Frequency Warping by Linear Transformation of Standard MFCC

Sankaran Panchapagesan

Department of Electrical Engineering
University of California, Los Angeles, U.S.A.
panchap @ icsl.ucla.edu

Abstract

A novel linear transform (LT) is proposed for frequency warping (FW) with standard filterbank based MFCC features. Here, we use the idea of spectral interpolation of [9] to perform a continuous warping in the log filterbank output domain, and incorporate both interpolation and warping into a single warped IDCT matrix. The new transformation matrix is thus mathematically simpler than in [9], and no modification of standard MFCC feature extraction is required like the previous approach. In VTLN experiments with maximum likelihood score (MLS) estimation of the FW parameter, the new LT outperformed regular VTLN implemented by warping the Mel filterbank. In speaker adaptation experiments using the new LT to transform HMM means, the results were significantly better than MLLR for limited adaptation data and comparable to those in [8], while using the computationally simpler MLS FW estimation.

Index Terms: speech recognition, speaker normalization, frequency warping, linear transformation, speaker adaptation

1. Introduction

Spectral frequency warping (FW) methods have proven to be very effective in reducing the acoustic mismatch between a speech recognition system and a new test speaker, particularly with limited adaptation data. FW is usually applied during feature extraction, as vocal tract length normalization (VTLN) ([1]-[4]).

One method of estimating the FW is by aligning formant frequencies or formant-like spectral peaks of the training and test speakers, particularly the third formant (F3) [2, 5, 8]. More commonly, the warp factor(s) controlling the FW is(are) estimated by optimizing a maximum likelihood (ML) criterion [1, 3, 4, 7].

FW of the spectrum may be shown to be equivalent to a linear transformation in the cepstral space ([3, 6]). This is also true for cepstral features which are based on Perceptual Linear Prediction (PLP) or by Mel warping of the frequency axis ([7, 4]).

The linearity of the transformation of cepstral features confers some important advantages. Firstly, for VTLN, one can apply the FW transform to previously computed features and not have to recompute features with different warp factors during FW estimation. This results in significant computational savings [9]. Secondly, the linearity enables one to take the expectation and thereby apply the same transformation to the means of the HMM distributions [5, 6]. In this way, different transforms can be estimated for different *classes* of HMM distributions, unlike VTLN where the same transformation is applied to all speech features [7].

Therefore, approximate linear transforms have also been developed for FW with standard Mel frequency cepstral coefficient (MFCC) features computed using a filterbank and the DCT [5, 8, 9]. Claes et al. derived an approximate linear transform for small warping factors [5]. Cui and Alwan [8] derived a simpler linear transform that may be shown to be a special case of Claes et al.'s transform (see Section 2), but was demonstrated to give better performance when used for speaker adaptation [8]. Their transform was in effect an "index mapping" on the filterbank outputs, i.e. one filterbank output was mapped to another, based on a FW estimated in the linear frequency domain by alignment of formant-like peaks.

Umesh et al. [9] showed that under the assumption of quefrency limitedness, the computation of the cepstral linear transformation in [3] could be considerably simplified using the idea of sinc interpolation of the log spectrum. They also extended the linear transformation to MFCCs by separating the filterbank smoothing (which leads to approximate quefrency limiting) and frequency warping operations thus modifying the standard MFCC feature extraction scheme.

In this paper, we develop a novel linear transform by using the idea of spectral interpolation in [9], to perform a continuous warping of the log filterbank outputs instead of the discrete mapping in [8]. The interpolation and warping are performed together using a single warped IDCT matrix, and the resulting transform is therefore mathematically simpler than that in [9] and unlike [9], no modification of the standard MFCC feature extraction scheme is required. The mathematical details are given in Section 3.

The warping in the IDCT matrix is parametrized and the parameter can be estimated directly by maximizing the likelihood score, without using the intermediate linear frequency spectrum as in [8]. With a smooth parametrization of the FW, there is also the possibility of estimating more flexible multiple parameter FWs by optimization techniques as in [7, 10].

