Frequency Warping for VTLN and Speaker Adaptation by Linear Transformation of Standard MFCC

Sankaran Panchapagesan *, Abeer Alwan

Department of Electrical Engineering, The Henry Samueli School of Engineering and Applied Science, 66-147E Engr. IV, 405 Hilgard Avenue, Box 951594, University of California, Los Angeles, CA 90095-1594, USA

Abstract

Vocal Tract Length Normalization (VTLN) for standard filterbank-based Mel Frequency Cepstral Coefficient (MFCC) features is usually implemented by warping the center frequencies of the Mel filterbank, and the warping factor is estimated using the maximum likelihood score (MLS) criterion (Lee and Rose, 1998). A linear transform (LT) equivalent for frequency warping (FW) would enable more efficient MLS estimation (Umesh et al., 2005). We recently proposed a novel LT to perform FW for VTLN and model adaptation with standard MFCC features (Panchapagesan, 2006). In this paper, we present the mathematical derivation of the LT and give a compact formula to calculate it for any FW function. We also show that our LT is very closely related to previously proposed LTs for FW (McDonough, 2000; Pitz et al., 2001; Umesh et al., 2005), and these LTs for FW are all found to be numerically almost identical for the sine-log all-pass transform (SLAPT) warping functions. Our formula for the transformation matrix is, however, computationally simpler and unlike other previous linear transform approaches to VTLN with MFCC features (Pitz and Ney, 2003; Umesh et al., 2005), no modification of the standard MFCC feature extraction scheme is required. In VTLN and Speaker Adaptive Modeling (Welling et al., 2002) experiments with the DARPA Resource Management (RM1) database, the performance of the new LT was comparable to that of regular VTLN implemented by warping the Mel filterbank, when the MLS criterion was used for FW estimation. This demonstrates that the approximations involved do not lead to any performance degradation. Performance comparable to front end VTLN was also obtained with LT adaptation of HMM means in the back end, combined with mean bias and variance adaptation according to the Maximum Likelihood Linear Regression (MLLR) framework. The FW methods performed significantly better than standard MLLR for very limited adaptation data (1 utterance), and were equally effective with unsupervised parameter estimation. We also performed Speaker Adaptive Training (SAT) with feature space LT denoted CLTFW. Global CLTFW SAT gave results comparable to SAM and VTLN. By estimating multiple CLTFW transforms using a regression tree, and including an additive bias, we obtained significantly improved results compared to VTLN, with increasing adaptation data.

Key words: automatic speech recognition, speaker normalization, VTLN, frequency warping, linear transformation, speaker adaptation

1 Introduction

Vocal Tract Length Normalization (VTLN) is a speaker normalization technique widely used to improve the accuracy of speech recognition systems. In VTLN, spectral mismatch caused by variation in vocal tract lengths of speakers is reduced by performing spectral frequency warping (FW) or its equivalent, typically during feature extraction. VTLN has proven to be particularly effective when only limited adaptation data from a test speaker is available, even in an unsupervised mode. The estimation and implementation of frequency warping have received much attention in recent years.

The parameters controlling the FW are commonly estimated by optimizing a maximum likelihood (ML) criterion over the adaptation data. The ML criterion could be the ASR likelihood score of the recognizer over the adaptation data (Lee and Rose, 1998; Pitz et al., 2001; Pitz and Ney, 2003), the EM auxiliary function (Dempster et al., 1977; McDonough, 2000; Loof et al., 2006), or likelihoods of Gaussian mixture models (GMMs) trained specifically for FW parameter estimation (Wegmann et al., 1996; Lee and Rose, 1998). Another FW estimation method is by alignment of formants or formant-like spectral peaks between the test speaker and a reference speaker from the training set (Gouvea and Stern, 1997; Claes et al., 1998; Cui and Alwan, 2006).

Maximizing the likelihood score is commonly performed using grid search over a set of warping factors, when the FW is described by a single parameter that controls the scaling of the frequency axis (Lee and Rose, 1998). More recently, optimization methods based on the gradient and higher order derivatives of the objective function have been used to estimate the FW function. This allows efficient estimation of multiple parameter FW functions like the All-Pass Transform (APT) FWs, which can give better recognition performance than single parameter FWs (McDonough, 2000; Panchapagesan and Alwan, 2006).

Frequency warping of the spectrum has been shown to correspond to a linear transformation in the cepstral space (McDonough et al., 1998; Pitz et al., 2001). This relationship confers some important advantages for speech recognition systems that use cepstral features. Firstly, one can apply the linear transform to previously com-

* Corresponding Author.
Email addresses: panchap@ee.ucla.edu (Sankaran Panchapagesan), alwan@ee.ucla.edu (Abeer Alwan).

puted unwarped features and not have to recompute features with different warp factors during VTLN estimation. This results in significant computational savings (Umesh et al., 2005), which would be important in embedded and distributed speech recognition (DSR) applications, where resources are limited. Given the recognition alignment of an utterance obtained with baseline models without VTLN, it can be shown by a rough calculation that parameter estimation for Regular VTLN is about 2.5 times as expensive as for LT VTLN, when the fixed alignment is used for VTLN estimation with the MLS criterion, with single Gaussian mixture HMMs and a grid search. The linear transform approach also has the advantage that one need not have access to any of the intermediate stages in the feature extraction during VTLN estimation. This aspect would have definite advantages in DSR, where feature extraction is performed at the client and recognition is performed at the server. During VTLN estimation using a grid search over warping factors, since it would be impractical for the client to recompute and transmit features for each warping factor, warped features would have to be computed at the server. With a linear transform, only the cepstral transformation matrices for each warping factor need to be applied to unwarped features to choose the best warping factor, while with VTLN by spectral warping, the linear frequency spectrum needs to be reconstructed and the warped features recomputed for each warping factor.

