Implementing Frequency-Warping and VTLN Through Linear Transformation of Conventional MFCC

S. Umesh

Department of Electrical Engineering
Indian Institute of Technology, Kanpur, India
sumesh@iitk.ac.in

A. Zolnay and H. Ney

Lehrstuhl für Informatik VI, Computer Sc. Dept.
RWTH-Aachen Univ., 52056 Aachen, Germany
{zolnay, ney}@informatik.rwth-aachen.de

Abstract

In this paper, we show that frequency-warping (including VTLN) can be implemented through linear transformation of conventional MFCC. Unlike the Pitz-Ney [1] continuous domain approach, we directly determine the relation between frequency-warping and the linear-transformation in the discrete-domain. The advantage of such an approach is that it can be applied to *any* frequency-warping and is not limited to cases where an analytical closed-form solution can be found. The proposed method exploits the bandlimited interpolation idea (in the frequency-domain) to do the necessary frequency-warping and yields exact results as long as the cepstral coefficients are que-frency limited. This idea of quefrency-limitedness shows the importance of the filter-bank smoothing of the spectra which has been ignored in [1, 2]. Furthermore, unlike [1], since we operate in the discrete domain, we can also apply the usual discrete-cosine transform (i.e. DCT-II) on the logarithm of the filter-bank output to get *conventional* MFCC features. Therefore, using our proposed method, we can linearly transform conventional MFCC cepstra to do VTLN and we do not require any recomputation of the warped-features. We provide experimental results in support of this approach.

1. Introduction

Vocal tract length normalization (VTLN) is an important approach to reduce inter-speaker variability in speaker-independent speech recognition. One of the most common approaches to VTLN involves appropriately warping the frequency axis and then obtaining the corresponding normalized features after application of mel-warping (motivated by psychoacoustic arguments) and DCT which approximates decorrelation of the features. The parameters of the warping function are usually estimated using a maximum-likelihood (ML) criterion. Since the ML estimation of warp-factor usually involves a grid-search, the features need to be recomputed as many times and this is expensive. If we can directly transform the original non-VTLN features for different warp-factors then this would be much more computationally efficient. Furthermore, the knowledge of the linear transformation would help us to compensate for the Jacobian of the linear transformation during warp factor estimation. In [3], the linear-transform for each warp-factor is itself estimated using a maximum-likelihood criterion. In [4], a different approach to frequency-warping is presented; and in this approach too, we can express the frequency-warping as a linear-transformation of the original features. However, the focus of this paper is on the approach proposed by Pitz and Ney [1], where they present a method for analytical computation of the linear-transformation in the *continuous* frequency domain.

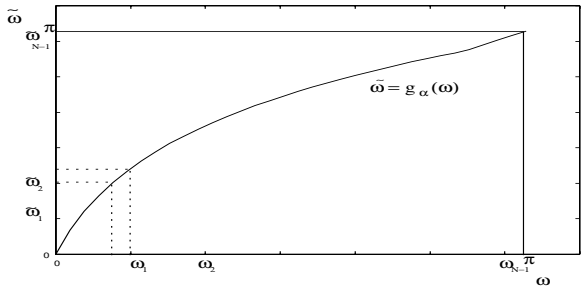

Figure 1: Figure shows that in discrete implementation, equally spaced frequencies in warped domain, often do not correspond to any of the equally-spaced frequencies in un-warped domain.

We first briefly review their method.

The cepstral coefficients c_k, $k = 0, \ldots, (N-1)$ of a spectrum $X(\omega)$ are defined by:

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega e^{i\omega k} \ln |X(\omega)|^2 = \frac{s_k}{\pi} \int_{0}^{\pi} d\omega \cos(\omega k) \ln |X(\omega)|^2 , \quad (1)$$

where s_k is appropriate scaling of either $\frac{1}{2}$ or 1.

The transformation for spectral warping is defined as $\omega \rightarrow \tilde{\omega} = g_\alpha(\omega)$, where $g_\alpha : [0, \pi] \rightarrow [0, \pi]$ and is assumed to be invertible. Since g_α is invertible, the value of the warped spectrum at any warped-frequency $\tilde{\omega}$ can be found from the unwarped spectrum, i.e. $\left|\tilde{X}_\alpha(\tilde{\omega})\right| := \left|X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|$.