The rest of this paper is organized as follows. In Section 2 we review previous work on FW as a linear transformation on MFCCs in more detail. The matrix for the new linear transformation is derived in Section 3. We then consider the estimation of FWs in Section 4 and experimental results are presented in Section 5.

2. FW as Linear Transformation of MFCC

Standard MFCC based features are computed as shown in Figure 1, using a filterbank which is usually as shown in Figure 2, with half-overlapping filters whose center frequencies are spaced equally apart on the Mel scale.

The MFCCs are therefore given by

$$\mathbf{c} = C \cdot \log(H \cdot \mathbf{S}) \tag{1}$$

where $\mathbf{S}$ is the power or magnitude (linear frequency) spectrum typically obtained as a vector for a given windowed speech frame

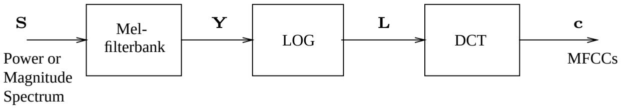

Figure 1: Standard MFCC computation.

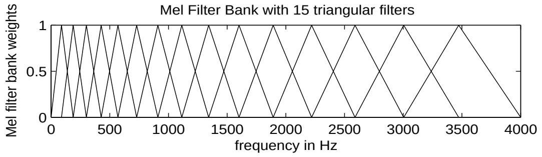

Figure 2: The shape of the Mel filter bank shown for the case when f_s is 8kHz and the number of filters is 15.

using the FFT, H is the Mel filterbank matrix, and C is the DCT matrix.

A non-linear FW transform for MFCCs may be derived as in [5]:

$$\hat{\mathbf{c}} = C \cdot \log\{H \cdot W \cdot H^{-1} \cdot \exp(C^{-1}\mathbf{c})\} \tag{2}$$

where W is the warping matrix in the linear frequency domain, and H^{-1} and C^{-1} are (approximate) inverses of H and C respectively.

Claes et al. [5] have shown that for small frequency scaling factors the transformation of the cepstrum in Equation 2 may be approximately linearized to

$$\hat{\mathbf{c}} \approx (C\bar{B}C^{-1})\mathbf{c} + C\mathbf{d} \tag{3}$$

where $\bar{B}$ is the matrix obtained from $B = H \cdot W \cdot H^{-1}$ by normalizing each of the rows of B so that the sum of the elements in each row is 1: $\bar{B}(i,j) = B(i,j)/\sum_j B(i,j)$, and $\mathbf{d}(i) = \log \sum_j B(i,j)$.

Cui and Alwan [8], approximated H, W and H^{-1} in Equation 2 by carefully chosen index mapping (IM) matrices, which are matrices in which each row contains only one nonzero element which is 1. Then, it is not difficult to see that the exponential and the logarithm in Equation 2 cancel each other out, and the transformation becomes

$$\hat{\mathbf{c}} = (CHWH^{-1}C^{-1}) \cdot \mathbf{c} \tag{4}$$

Note that when H, W and H^{-1} are IM matrices, so is $B = H \cdot W \cdot H^{-1}$ and so in Equation 3, $\bar{B} = B$ and $\mathbf{d} = 0$ and therefore Equation 3 reduces to Equation 4. Cui and Alwan's linear transform is therefore mathematically a special case of Claes et al's transform.

Since HWH^{-1} is an IM matrix, the transformation amounts to an index mapping on the approximate log-filterbank output $\mathbf{L} = C^{-1}\mathbf{c}$. i.e., filterbank outputs are just replaced by other filterbank outputs depending on W which is estimated in the linear frequency domain by alignment of formant-like peaks.

This suggests the possibility of estimating and applying FWs directly on the log Mel spectrum $\mathbf{L}$. This would eliminate the need to estimate a linear frequency spectrum $\mathbf{S}$ using an approximate inverse of the filterbank. Also, since the dimension of $\mathbf{L}$ is usually much smaller than that of $\mathbf{S}$, this also reduces the computational requirement in the estimation and application of the adaptation transform.

