NORMALIZATION OF SPEAKER VARIABILITY BY SPECTRUM WARPING FOR ROBUST SPEECH RECOGNITION

Y.C. Chu, Charlie Jie, Vincent Tung, Ben Lin and Richard Lee
Technology Center Philips Taiwan
P.O. Box 22978, Taipei, Taiwan, R.O.C.
Tel. +886 2 382 3207, FAX: +886 2 382 4598, E-mail: y.c.chu@tw.ccmail.philips.com

ABSTRACT

This paper examines techniques for normalization of unseen speakers in recognition. Two implementations of linear spectrum warping were examined: time domain resampling and filter bank scaling. It is shown that for seen speakers, the models trained by unwarped utterances are less sensitive to spectrum warping by filter bank scaling than by resampling. A pitch-based scheme for warping factor estimation has been proposed. The method is shown to be cost-effective in reducing the variability of unseen speakers compared to the ML-based methods. In particular the combination of filter bank scaling with the pitch-based warping factor estimation reduces the error rate of isolated Mandarin digit recognition by more than 30% for unseen speakers.

1. INTRODUCTION

Speaker variability is one of the major challenges in speaker independent speech recognition. Typically, the majority of error comes from just a few difficult speakers whose spectral characteristics are not well represented in the training population. A common approach to this problem is to collect a large amount of training data, which is time consuming and costly. When only limited training data is available, supervised or unsupervised speaker adaptation, or speaker normalization techniques must be used to achieve a satisfactory recognition rate for unseen speakers.

Vocal tract normalization by linearly warping the input utterance spectrum [1-4] was shown to be an effective way to reduce the speaker variability. Two different implementations have been proposed for linear spectrum warping. One is to resample the speech data in the time domain [1] and the other is to scale the center and bandwidth of Mel-frequency filters [2]. The former can be applied to any recognizer front end, while the latter is restricted to the filter bank types of pre-processors. Most of the previous efforts focused on using the spectrum warping techniques in both training and recognition to reduce the error rate for a given benchmark task. When used with speaker adaptation [4], spectrum warping was shown to be effective for the case where the acoustic mismatch is significant between the speaker population used in training and in recognition. However, it is unclear how to effectively apply spectrum warping alone during recognition to normalize unseen speakers having a significant acoustic mismatch with the training population. Two issues of particular interests are which implementation is best suited for such a purpose and how to determine the warping factor at a minimum computational cost. This paper reports an experimental investigation of the above two aspects. The next section describes the results of gender mismatched experiments for the comparison of different warping implementations. Next, a pitch-based scheme is presented in section 3 for optimal selection of the warping factor in a cost-effective way. The results using the pitch-based estimation are compared with those using the maximum likelihood (ML) methods [1-3]. It is shown that the pitch-based method gives a recognition rate comparable to the ML-based methods but at a much lower computational cost.

2. COMPARISON OF DIFFERENT WARPING IMPLEMENTATIONS

2.1. Frond-end Warping Implementation

The features used in our speech recognition experiment are 12 Mel-frequency Cepstrum coefficients (MFCC) and 12 delta coefficients. A set of 18 Mel-scale filters is used in the spectral analysis front end. When resampling is used to warp the speech spectrum, the input utterance is interpolated according to the given warping factor (W_t) before passed to the feature extraction front end. Thus the warping process is completely independent of the feature extraction module.

When filter bank scaling is used for spectrum warping, the center and bandwidth of each filter are multiplied by the warping factor ($W_b=1/W_t$). Note that in this type of implementation, the center frequency of the highest filter may be scaled beyond the signal bandwidth. In this case, we simply replace the output of the highest filter with that of the next highest one [3]. As will be seen later, such a process has a profound effect on the recognition results.

2.2. Gender Mismatched Experiment

A multi-speaker (50 males and 50 females) isolated Mandarin digit database was used in the recognition experiment. Each speaker uttered one utterance per digit. The speech signal was sampled at 10 kHz with a 32 ms frame size and 16 ms shift. First, the male utterances were used in training while both male and female utterances were used in recognition. The acoustic models used in the

experiment are whole word mixture Gaussian HMMs. The resultant recognition rates for different warping factors are shown in Fig. 1.

Although the models were trained from unwarped male utterances, the male recognition rate remains nearly the same for ±5% warping, Beyond this range, the male recognition rate decreases rapidly for the resampling case while much slower for filter bank scaling. Spectrum warping by resampling will affect the utterance speaking rate. For warping greater than 10 %, the change in the speaking rate becomes noticeable. This may explain the rapid decrease in the male recognition rate. Note that for the case of filter bank scaling, there is a consistent jump in recognition rates for different populations at W_b =1.15. Above this warping factor, the center of the highest filter is scaled beyond the signal bandwidth and the output of the next highest filter is used to replace that of the highest one. It seems that such replacement causes a significant increase in the recognition rate. For unseen speakers (females), filter bank scaling outperforms resampling in the expected warping range ($0.8<W_t<1$; $1<W_b<1.25$), while the maximum overall recognition rate seems to be similar between these two.