Hence, the $n-$th cepstral coefficient $\tilde{c}_n(\alpha)$, $n = 0, \ldots, (N-1)$ of the *warped* spectrum is given by

$$\tilde{c}_n(\alpha) = \frac{s_k}{\pi} \int_{0}^{\pi} d\tilde{\omega} \ln \left|X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|^2 \cos(\tilde{\omega} n) . \quad (2)$$

The spectrum $\ln \left|X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|^2$ is expanded in terms of the *unwarped* cepstrum, c_k, using the inverse of the relation of Eq. 1 and inserted into Eq. (2), to obtain $\tilde{c}_n(\alpha) = \sum_{k=0}^{K} A_{nk}(\alpha)\ c_k$ with

$$A_{nk}(\alpha) = \frac{2s_k}{\pi} \int_{0}^{\pi} d\tilde{\omega}\ \cos(\tilde{\omega} n) \cos(g_\alpha^{(-1)}(\tilde{\omega}) k) . \quad (3)$$

However, when operating in the discrete domain, the relation between discrete-warped spectra and discrete-original spectra is not so straightforward and can only be made by making some additional assumption on the que-frency property of the cepstra.

In discrete implementation, the value of the discrete unwarped-spectra, $\ln|X(\omega)|^2$, is known only at a set of equally spaced frequencies $\omega_q = \frac{2\pi q}{N}$. In order to compute the *warped* cepstrum, $\tilde{c}_n(\alpha)$, we need to know the value of warped spectra, $\ln\left|\tilde{X}_\alpha(\tilde{\omega})\right|^2$, at equally spaced set of frequencies, $\tilde{\omega}_l = \frac{2\pi l}{N}$. As seen in Fig. 1, $\tilde{\omega}_l$ may not correspond to any of the known discrete un-warped frequencies, i.e. $g_\alpha^{(-1)}(\tilde{\omega}_l)$ may not correspond to any of the discrete frequencies ω_q. It is, therefore, important to understand the conditions under which we can we exactly recover $\ln\left|\tilde{X}_\alpha(\tilde{\omega})\right|^2$ with the knowledge of $\ln|X(\omega)|^2$, at equally spaced frequencies $\omega_q = \frac{2\pi j}{N}$. There is an approach suggested in [2], which uses un-warped $\ln|X(\omega)|^2$ to directly obtain the *warped* spectrum, but this is basically non-uniform DFT and is only an approximation.

If $\ln|X(\omega)|^2$ and c_k are thought of as a discrete-time Fourier transform (DTFT) pair, then sampling of $\ln|X(\omega)|^2$ would result in periodic repetition of c_k. As long as c_k is strictly que-frency limited and the sampling rate is sufficiently high, then there is no aliasing in the cepstral domain. In such a case, the value of $\ln|X(\omega)|^2$ at any frequency (including $g_\alpha^{(-1)}(\tilde{\omega}_l)$) can be found from its discrete samples through bandlimited interpolation. This is basically exploiting the sampling theorem, where a signal (in this case frequency-domain signal) can be reconstructed from its samples by using sinc-interpolation. This relationship can be seen by inserting Eq. 4 into Eq.7.

With this motivation, we address the following issues:

- In the discrete case, with proper assumptions, linear transformation matrices can be found between original cepstra and the warped cepstra *without* the need for analytical calculation of the matrix.

- In [1], the front-end signal processing is based on the work in [2], and the cepstra is obtained without any smoothing of the spectra. Therefore, in case of voiced frames, the unwarped plain cepstra would have a periodic high que-frency component which may violate the assumption of que-frency limitedness in the discrete case. And further, in the case of mel-warping (or any non-linear VTLN warping), the pitch harmonics are no longer equally spaced and may manifest in "broad que-frency" components. Therefore, to implement frequency-warping through linear transformation of cepstra, we show that spectral smoothing is necessary. This can be in the form of filter-bank prior to the computation of the cepstra and would serve an "anti-aliasing" function.

