AN INVESTIGATION INTO FRONT-END SIGNAL PROCESSING FOR SPEAKER NORMALIZATION

S. Umesh and Rohit Sinha

Department of Electrical Engineering
Indian Institute of Technology
Kanpur - 208 016, INDIA
{sumesh, srohit}@iitk.ac.in

S. V. Bharath Kumar

Imaging Technology Lab
General Electric - Global Research
JFWTC, Bangalore - 560 086, INDIA
bharath.sv@geind.ge.com

ABSTRACT

Our investigation into the front-end signal processing for maximum likelihood based speaker normalization reveals that in the linear scaling model, it is more appropriate (and evidently more correct) to assume that the *spectral envelopes* of any two speakers for same sound are linearly scaled versions of one and another, rather than assuming that the whole magnitude spectra (including pitch harmonics) are scaled. The use of the proposed model and its implementation results in about 4% and 7% relative improvement for adults and children respectively on a digit recognition task.

1. INTRODUCTION

Andreou *et. al.* [1] proposed a maximum likelihood (ML) based speaker normalization procedure to extract and use acoustic features that are robust to variations in vocal tract length and, unlike earlier methods, it did not require estimation of formant frequencies. However, their method involved resampling of the speech data while doing the grid search to estimate ML warp factors. Later Lee and Rose [2] extended Andreou *et. al.*'s procedure by efficiently incorporating the linear warping into the front-end computation of Mel frequency cepstral coefficient (MFCC) feature by scaling the center frequency and bandwidth of the filters in filter bank instead of resampling the data.

Recently, we have shown that the linear speaker normalization can alternatively be performed through ML shifting in Log-warped spectral domain [3]. On comparing the performance of the conventional ML linear scaling based speaker normalization procedures with our proposed shift-based method, we noticed significant difference in performance particularly for children on a connected digit recognition task.

In this paper, we present a study undertaken to understand the reason for the difference in performance of two theoretically equivalent speaker normalization methods and it indicates that in linear scaling model for speaker normalization it is more appropriate to assume that only *spectral envelope* are scaled versions of one and another, rather than the *complete magnitude spectrum* (including pitch harmonics) as is usually assumed in conventional speaker normalization methods [1, 2].

2. EXPERIMENTAL SETUP

The recognition experiments are performed on a telephone based connected digit recognition task. The speech data for training the recognizer is derived from the Numbers corpus v1.0cd of OGI. The training set consist of 6078 utterances from adult male and female speakers. Two test sets are used: *matched* test set which is derived from Numbers corpus and consists of 2169 utterances from adult male and female speakers and *mismatched* test set consisting of 2798 utterances from speaker having age between 6 to 18 years. Through out this paper word error rate is used to evaluate the performance of the different methods.

The digit recognizer was developed using HTK HMM Toolkit. The digits are modeled as whole word simple left-to-right HMMs without skips and have 16 states per word with 5 diagonal covariance Gaussian mixtures per state. The silence is modeled using 3 state HMM model having 6 Gaussian models per state. A single state short pause model tied to middle state of silence model is also used. The feature vector comprising normalized energy, C_1 to C_{12} static cepstra and their first and second order derivatives is used and cepstral mean subtraction is also performed.

2.1. Recognition Performance

We briefly review our shift-based approach to speaker normalization. In this approach, we assume that the spectral envelope (and not magnitude spectrum) of any two speakers A and B are related by $S_A(f) = S_B(\alpha_{AB} f)$. If we Log-warp the frequency axis, $\lambda = \ln f$, then in the warped domain we have $s_a(\lambda) = s_b(\lambda + \ln \alpha_{AB})$. So the normalization can be performed by ML estimation of shift fac-

Condition	Shift based		Warp based (conventional MFCC)	
	Adults	Children	Adults	Children
Baseline	3.18	14.01	3.42	15.35
Norm.	2.67	9.37	2.70	9.52

Table 1. Word error rate before and after applying shift and warp based linear speaker normalization methods.

