figure 3: Average warping factors per age and gender.

transformation (linear warping) the speaker normalization algorithm can provide word error rate reduction up to 70% for mismatched training and testing speaker populations.

In Fig. 3, the average optimal warping factors $\hat{\alpha}$ are shown per speaker's age and gender. Note that $\alpha = 1$ corresponds to no warping (data matched with the adult male model), while $\alpha = 0.8$ corresponds to 20% compression of the frequency scale. The elevated slope of the average warping factor curve for young male speakers corresponds to the rapid vocal tract growth during puberty. One can further infer from this plot that the average formant frequency values for adult male speakers is approximately 15% smaller than those of adult female speakers.

5. ISSUES IN FREQUENCY WARPING

In this section, we attempt to better understand the potential performance gains that can by achieved by frequency warping normalization. It was noted in Section 3.1 that one can create a larger ensemble of observations by generating multiple "warpings" of a sequence of observations. Further, by separately decoding each of the warped observation sequences one can generate a larger ensemble of candidate decodings to choose from. In the following, we investigate the richness of alternate string hypotheses generated from the warping algorithm and evaluate the performance of the maximum likelihood criterion for selecting the decoded string corresponding to the correct transcription.

In order to characterize the richness of the ensemble of alternate string hypothesis generated by the warping algorithm, we measure how often the correct string is included in the set of candidate strings decoded from the warped observation sequences. This is similar to observing how often the correct string is included in the top-N sentence hypothesis list for an N-best decoder. In Table 3 we display the percent utterances where the correct string is contained in the list of N-top candidates when using the frequency warping speaker normalization algorithm (second column) or an N-best decoder (third column). Note that in the case of warping, N corresponds to the number of alternate warping indexes considered, not the total number of (unique) decoded strings. The databases and HMMs used for this experiment are identical to those used in Section 3.2.

It is interesting to note from Table 3 that by evaluating over an ensemble of at least four possible warping values, the string recognition rate could improve from 91% to 96%. The fact that a smaller improvement is achieved when using the procedure described in Section 2 is an indication that the criterion used for selecting the warping factor is not performing as well as we would like it to.

Ensemble Size N	Warping	N-best
1	91.0 %	91.0 %
2	94.2 %	95.8 %
3	95.6 %	97.0 %
4	96.0 %	97.8 %

Table 3: Percent of correct strings in N-top candidates.

The string recognition performance in the second column of Table 3 can be directly compared with the corresponding performance obtained using the frequency warping procedure in Section 3.2. The string recognition rate obtained using frequency warping was 93.3%. This corresponds to the digit error rate of 2.5% given in the third row of Table 1. This is significantly worse performance than the potential 96% string recognition rate (in Table 3) if the warping factor selection criterion were perfect. Furthermore, we observed that by using the correct transcription H_c in Eq. 1 to estimate the optimum warping factor the string correct rate increased from 93.3% to 94.3%. This is still far from the potential 96% string correct. Thus, errors in selecting the warping index are only partially due to poor alignment between HMM states and the observation sequence. Further investigation is necessary to improve the selection criterion.

6. SUMMARY

In this paper, we have presented evidence that the improvement in recognition performances achieved by frequency warping and spectral shaping adaptation are independent. By combining frequency warping and spectral shaping during training and testing 40% reduction in word error rate was achieved for our task. Further, the power of frequency warping was demonstrated for both matched and, especially, for mismatched training and testing speaker populations.

7. ACKNOWLEDGMENTS

The authors would like to express their sincere appreciation to the members of Furui Lab at NTT for their generous assistance and helpful advice during the early stages of this work. In particular, we would like to thank T. Matsui for helpful discussions concerning speaker adaptation.

8. 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 and R.C. Rose, "Efficient Frequency Warping Procedures for Telephone Based Speech Recognition," in Proc. *ICASSP*, 1996.

[4] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[5] T. Matsui and S. Furui,"N-best-based instantaneous speaker adaptation method for speech recognition", Proc. *ICSLP*, 1996.

[6] D. C. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in Proc. *ICSLP*, 1996.