In [9], Umesh et al. use the ideas of quefrency limitedness and sinc interpolation to show that any linear frequency warping may be applied as a linear transformation on plain cepstra $\mathbf{C}$:

$$\tilde{\mathbf{C}} = D \cdot \mathbf{C} \tag{5}$$

where the matrix D depends on the linear FW $g(\omega)$.

In their approach, the MFCC feature extraction scheme was modified by separating the filterbank smoothing and Mel warping operations. The filters of their filterbank were uniformly spaced in the linear frequency domain, but of uniform bandwidth in the Mel domain. Unwarped DCT-cepstra are computed using a DCT on the log of the filterbank output $\mathbf{l} = \{\log |X_{FB}[q]|^2,\ q = 0, 1, \ldots, M-1\}$:

$$\mathbf{d} = T_1 \cdot \mathbf{l} \tag{6}$$

where $T_1 = C$ is the DCT matrix used in standard MFCC computation. The warping, including Mel and VTLN warping is to be applied as a matrix on $\mathbf{d}$. But for the purpose of frequency warping, the "plain" cepstra are first computed as

$$\mathbf{C} = T_2 \cdot \mathbf{l} \tag{7}$$

where T_2 is a DCT matrix different from T_1.

Mel and/or VTLN warping were applied on $\mathbf{C}$ by transforming them as in Equation 5 and then recomputing the DCT-cepstra after computing the warped log filterbank outputs. The final warped cepstra may be shown to be:

$$\tilde{\mathbf{d}} = (T_1 T_2^{-1} D T_2 T_1^{-1})\mathbf{d} \tag{8}$$

$D = D(mel)$ for regular Mel warping and $D = D(mel, \alpha)$ for Mel warping followed by VTLN warping. Since we must have $D(mel, \alpha) = D(\alpha)D(mel)$ it can be shown that the Mel & VTLN warped cepstra $\mathbf{d}(mel, \alpha)$ are related to the Mel warped cepstra $\mathbf{d}(mel)$ by:

$$\mathbf{d}(mel, \alpha) \;=\; (T_1 T_2^{-1} D(\alpha) T_2 T_1^{-1}) \cdot \mathbf{d}(mel) \tag{9}$$

The cepstral transformation matrix in Equation 9 is quite complicated because of the use of the "plain" cepstrum for FW. However, a much simpler matrix for linear transformation of cepstra may be derived by direct warping of the *cosine* interpolated log Mel spectrum. We will discuss this next.

3. Derivation of the Transformation Matrix

With a unitary DCT matrix C, we have $C^{-1} = C^T$, and the equation $\mathbf{L} = C^{-1}\mathbf{c} = C^T\mathbf{c}$ may be written in expanded form as

$$\mathbf{L}(m) = \sum_{k=0}^{N-1} \mathbf{c}(k)\alpha_k \cos\left(\frac{\pi(2n-1)k}{2M}\right),\ m = 1, 2, \ldots, M \tag{10}$$

where $\mathbf{c}(k), k = 0, 1, \ldots, N-1$, are the MFCCs and

$$\alpha_k = \begin{cases} \sqrt{\frac{1}{M}}, & k = 0 \\ \sqrt{\frac{2}{M}}, & k = 1, 2, \ldots, N-1 \end{cases}$$

is a factor that ensures that the DCT is unitary.

Using the idea of cosine interpolation one can consider the IDCT approximation of Equation 10 to describe a continuous log Mel spectrum $L(u)$, where u is a continuous "Mel" frequency variable:

$$L(u) = \sum_{k=0}^{N-1} \mathbf{c}(k)\alpha_k \cos\left(\frac{\pi(2u-1)k}{2M}\right) \tag{11}$$

with

$$\mathbf{L}(m) = L(u)|_{u=m}, m = 1, 2, \ldots, M$$

One detail is the range of values that u can take. $L(u)$ as described in Equation 11 above is periodic with a period of $2M$, and is symmetric about the point $u = M + \frac{1}{2}$. Therefore, we may take the range of u to be $\dfrac{1}{2} \leq u \leq M + \dfrac{1}{2}$.