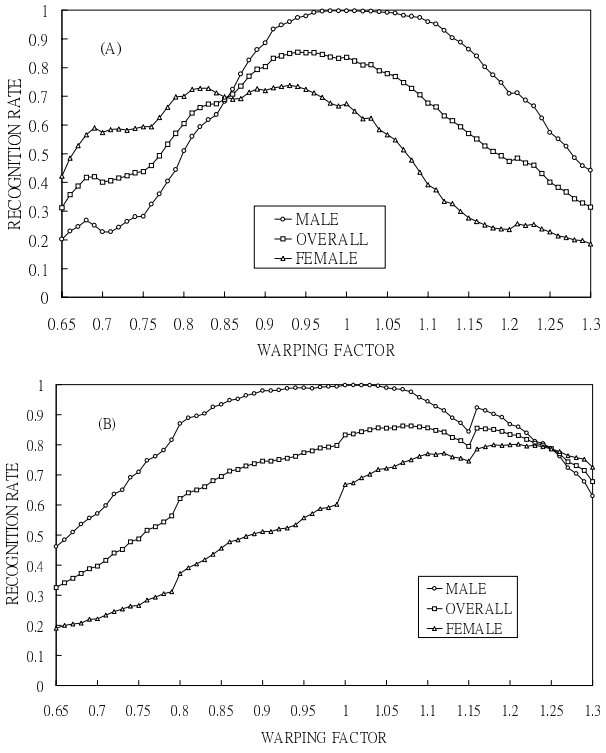

Figure 1: Recognition rate as a function of the warping factor with acoustic models trained by male utterances. The speech data is warped by (a) resampling and (b) filter bank scaling.

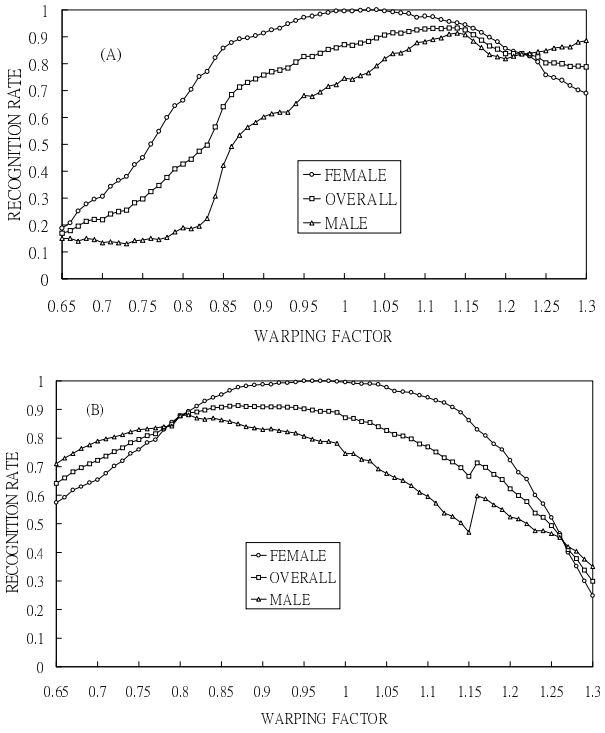

Figure 2: Recognition rate as a function of the warping factor with acoustic models trained by female utterances. The speech data is warped by (a) resampling and (b) filter bank scaling.

The recognition results using models trained by female utterances are shown in Fig. 2. Again for seen speakers, the models are less sensitive to spectrum warping by filter bank scaling than by resampling. For unseen speakers, the results by filter bank scaling tend to be stable (around 80 percent) in the expected warping range, while the results by resampling reach a peak value of 91.4% at W_t =1.14. The jump found in Fig. 1 by the filter output replacement is clearly seen for the male recognition rate while not noticeable for the female recognition rate. Table 1 summarizes the maximum recognition rates for different speaker populations and their corresponding warping factors.

Table 1: The maximum recognition rates for different speaker populations. The numbers within the parentheses denote the corresponding warping factors.