- Note that, in [1, 2], the cepstral coefficients are obtained by using the inverse Fourier transform (IFT). The fact that $\ln|X(\omega)|^2$ is a real and even function reduces the IFT to be expressed in terms of the cosine basis functions. However, this is not the usual discrete-cosine transform (DCT), even when operating in the discrete domain, because of the crucial half-sample shift. Normally, in speech recognition, the DCT (with half-sample shift) is used due to their approximate de-correlating properties. In this paper, we use this conventional DCT (with half-sample shift) and show a direct linear relation between conventional MFCC and VTLN-warped MFCC.

2. Linear Transformation of plain Cepstra

In this section, we present the discrete implementation of the linear transformation required for warping the "plain" cepstra, i.e. without any filter-bank smoothing or DCT. Given a frame of speech (with appropriate pre-emphasis and windowing), $x[m]\}_{m=0}^{N-1}$, the corresponding DFT is computed to get $X[q]\}_{q=0}^{N-1}$. The corresponding cepstra is computed by taking the logarithm of the magnitude of $X[q]$, i.e.

$$\hat{C}_k = \frac{1}{N}\sum_{q=0}^{N-1} \ln|X[q]|^2\, e^{+j\frac{2\pi}{N}qk}. \tag{4}$$

Note that $x[m]$ and the continuous spectra $X(\omega)$ (without magnitude and logarithm) form a DTFT pair. In conventional implementation, to get the warped spectra $\tilde{X}_\alpha(\tilde{\omega}_l = \frac{2\pi}{N}l)$, we can use the DTFT relation, i.e.

$$\tilde{X}_\alpha[l] = \tilde{X}_\alpha(\tilde{\omega}_l) = X\left(g_\alpha^{(-1)}(\tilde{\omega}_l)\right) = \sum_{m=0}^{N-1} x[m]e^{-jg_\alpha^{(-1)}(\tilde{\omega}_l)m}. \tag{5}$$

We then apply the magnitude and the logarithm operation on $\tilde{X}_\alpha[l]$ to get the warped cepstra as

$$\hat{C}_n(\alpha) = \frac{1}{N}\sum_{l=0}^{N-1} \ln\left|\tilde{X}_\alpha[l]\right|^2 e^{+j\frac{2\pi}{N}ln}. \tag{6}$$

Therefore, for every warp factor, we have to compute the warped spectra $\tilde{X}_\alpha[l] = \tilde{X}_\alpha(\tilde{\omega}_l)$ and then find the corresponding warped-cepstra $\hat{C}_n(\alpha)$ through an inverse DFT (IDFT). However, it would be computationally efficient if we could obtain $\hat{C}_n(\alpha)$ directly from $\hat{C}_k$. Note that knowledge of $\hat{C}_k$ implies knowledge of $\ln|X[q]|^2$ since they form a DFT pair. However, with this knowledge we cannot recover $x[m]$ or $X[q]$ because of the magnitude operation, and therefore we cannot compute $\tilde{X}_\alpha[l]$. However, since our interest is only in $\ln|\tilde{X}_\alpha[l]|^2$, if we assume that $\hat{C}_k$ is que-frency limited and unaliased, then we can exactly determine $\ln|\tilde{X}_\alpha[l]|^2$ from $\ln|X[q]|^2\}_{q=0}^{N-1}$ through sinc-interpolation. Or, we can directly determine it from $\hat{C}_k$ by

$$\begin{aligned}\ln|\tilde{X}_\alpha[l]|^2 &= \ln|\tilde{X}_\alpha(\tilde{\omega}_l)|^2 = \ln|X\left(g_\alpha^{(-1)}(\tilde{\omega}_l)\right)|^2 \\ &= \sum_{k=0}^{N-1}\hat{C}_k e^{-j\frac{2\pi}{N}g_\alpha^{(-1)}(\tilde{\omega}_l)k}\end{aligned} \tag{7}$$

Note that this equation will not be exact if there is aliasing, in which case the values will only match at $\ln|X(\omega_q)|^2$. Now, substituting Eq.7 in Eq.6, we get the linear transformation relation between $\hat{C}_n(\alpha)$ and $\hat{C}_k$, i.e.

$$\begin{aligned}\hat{C}_n(\alpha) &= \sum_{k=0}^{N-1}\hat{C}_k\frac{1}{N}\sum_{l=0}^{N-1} e^{-j\frac{2\pi}{N}g_\alpha^{(-1)}(\tilde{\omega}_l)k}e^{+j\frac{2\pi}{N}ln} \\ &= \sum_{k=0}^{N-1} A_{n,k}\hat{C}_k\end{aligned} \tag{8}$$

The above equations reveal the benefit of using the discrete approach. The basic approach is to start with equally spaced samples in the warped domain, and then map it back to corresponding discrete values in the physical frequency or ω domain. From these discrete values, the matrices can be formed as shown in Eq. 8. This is in contrast to the continuous domain

approach of [1], where we have to use the analytical formula for the warping function and then analytically solve for the linear transformation through (3).