We will apply frequency warping functions on u, which are obtained by as follows. Let λ be a normalized frequency with $0 \leq \lambda \leq 1$. We can pass from the continuous Mel domain u to the normalized frequency domain λ and vice versa by the transformations

$$\begin{aligned} u &\rightarrow \lambda = \frac{u - 1/2}{M}, \;\; \frac{1}{2} \leq u \leq M + \frac{1}{2} \\ \lambda &\rightarrow u = \frac{1}{2} + \lambda M, \;\; 0 \leq \lambda \leq 1 \end{aligned}$$

Let $\theta_p(\lambda)$ be a normalized FW function controlled by parameter(s) p (see Eqs. 22 and 23). Then we can obtain a warping $\psi_p(u)$ on u, using

$$\psi_p(u) = \frac{1}{2} + M \cdot \theta_p\left(\frac{u - 1/2}{M}\right) \tag{12}$$

Note that if $\lambda = 0$ and $\lambda = 1$ are fixed points of $\theta_p(\lambda)$, then $u = \frac{1}{2}$ and $u = M + \frac{1}{2}$ as fixed points of $\psi_p(u)$.

We now take the *inverse* FW function to be applied to u to be $\psi_p(u)$. The warped log Mel spectrum is then:

$$\begin{aligned} \hat{L}(u) &= L(\psi_p(u)) \\ &= \sum_{k=0}^{N-1} \mathbf{c}(k)\alpha_k \cos\left(\frac{\pi(2\psi_p(u) - 1)k}{2M}\right) \end{aligned} \tag{13}$$

The warped Log filterbank output is

$$\hat{\mathbf{L}}(m) = \hat{L}(u)|_{u=m}, \; m = 1, 2, \ldots, M \tag{14}$$

In vector form,

$$\hat{\mathbf{L}} = \tilde{C}_p \cdot \mathbf{c}$$

where $\tilde{C}_p$ is the *warped IDCT matrix*:

$$\tilde{C}_p = \left[\alpha_{j-1} \cos\left(\frac{\pi(2\psi_p(i) - 1)(j-1)}{2M}\right)\right]_{\substack{1 \leq i \leq M \\ 1 \leq j \leq N}} \tag{15}$$

The transformed MFCCs are given by

$$\begin{aligned} \hat{\mathbf{c}} &= C\,\hat{\mathbf{L}} = (C\tilde{C}_p)\,\mathbf{c} \\ &= T_p\,\mathbf{c} \end{aligned} \tag{16}$$

Hence, the warped MFCCs may be obtained by a linear transformation of the original MFCCs, and the transformation matrix is given by

$$T_p = (C\tilde{C}_p) \tag{17}$$

where $\tilde{C}_p$ is the warped IDCT matrix given in equation 15.

Comparison with previous transforms:

Comparing our linear transform in Equation 17 with that of [9] in Equation 9, it is clear that our linear transformation is mathematically much simpler without any change to the standard MFCC feature extraction scheme as in [9].

Also since the warping is incorporated directly into the IDCT matrix, the FW parameter(s) p can be estimated directly using an MLS criterion (see Section 4), without using the intermediate linear frequency domain as for [8] (Equation 4).

Transformation of Features and HMM means:

The final feature vector $\mathbf{x}$ consists of the MFCCs and their first and second time derivatives. The transform on the time derivatives of the cepstral features will also be linear [5, 8]:

$$\begin{aligned} \widehat{\Delta\mathbf{c}} &= T_p\Delta\mathbf{c} \quad (18) \\ \widehat{\Delta^2\mathbf{c}} &= T_p\Delta^2\mathbf{c} \quad (19) \end{aligned}$$

Therefore, the feature vector $\mathbf{x} = \begin{bmatrix} \mathbf{c} \\ \Delta\mathbf{c} \\ \Delta^2\mathbf{c} \end{bmatrix}$ may be transformed as:

$$\mathbf{x}^p = A_p\mathbf{x}, \;\; \text{where} \;\; A_p = \begin{bmatrix} T_p & 0 & 0 \\ 0 & T_p & 0 \\ 0 & 0 & T_p \end{bmatrix} \tag{20}$$

where the transformed feature vector $\mathbf{x}^p$ is now a function of the FW parameters, p. Taking the expectation, the mean μ of a given HMM distribution may be transformed as [5, 8]:

$$\hat{\mu} = A_p\mu, \tag{21}$$

Examples of Frequency Warping Functions:

1. **Piecewise Linear** These are the type of FW functions that are most commonly used in vocal tract length normalization (VTLN) in the front-end as in [3].

$$\theta_p(\lambda) = \begin{cases} p\lambda, & 0 \leq \lambda \leq \lambda_0 \\ p\lambda_0 + \left(\frac{1 - p\lambda_0}{1 - \lambda_0}\right)(\lambda - \lambda_0), & \lambda_0 < \lambda \leq 1 \end{cases} \tag{22}$$

2. **Linear** This FW can be used for adaptation from adult models to children's models, where the original models have more spectral information than necessary for children's speech. They may be used for a global adaptation of all the means before subsequent multi-class adaptation. For $p \leq 1$,

$$\theta_p(\lambda) = p\lambda, \; 0 \leq \lambda \leq 1 \tag{23}$$

4. Estimation of the FW function

The maximum likelihood score (MLS) criterion is commonly used for VTLN estimation ([1, 3]), to estimate the optimal FW parameters $\hat{p}$:

$$\hat{p} = \arg\max_p \left[\log P(\mathbf{X}^p, \Theta^p | W, \Lambda) + T \log |A_p|\right] \tag{24}$$

where p is(are) the FW parameter(s), $\mathbf{x}^p = A_p\mathbf{x}$ is a normalized feature vector, $|A_p|$ is the determinant of A_p, $\mathbf{X}^p = \{\mathbf{x}_1^p, \mathbf{x}_2^p, \ldots, \mathbf{x}_T^p\}$ is the normalized adaptation data, W is the word (or other unit) transcription, Λ are the corresponding HMMs, and Θ^p is the ML HMM state sequence with which $\mathbf{X}^p$ are aligned to Λ^p by the Viterbi algorithm during ASR decoding.

For regular VTLN by Mel bin center frequency warping [1], the objective function only includes the first term in Equation 24. In our experiments with the Linear Transformation too, the determinant term was not used since better results were obtained without it.

The same MLS criterion can also to be used to estimate the FW parameters to be used to transform the means of the HMMs as in Equation 21:

$$\hat{p} = \arg\max_p \left[\log P(\mathbf{X}, \Theta^p | W, \Lambda^p)\right] \tag{25}$$

where the variables are as explained above for Equation 24 except that here it is not the adaptation data but the HMMs Λ that are modified to Λ^p for FW parameters p.

5. Experimental Results

We tested the developed linear transform on connected digit recognition of children's speech using the TIDIGITS database. The baseline system was the same as in [8]. 20 HMMs including 18 monophone models and the silence and short pause models, were trained for connected digit recognition from the adult male speakers in TIDIGITS. The number of states per monophone varied from 2 to 4 with 6 Gaussian mixtures in each state. The features used for speech recognition consisted of the first 13 MFCCs and their first and second time derivatives. Ten children, five boys and five girls were selected for testing. The baseline recognition word accuracy was 38.9 %. Tables 1 and 2 show the results of VTLN and speaker adaptation experiments respectively, with 1, 5 and 11 digits used for adaptation.

Table 1: Recognition Accuracy in VTLN Experiments: (1) MLLR (2) Regular VTLN by Mel Bin Center Frequency Warping (3) VTLN with Our Linear Transform (LT-VTLN)

	Number of adaptation digits			
Algorithm	0	1	5	11
MLLR	38.9	40.6	63.4	90.9
Regular VTLN	38.9	80.8	82.7	86.9
New LT-VTLN	38.9	89.1	90.4	90.9