Condition	WOSA-MFCC based	
	Adults	Children
Baseline	3.26	14.26
Norm.	2.59	8.59

Table 2. Performance of warp based speaker normalization using WOSA-MFCC feature. This may be compared with that of conventional MFCC feature shown in Table 1.

tor, $\ln \alpha_{AB}$, by doing a grid search. The spectral envelopes (s_a, s_b) are obtained by the weighted overlapped spectral averaging (WOSA) [4]. The details of the front-ends used for the implementation of conventional filter bank based linear speaker normalization approach and our proposed shift based speaker normalization approach are described in [3]. The parameters of the front-end signal processing of two methods have been chosen such that the up and down shift by 3 in the shift based method correspond to linear frequency warping by 7 warping factors chosen between range 0.88-1.12 in steps of 0.04 which is same as that used in conventional warp based normalization method. In shift based method, the subframe length and overlap are chosen to be 80 and 60 samples respectively in WOSA procedure [3].

The performances of the two normalization methods are shown in Table 1. It can be noticed that the shift based normalization method provides better baseline and after normalization performances compared to warp based method. This improvement is particularly significant for children considering the fact that the front-end of shift based method uses *Log-warping* whereas that of warp based method involves *Mel-warping* which is known to provide improvement for children [5].

While in the above experiment, the two methods are broadly equivalent in theory, there are certain differences in their implementation. Therefore, in this paper, we have tried to match the implementation of two methods as close as possible so that a more definitive conclusion can be made. We now describe the details of a signal processing front-end that we have used to match the implementations.

3. NORMALIZATION USING WOSA-MFCC FEATURE

We computed the conventional MFCC features using WOSA spectral smoothing procedure instead of using filter bank smoothing and the resulting feature is referred to as *WOSA-MFCC feature* in this work.

In the computation of WOSA-MFCC feature, the given frame of speech is smoothed using WOSA procedure [3] and the resulting smoothed auto-correlation estimates are then converted to Mel warped spectrum through non-uniform DFT which is computed on the frequencies that are same as the center frequencies of Mel-scaled filter bank used in case of un-warped MFCC feature computation. Finally similar to MFCC feature computation, the above computed Mel-spaced spectrum is log compressed and converted to cepstral coefficients using DCT. Note that there is no explicit filter bank in this method, but a filtering interpretation can be given as discussed in Section 3.1. Thus WOSA-MFCC and conventional MFCC front-ends are same in all respects except for the spectral smoothing procedure used.

3.0.1. Computation of Warped WOSA-MFCC Feature during normalization

For implementing the warping of the WOSA-MFCC feature during normalization, the points on which Mel-warped spectra is computed (using non-uniform DFT) are scaled by appropriate values of scale factor before being converted to cepstral coefficients.

3.0.2. Recognition Performance

Table 2 shows the performance of linear speaker normalization method using WOSA-MFCC feature. On comparing the performance of normalization method using WOSA-MFCC feature with that of normalization method using conventional MFCC feature, we notice that it has provided about 4% reduction in word error rate for adults and a *significant* reduction of 10% in word error rate for children.

3.1. Filtering Interpretation of WOSA Procedure

In order to clearly understand the effect of spectral smoothing procedure on the performance, the WOSA smoothing procedure is interpreted as a filtering operation.

Since WOSA smoothing procedure is a variant of averaged periodogram spectral estimation method so similar to averaged periodogram method it can also be given a filtering implementation. From Nuttall and Carter [4], in WOSA method, the relationship between power spectra, $\hat{G}_{av}(f)$, computed by Fourier transform of averaged auto-correlation estimates and the true power power spectra, $G(f)$, of the speech frame can be expressed as follows,

$$\hat{G}_{av}(f) = G(f) \otimes \{\mathcal{F}[w(t)]\}^2 \quad (1)$$

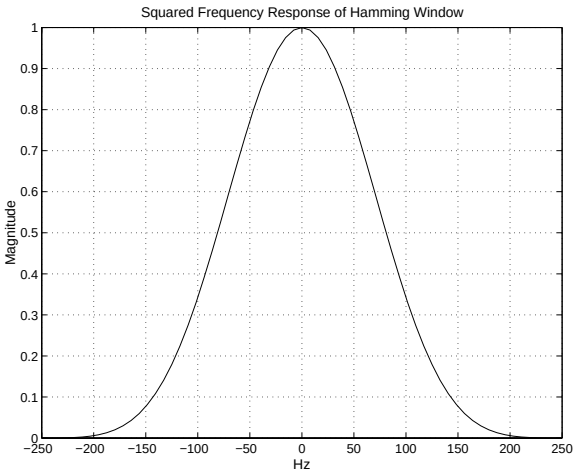

Fig. 1. Plot of normalized magnitude squared Fourier transform of Hamming window of length 80 samples.

where '$\otimes$' denotes convolution and $\mathcal{F}[w(t)]$ denotes Fourier transform of (Hamming) window used on each sub-frame.