In practice, piece-wise linear-warping associated with VTLN is normally applied first, followed by mel-warping to get the final VTL-normalized mel-warped cepstra, i.e.

$$\tilde{\omega}_{mel,\alpha} = g_{mel}(\lambda) = g_{mel}(g_\alpha(\omega)) = g_{mel,\alpha}(\omega). \quad (9)$$

As the above equation indicates, we start from equally spaced samples in the $\tilde{\omega}_{mel,\alpha}$, do inverse-mel warping, followed by inverse scaling to get the corresponding non-uniformly spaced discrete frequencies. Using these discrete frequencies, the transformation between plain cepstral coefficients and mel+linear-warped cepstral coefficients can be found as

$$\begin{aligned}\hat{C}_n(mel,\alpha) &= \sum_{k=0}^{N-1} \hat{C}_k \frac{1}{N} \sum_{l=0}^{N-1} e^{-j\frac{2\pi}{N} g_{mel}^{\alpha}{}^{(-1)}(\frac{2\pi}{N}l)k} e^{+j\frac{2\pi}{N}ln} \\ &= \sum_{k=0}^{N-1} D_{n,k}^{mel,\alpha} \hat{C}_k. \end{aligned} \quad (10)$$

Of course, what is of practical importance is the relation between mel-warped cepstra and VTLN-warped-mel-warped cepstra, i.e.

$$\hat{C}_n(mel,\alpha) = \sum_{k=0}^{N-1} D_{n,k}^{mel,\alpha} \sum_{m=0}^{N-1} F^{-1}{}_{k,m}^{mel} \hat{C}_m(mel), \quad (11)$$

where we make use of the relation $\hat{C}_k = \sum_{m=0}^{N-1} F^{-1}{}_{k,m}^{mel} \hat{C}_m(mel)$.

3. Effect of Spectral Smoothing on warped coefficients obtained by Linear Transformation

In this section, we show the importance of pre-smoothing of spectra. To begin with, we will adopt the front-end signal processing discussed in [2], where there is no smoothing of spectra and the logarithm is directly applied on the magnitude of spectra followed by inverse Fourier transform. Since there is no spectral smoothing, the corresponding cepstra is not strictly que-frency limited, especially when pitch harmonics are present. In fact, if there is some warping of the frequency-axis, then the pitch periodicity is also destroyed and the cepstra will be smeared by the effect of the broad-quefrency effect of the pitch. In Table 1, we compare the warped-cepstral coefficients obtained by our method with the coefficients obtained by actual spectral warping using (5) and (6). In this example, we have used both mel-warping and VTLN warping with a warp-factor of 0.90. For this case, a closed-form solution can be found for the method in [1] and therefore the linear-transformation matrices are identical. From Table 1, we can see that, for the un-smoothed case, the linearly transformed cepstra corresponding to mel-warping with VTLN-warping are significantly different from the values of the cepstra obtained by *directly* warping the spectra.

Next, we smoothed the spectra so that the corresponding cepstra becomes que-frency limited. We have adopted the procedure described in [5], which essentially involves smoothing the spectra with filters that have a hamming-window like shape. These filters are uniformly spaced and have uniform bandwidth in the physical frequency domain. We then applied the logarithm on the magnitude of the filter-bank output and computed the cepstra. In Table 1, we can see that, for the smoothed case, the cepstra obtained by linear-transformation and that obtained directly by spectral-warping are identical.