(a) acoustic models trained by male utterances

	Male	Female	Overall
Resampling	0.998 (1)	0.738 (0.93)	0.854 (0.94)
Filter scaling	0.998 (1)	0.802 (1.21)	0.863 (1.08)

(b) acoustic models trained by female utterances

	Male	Female	Overall
Resampling	0.914 (1.14)	1.0 (1.03)	0.933 (1.14)
Filter scaling	0.882 (0.81)	1.0 (0.98)	0.914 (0.87)

3. A PITCH-BASED SCHEME FOR WARPING FACTOR ESTIMATION

Estimation of the warping factor is the most critical issue for the implementation of spectrum warping. In the ML method [1], the warping factor is estimated by searching over a number of possible factors and choosing the one that maximizes the data likelihood to the acoustic models. The method is sound but requires extensive computation. A simplified ML approach [2,3] was suggested to reduce the computation complexity. In the simplified method, one first decodes the unwarped utterance and then uses the transcription to find the warping factor according the ML criterion. However, if the unseen speakers have a large acoustic mismatch with the training population, the transcriptions of unwarped utterances may not be reliable and nor is the warping factor estimated based on them. A more direct and cost-effective way is to estimate the warping factor based on the spectral characteristics of the testing utterance and the training population. A formant-based method, using the formant ratio directly as the warping factor, has been suggested [5,6]. However, the method does not guarantee a better recognition rate for difficult speakers [6].

Our experimental results show that for most utterances there is a desired warping range within which the warped utterance can be correctly recognized. Thus, the task can be seen as finding a link function that maps a given spectral feature, such as pitch or formant, to a factor within the desired warping range. In this study, we choose the pitch frequency as the spectral feature because it can be more easily and reliably estimated than the formant. A simple autocorrelation pitch detector [7] was used for pitch estimation. As an example, Fig. 3(a) shows the histograms for the upper and lower bounds of the desired warping range by resampling for female utterances when the models are trained by male utterances. Also shown in the figure is the histogram of the pitch ratio R (the average pitch frequency of the input utterance divided by the average pitch frequency of the training population) for female utterances. Fig. 3(b) shows the approximate Gaussian distribution of those shown in Fig. 3(a). The purpose of the link function is to map the pitch ratio distribution to somewhere between the distributions of the lower and upper bounds.

An appropriate link function must meet three conditions: mapping the pitch ratio to the desired warping range, keeping the warping direction suggested by the pitch ratio, and being tolerable to error in pitch estimation. The link function can be as simple as

$$W_t, 1/W_b = (1+R)/2; \quad L < W_t, 1/W_b < U \qquad (1)$$

or more complicated like

$$W_t, 1/W_b = R^k; \quad 0 < k < 1; L < W_t, 1/W_b < U \qquad (2)$$

where L and U denote the lower and upper thresholds for the estimated warping factor.

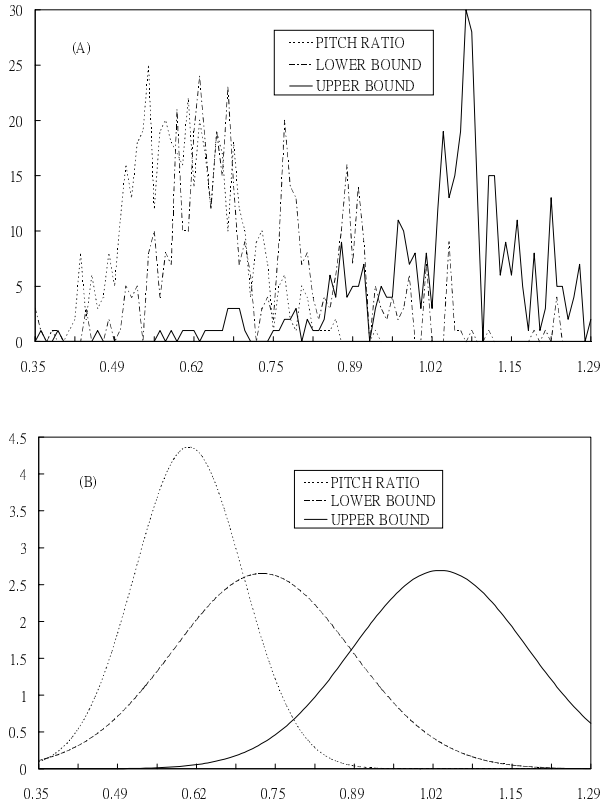

Figure 3: (a) histograms and (b) distributions of the lower and upper bounds of the desired warping range and the pitch ratios for female utterances when the models are trained by male utterances.

Table 2: Comparison of recognition rates for different link functions and warping methods.