So we can argue that, in WOSA procedure, the smooth power spectrum estimate at any frequency is obtained by a bandpass filtering the true power spectrum at that frequency. The bandpass filter has frequency response equal to that of square of Fourier transform of Hamming window as shown in Fig. 1 and is called as "WOSA-filter" in this work. For parameters chosen for WOSA processing, the bandwidth of WOSA-filter is approximately $250Hz$.

Since in WOSA-MFCC feature, we have computed the power spectra at frequencies spaced on Mel scale so it can be argued that WOSA based feature also uses a filter bank, similar to Mel filter bank used in MFCC feature computation, except its constituent filters are of *uniform bandwidth* and have frequency response equal to that of WOSA-filter as shown in Fig. 1. In Fig. 2, we show the *explicit* triangular Mel filter bank mask used in conventional MFCC feature computation and the *implicit* filter bank argued in WOSA-MFCC feature. Note that the bandwidth of filters and in turn the resulting spectral smoothing increases with increasing frequency in conventional MFCC filterbank unlike in WOSA-MFCC case.

Further, during normalization, the bandwidth of the filters along with center frequencies are scaled in the conventional speaker normalization implementation [2]. It can easily be shown that the bandwidth scaling is necessary, if we assume that the entire magnitude spectrum (including pitch harmonics) is scaled and is a more efficient implementation than resampling the speech data. However, if we assume only the spectral envelopes are scaled, then the scaling of the bandwidths are not necessary (only the center frequencies have to be moved appropriately) as done in WOSA-MFCC based normalization method. We will discuss these issues in greater detail in the next section.

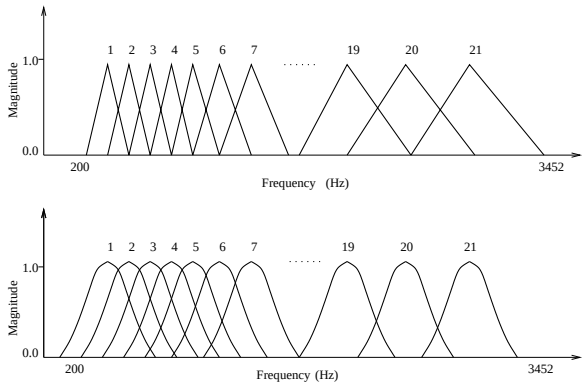

Fig. 2. Illustration of the explicit filter bank used in conventional MFCC feature computation (shown on top of figure) and the implicit filter bank argued in WOSA-MFCC feature computation (shown in the bottom of figure).

Condition	MFCC(magnitude)		MFCC(power)	
	Adults	Children	Adults	Children
Baseline	3.42	15.35	3.19	15.22
Norm.	2.70	9.52	2.62	9.18

Table 3. Performance of warp based speaker normalization methods using MFCC feature derived using magnitude and power spectrum.

4. STUDY INTO SIGNAL PROCESSING OF NORMALIZATION METHODS

There are three differences in our implementation of conventional MFCC and WOSA-MFCC feature based normalization methods. These are (1) use of power versus magnitude spectrum (a difference due to our implementation), (2) use of uniform bandwidth filter bank versus constant-Q filter bank and (3) scaling and non-scaling the bandwidth of filters of filter bank during linear warping.

We now investigate each of them by modifying the conventional MFCC to accommodate each of these differences.

4.0.1. Power Vs Magnitude Spectrum

In our implementation, we have computed MFCC feature using magnitude spectrum whereas WOSA-MFCC feature computation involves use of power spectrum so this difference needed to be studied.

Using power spectrum instead of magnitude spectrum in computation of MFCC feature does seem to improve the performance for all cases (i.e., adults or children, baseline or after normalization) as can be seen from Table 3.