	Un-Smoothed		Smoothed	
Co.#	Warp	L.T.	Warp	L.T.
c0	3.277	3.242	5.054	5.054
c1	0.391	0.341	0.246	0.246
c2	0.063	-0.049	-0.078	-0.078
c3	0.188	0.183	0.519	0.519
c4	-0.026	-0.006	-0.071	-0.071
c5	-0.115	-0.102	-0.386	-0.386
c6	-0.028	-0.008	-0.087	-0.087
c7	-0.029	-0.006	0.007	0.007
c8	-0.063	-0.035	-0.082	-0.082
c9	-0.005	-0.001	-0.007	-0.007
c10	-0.011	0.009	-0.062	-0.062
c11	0.009	0.008	0.061	0.061
c12	0.008	0.002	0.017	0.017

Table 1: Warped cepstral coefficients *with* and *without* smoothing of spectra for a frame of voiced speech. As seen, when there is no spectral-smoothing, the cepstra from the linear-transformation approach differ from the cepstra obtained by actual spectral-warping. Here "Warp" indicates the use of *actual* spectral warping while "L.T." indicates cepstral coefficients obtained by linear-transformation method

In this section we have shown that, with smoothing of spectra, we can assure that the cepstra is que-frency limited and in this case, the cepstra obtained by actual spectral-warping and by the liner-transformation method are identical. In the next section, we will extend the idea of linear transformation of cepstra to the case where DCT has been applied on the logarithm of the filter-bank smoothed spectra.

4. Linear Transformation of cepstra obtained by applying DCT

In most automatic speech recognition systems, diagonal covariance matrices are used based on the assumption that the features are approximately decorrelated. This is often approximately ensured by the application of discrete-cosine transform (DCT) while computing the cepstral coefficients.

Therefore, in this section, the cepstra (without mel-warping) is obtained by applying the DCT on the logarithm of the filter-bank smoothed spectra, $\ln |X_{FB}[q]|^2$, i.e. :

$$d_k = \sum_{q=0}^{M-1} \ln |X_{FB}[q]|^2 \cos\left(\frac{(2q+1)\,k\,\pi}{2M}\right), \quad (12)$$

where M is the number of outputs of filter-bank. Similarly, we can write the "plain" cepstra of *filter-bank* output (with abuse of notation) as

$$\hat{C}_k = \frac{1}{2(M-1)} \sum_{q=0}^{M-1} \beta_q \ln |X_{FB}[q]|^2 \cos(\frac{\pi}{M-1} qk), \quad (13)$$

where $\beta_q = 2$ for $q \neq 0, (M-1)$. This relation will be clear if we look at the analogy with Eq.4 and set $M = \frac{N}{2} + 1$. Using the above two equations, we can write the linear transformation relation between d_k and $\hat{C}_k$. But, from Section 2, $\hat{C}_k$ is itself related through linear transformation to $\hat{C}_m(mel)$ and $\hat{C}_n(mel,\alpha)$. Using inverse DCT relations, we can write

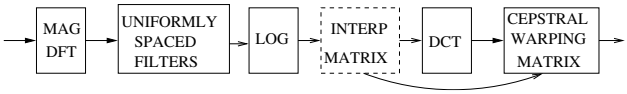

Figure 2: The signal processing blocks in our linear transformation approach to computing warped features.

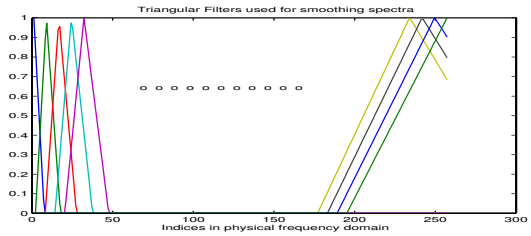

Figure 3: Figure shows the uniformly spaced but varying bandwidth filters that are used for smoothing the spectra.

the linear transformation between $\hat{C}_m(mel)$ and $d_m(mel)$, and $\hat{C}_n(mel, \alpha)$ and $d_n(mel, \alpha)$. Therefore, with the knowledge of d_k and assuming proper smoothing, we can do any type of frequency-warping and obtain the corresponding MFCC coefficients.

In conventional MFCC implementation, the mel-warping and the filter-bank are integrated to have a mel-warped filter-bank. Any frequency-warping operation that is necessary is therefore done by appropriate modification of the filter-bank. However, here (see Fig.2) we have separated the filter-bank smoothing and the frequency-warping operation, and any frequency-warping operation (including mel) can be implemented by either (i) integrating warping into sinc-interpolation of log-filter bank output or (ii) as a linear-transformation of the DCT cepstra.