(a) acoustic models trained by male utterances

	Resampling		Filter scaling	
	$(1+R)/2$	$\sqrt{R}$	$(1+R)/2$	$\sqrt{R}$
Male	0.962	0.962	0.976	0.974
Female	0.744	0.73	0.784	0.786
Overall	0.853	0.846	0.88	0.88

(b) acoustic models trained by female utterances

	Resampling		Filter scaling	
	$(1+R)/2$	$\sqrt{R}$	$(1+R)/2$	$\sqrt{R}$
Male	0.848	0.844	0.864	0.864
Female	0.982	0.982	0.988	0.988
Overall	0.915	0.913	0.926	0.926

When the estimate exceeds the threshold, the threshold value will be used as the warping factor. This is necessary to avoid error in pitch estimation. Typically ±20% is a reasonable warping range as seen from Fig. 1. The key issue is to find the best combination of the link function and the warping method. Table 2 compares the recognition rates with various link functions and warping methods. The thresholds for the experiment were set to be

20% toward the expected warping direction for the unseen speakers while 5% in the opposite direction. The results show that filter bank scaling is consistently better than resampling for different populations. The difference between two link functions, however, is very small. We chose *(1+R)/2* as the link function in our subsequent experiment for its simplicity.

4. COMPARISON OF PITCH-BASED AND ML-BASED APPROACHES

To evaluate the proposed pitch-based estimation scheme, we compared the recognition rates obtained using the present approach with those by the ML and simplified ML methods. The filter bank scaling method is used in the experiment for spectrum warping. The results are shown in Table 3 where the computational cost of each method normalized with respect to the baseline is also given. All three methods give a considerable gain in the overall recognition rate over the baseline. The ML method yields the best results with the highest computational cost. For seen speakers, both ML-based methods give a better recognition rate compared to the baseline, while the results by the pitch-based method are lower than the baseline. This is expected since some of warping factors estimated by the pitch-based method are out of the tolerable warping range of the models, thus creating a large acoustic mismatch between the warped utterance and the models. For unseen speakers, the results by the pitch-based method are comparable to those by the ML method, while the simplified ML method gives a lower recognition rate. As discussed in previous section, the simplified ML method uses the transcription of the unwarped utterance to find the warping factor. Due to the significant acoustic mismatch between the unseen speakers and the training population, the transcription of the unwarped utterance may be corrupted and the error affects consequently the warping factor estimation.

Table 3: Comparison of recognition rates by pitch-based and ML-based approaches for warping factor estimation. The bottom row gives the computational cost of each method.

(a) acoustic models trained by male utterances

	Baseline	Simplified ML	Pitch	ML
Male	0.998	1.0	0.976	1.0
Female	0.668	0.736	0.784	0.798
Overall	0.833	0.868	0.88	0.899
Cost	1.0	11.7	<1.1	51

(b) acoustic models trained by female utterances

	Baseline	Simplified ML	Pitch	ML
Male	0.746	0.798	0.864	0.868
Female	0.996	0.998	0.988	1.0
Overall	0.871	0.898	0.926	0.934
Cost	1.0	11.7	<1.1	51

The computational cost of each method normalized with respect to the baseline is given in the last raw of the table. The pitch-based method adds only a marginal cost for the pitch estimation compared to the baseline system. It is clear that the pitch-based method gives an overall recognition rate comparable to those given by the ML-based methods but at a significantly lower cost.

5. CONCLUSION

In this paper, we have investigated the use of linear spectrum warping for normalization of unseen speakers during recognition. Two warping implementations have been examined and the results show that spectrum warping by filter bank scaling for seen speakers is better tolerated by acoustic models trained from unwarped utterances. A pitch-based scheme for warping factor estimation has been proposed. The method is shown to be cost-effective in reducing the variability of unseen speakers. In particular, the combination of filter bank scaling with the pitch-based warping factor estimation reduces the error rate of isolated Mandarin digit recognition by more than 30% for unseen speakers at a minimum computational cost.

REFERENCES

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

[2] L. Lee and R.C. Rose, "*Speaker Normalization Using Efficient Frequency Warping Procedures*," Proc. ICASSP'96, pp. 353-356, Atlanta, 1996.

[3] D. Pye and P.C. Woodland, "*Experiments in Speaker Normalization and Adaptation for Large Vocabulary Speech Recognition*," Proc. ICASSP'97, pp. 1047-1050, Munich, 1997.

[4] A. Potamianos and R.C. Rose, "*On Combining Frequency warping and Spectral Shaping in HMM Based Speech Recognition*," Proc. ICASSP'97, pp. 1275-1278, Munich, 1997.

[5] E. Eide and H. Gish, "*A Parametric Approach to Vocal Tract Length Normalization*," Proc. ICASSP'96, pp. 346-348, Atlanta, 1996.

[6] P. Zhan and M. Westphal, "*Speaker Normalization Based on Frequency Warping*," Proc. ICASSP'97, pp. 1039-1042, Munich, 1997.

[7] L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals," Prentice Hall, Englewood Cliffs, NJ, 1978.