Condition	Conventional MFCC (constant-Q FB) {BW scaled}		Modified MFCC (uniform-BW FB) {BW not scaled}	
	Adults	Children	Adults	Children
Baseline	3.19	15.22	3.19	14.43
Norm.	2.62	9.18	2.59	8.49

Table 4. Performance of speaker normalization methods using conventional MFCC feature derived involving constant-Q filter bank and modified MFCC feature derived using uniform filter bank, both methods used power spectra.

4.0.2. Uniform Vs Constant-Q Bandwidth Filter Bank

In this subsection, we derive the uniform bandwidth filter bank based MFCC feature by modifying the constant-Q Mel filter bank used in conventional MFCC feature such that the constituent filters are now have a constant bandwidth of 250 Hz (equal to WOSA-filter). Further, these filters are not scaled during warping; only center frequencies are scaled.

The recognition performance for linear normalization methods using conventional and modified uniform bandwidth filter bank based MFCC features are shown in Table 4. We observed 5 % reduction in word error rate for baseline and 7.5 % reduction in word error rate after normalization for modified MFCC feature case compared to conventional MFCC case particularly for children. Although, for adults, no significant change was observed in baseline and after normalization performances.

The improvement in the performance for children can be attributed to following: (1) Less smoothing effected by uniform bandwidth filter bank compared to constant-Q filter bank particularly at high frequencies which may help preserve the formant structure. (2) Non-scaling of filter bandwidths during warping in case of uniform bandwidth filter bank case seems to avoid the over-smoothing possible in the conventional constant-Q filter bank case. Since the above mentioned possible over-smoothing would be critical only at higher frequency region so this could be the reason why performances for adult did not show any significant change.

4.0.3. Bandwidth Scaled Vs Not Scaled during Warping

In this subsection, we study the effect of scaling the filter bandwidths during warping in two normalization methods.

Table 5 shows that just *non-scaling* of bandwidth results in about 7 % *drop* in the word error rate in conventional MFCC based method for children whereas in modified MFCC based method, the *scaling* of the bandwidth during warping results in about 7 % *increase* in the word error rate for children.

Therefore, we conclude that the bandwidth of the filter should not be scaled during normalization. This, of course, can be mathematically justified only if we assume that the spectral envelopes are scaled and not if the entire magnitude spectrum is assumed to be scaled.

Condition	Conventional MFCC {BW not scaled}		Modified MFCC {BW scaled}	
	Adults	Children	Adults	Children
Baseline	3.19	15.22	3.19	14.43
Norm.	2.68	8.54	2.53	9.18

Table 5. Performance of speaker normalization methods using modified MFCC feature with filter bandwidth scaling and conventional MFCC feature with no filter bandwidth scaling during warping, both methods used power spectra.

5. CONCLUSION

In this work, we present a study into linear scaling model assumed in the widely used ML based speaker normalization method which indicates that it is more appropriate (and obviously more correct) to assume that the spectral envelope of any two speakers are scaled version of one another rather than whole magnitude spectrum including pitch harmonics.

The motivation to the proposed modification is provided by our recently proposed shift based speaker normalization approach and the proposed modification results in 4% and 10% relative improvement in the normalization performance for adults and children respectively. We also show that simply non-scaling of filter bandwidth in conventional Mel filter bank based method results in about 7% relative improvement in the performance for children without affecting the performance for adults.

6. 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] Li Lee and Richard Rose, "Frequency Warping Approach to Speaker Normalization," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, pp. 49–59, Jan. 1998.

[3] Rohit Sinha and S. Umesh, "Non-Uniform Scaling Based Speaker Normalization," in *Proc. of IEEE ICASSP'02*, May 2002, vol. 1, pp. 589–592.

[4] A. H. Nuttall and G. C. Carter, "Spectral Estimation using Combined Time and Lag Weighting," *Proceedings of the IEEE*, vol. 70, pp. 1115–1125, Sept. 1982.

[5] J. G. Wilpon and C. N. Jacobsen, "A Study of Speech Recognition for Children and the Elderly," in *Proc. ICASSP'96*, May 1996, vol. 1, pp. 349–352.