5. Recognition Experiments

We demonstrate this linear transformation approach to warping on a large vocabulary speech recognition task. In conventional MFCC, we take M equally-spaced and uniform bandwidth triangular filters in the *mel* domain. This corresponds to non-uniformly spaced (because of mel-warping) and non-uniform bandwidth filters in the physical frequency domain. We then apply DCT on the logarithm of the filter-bank output to get MFCC features. In our linear-transformation approach, we have M uniformly spaced but non-uniform bandwidth filters in the physical frequency domain as seen in Fig. 3. This is because the filters are matched to have the same uniform bandwidth in the mel-domain as conventional MFCC. The warping (including mel) is implemented by appropriate linear transformation as seen in Fig. 2. The linear transformation, therefore, implicitly does the warping of filter-bank center-frequencies.

The triangular filters are not very "smooth" filters, and the cepstra do not sufficiently decay when we use only 20 filters in the standard RWTH implementation of MFCC. Therefore, we have increased the number of filters to 33, and have appropriately increased the filter-width to substantially reduce the aliasing effects. Under these "alias-free' conditions, conventional MFCC obtained by integrated mel-warped filter-bank and our linear transformation approach should yield the same results.

In this paper, as an illustration of our method, we show the results for mel-warped cepstra obtained by using linear transformation between d_k and $d_m(mel)$. This approach is compared with conventional MFCC with the integrated mel-warped filter-bank. In both cases, we use 33 filters and similar filter-widths. We have used 16 MFCC features and the corresponding first derivatives and also the second derivative of zeroth cepstral coefficient to form a 33 dimensional feature vector. We have used decision-tree state clustering and the tree contains 4500 states. A single global variance vector is used and the cross-word triphone HMM models have been trained using 41 hours of English European Parliamentary (EPPS) data. The system was evaluated using about 4 hours of development data – dev05. We see from Table 2 that the word-error rates are almost identical showing that conventional MFCC can be implemented either by integrated mel-warped filter-bank or by our approach.

Signal Processing Method	dev05
Conventional Mel-warped FB Cepstral Coeff.	17.93
Mel-warping by Linear Transformation	17.96

Table 2: We compare the word-error rates (WERs) on English EPPS development set – dev05. As we can see that the WERs are comparable, with the minor difference being due to the cepstra being not completely que-frency limited due to the use of triangular filters.

The issue of VTLN warping through linear transformation of conventional MFCC, i.e. from $d_m(mel)$ to $d_n(mel, \alpha)$, and the issue of Jacobian will be discussed in future work using the transformations derived in this paper.

6. Discussion

In this paper, we have shown that we can linearly transform cepstra obtained from filter-bank smoothed spectra to obtain different warping operations. This has the advantage that the cepstra need not be recomputed for each warping operation but can be linearly transformed from the original cepstra as long as the cepstra is que-frency limited. This illustrates the necessity of the smoothing of the spectra before the computation of cepstra and serves as an "anti-aliasing" function. In all our experiments we have attempted to mimic the signal processing used in conventional MFCC and have shown that similar results can be obtained either by linear-transformation approach or frequency-warping approach.

7. Acknowledgements

S. Umesh gratefully acknowledges the Humboldt Research Foundation for supporting this work through the Humboldt Research Fellowship.

8. References

[1] M. Pitz and H. Ney, "Vocal Tract Length Normalization Equals Linear Transformation in Cepstral Space," *IEEE Trans. Speech Audio Processing*, 2005.

[2] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing Mel- Frequency Cepstral Coefficients on the Power Spectrum," in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, 2001, pp. 73–76.

[3] D. Kim, S. Umesh, M. Gales, T. Hain, and P. Woodland, "VTLN in Broadcast News Transcription," in *Proc. of Interspeech 2004*, 2004.

[4] R. Sinha, "Front End Signal Processing For Speaker Normalization," Ph.D. dissertation, Indian Institute of Technology, Kanpur, India, June 2004.

[5] S. Umesh, R. Sinha, and S. V. B. Kumar, "An Investigation into Front-End Signal Processing for Speaker Normalization," in *Proc. ICASSP*, 2004, pp. 345–348.