The results in Table 1 demonstrate the effectiveness of VTLN using the new linear transformation (LT-VTLN) over regular VTLN performed by warping the center frequencies of the Mel filterbank. In both cases, an optimal speaker-specific warp factor for the piecewise-linear FW was estimated from the adaptation data, using a grid search to optimize the MLS criterion of Section 4. In [8], VTLN warp factors were estimated on a per utterrance basis, therefore with different recognition results. A speaker-specific warp factor is computationally simpler and was also considered more appropriate for a comparison with MLLR. For the LT-VTLN, the Jacobian normalization actually resulted in worse performance and was therefore not used. The results are compared with Maximum Likelihood Linear Regression (MLLR) [11] with a full regression matrix (which gave better results than the 3-block structure used in [8]). LT-VTLN consistently outperforms regular VTLN, and both outperform MLLR for small amounts of data, as expected. For more than 11 adaptation digits, MLLR starts to perform better than LT-VTLN.

Table 2: Recognition Accuracy in Speaker Adaptation Experiments

	Number of adaptation digits			
Algorithm	0	1	5	11
MLLR	38.9	40.6	63.4	90.9
Peak Alignment	38.9	86.7	88.3	89.2
New LT, MLS	38.9	87.0	89.2	89.4

The results in Table 2 demonstrate the effectiveness of adaptation of HMM Means in the back end using the new linear transformation. The performance is compared with that of the Peak Alignment based adaptation approach of [8] (with only the FW transform), and the results are comparable. However, the FW parameter for our new LT was directly estimated by a grid search with the computationally simpler MLS criterion while the Peak Alignment approach used alignment of F3 estimated from the intermediate linear frequency spectrum.

6. Conclusions

A novel linear transform was developed for frequency warping with standard filterbank based MFCC features by a smooth parametrization of the discrete FW transform introduced in [8], using the spectral interpolation idea in [9]. Using cosine interpolation in the log filterbank output domain, the derived transformation matrix was much simpler mathematically, than that in [9]. In Vocal Tract Length Normalization (VTLN) experiments the new linear transform outperformed regular VTLN implemented by warping the center frequencies of the Mel filterbank, when a maximum likelihood score (MLS) criterion was used to estimate FW parameter for both methods. Results comparable to those in [8] were obtained for adaptation of HMM means using the new Linear Transform with the computationally simpler MLS criterion used to estimate the FW parameter.

7. Acknowledgements

The author wishes to thank Abeer Alwan and Xiaodong Cui and for discussions on this work and for the use of their recognition and adaptation systems. This work was supported in part by the NSF.

8. References

[1] L. Lee and R. C. Rose, ”Speaker normalization using efficient frequency warping procedures,” *ICASSP*-96, pp.353-356.

[2] E. B. Gouvea and R. M. Stern, ”Speaker normalization through formant-based warping of the frequency scale,” *Eurospeech* 1997, vol. 3, pp.1139-1142.

[3] M. Pitz, S. Molau, R. Schlueter, H. Ney, ”Vocal Tract normalization equals linear transformation in cepstral space”, *Eurospeech 2001*, pp.721-724.

[4] M. Pitz and H. Ney, ”Vocal Tract normalization as linear transformation of MFCC”, *Eurospeech 2003*, pp.1445-1448.

[5] T. Claes, I. Dologlou, L. ten Bosch, and D. Van Compernolle, ”A novel feature transformation for vocal tract length normalization in automatic speech recognition,” *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 6, pp.549-557, November 1998.

[6] J. McDonough and W. Byrne, ”Speaker adaptation with all-pass transforms,” *Proc. ICASSP*, Vol.2 , pp.757-60, 1999.

[7] J. W. McDonough, ”Speaker compensation with all-pass transforms,” Ph.D. dissertation, Johns Hopkins University, Baltimore, Maryland, 2000.

[8] X. Cui and A. Alwan, ”MLLR-Like Speaker Adaptation Based on Linearization of VTLN with MFCC features,” *Interspeech 2005*, pg. 273-276.

[9] S. Umesh, A. Zolnay and H. Ney ”Implementing frequency-warping and VTLN through linear transformation of conventional MFCC,” *INTERSPEECH* 2005.

[10] S. Panchapagesan and A. Alwan, ”Multi-parameter Frequency warping for VTLN by gradient search,” *ICASSP 2006*, I-1181.

[11] 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.