The linearity also enables one to take the expectation and thereby apply the linear transformation to the means of HMM distributions (Claes et al., 1998; McDonough and Byrne, 1999). Different transforms could then be estimated for different phonemes or classes of HMM distributions, unlike VTLN where the same global transformation is applied to all speech features (McDonough, 2000).

Mel frequency cepstral coefficients (MFCCs) computed using a filterbank and the DCT (Davis and Mermelstein, 1980), are a very popular choice of features for speech recognition. The equivalence of FW to linear transformation, though true also for cepstral features which are based on Perceptual Linear Prediction (PLP) or by Mel warping of the frequency axis (McDonough, 2000; Pitz and Ney, 2003), does not hold exactly for standard MFCC features. In fact, for standard MFCC features, because of the non-invertible filterbank with non-uniform filter widths, even with the assumption of quefrency limitedness, the MFCC features after warping cannot even be expressed as a function (linear or non-linear) of the unwarped MFCC features. i.e., for a given warping of the linear frequency signal spectrum, there is not a single function (for all possible cepstra) that will give the warped cepstra from the unwarped cepstra. Hence, approximate linear transforms have been developed for FW with MFCC features (Claes et al., 1998; Cui and Alwan, 2006; Umesh et al., 2005).

Claes et al. (1998) were the first to derive an approximate linear transform which was used to perform model adaptation with some success. Cui and Alwan (2005, 2006) derived a simpler linear transform that is essentially an “index mapping” on the Mel filterbank outputs, i.e. one filterbank output is mapped to another. In fact, it may be shown to be mathematically a special case of Claes et al.’s transform (see Section

2) but was demonstrated to give better performance (Cui and Alwan, 2005). In both Claes et al. (1998) and Cui and Alwan (2006), the FW was estimated by alignment of formants or formant-like peaks in the linear frequency domain.

Umesh et al. (2005) showed that the formula for computing the linear transform for ordinary cepstra, derived in Pitz et al. (2001), could be considerably simplified under the assumption of quefrency limitedness of the cepstra, when the log spectrum can be obtained from samples by sinc interpolation. They also developed non-standard filterbank based MFCC features, to which the linear transformation was extended. In their modified filterbank, the filter center frequencies were uniformly spaced in the linear frequency domain but filter bandwidths were uniform in the Mel domain. Their transformation formula (discussed further in Section 4) was, however, complicated by the use of two different DCT matrices, one for warping purposes and the other for computing the cepstra.

In Panchapagesan (2006), we introduced a novel linear transform for MFCCs that required no modification of the standard MFCC feature extraction scheme. The main idea was to directly warp the continuous log filterbank output obtained by cosine interpolation with the IDCT. This approach can be viewed as using the idea of spectral interpolation of Umesh et al. (2005), to perform a continuous warping of the log filterbank outputs instead of the discrete mapping in Cui and Alwan (2006). However, a single warped IDCT matrix was used to perform both the interpolation and warping, thus resulting in a simpler mathematical formula for computing the transform compared to Umesh et al. (2005). Also, the warping in the IDCT matrix is parametrized and the parameter can be estimated directly by optimizing an objective criterion, without using the intermediate linear frequency spectrum as in the *Peak Alignment* method of Cui and Alwan (2006). As mentioned above, this would be advantageous in distributed speech recognition, where intermediate variables in the feature extraction have to be reconstructed at the recognizer. Also, with a smooth parametrization of the FW, it is possible to estimate the FW parameters by faster optimization techniques as in McDonough (2000) and Panchapagesan and Alwan (2006) instead of the commonly used grid search, and also perform simultaneous optimization of several parameters.

In Panchapagesan (2006), we validated the technique on connected digit recognition of children's speech, and showed that for that task, it performed favorably compared to regular VTLN by warping the filterbank. We also compared the method in the back end with the Peak Alignment method (Cui and Alwan, 2006), and showed comparable and slightly better results.

In this paper, the mathematical derivation of our Linear Transform (LT) is presented in more detail, and the final formula for computing the LT for any given frequency warping function and parameter is expressed in a simple and compact form. We validate the LT further by demonstrating its effectiveness in continuous speech recognition using the DARPA Resource Management (RM1) database. These include ex-

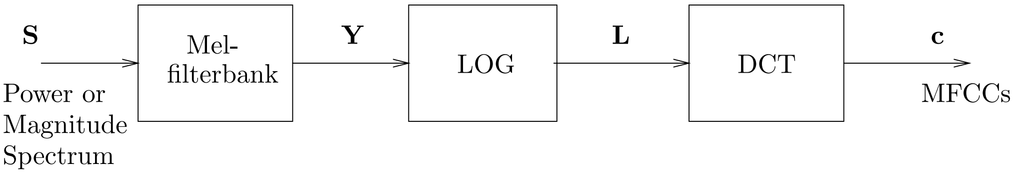

Fig. 1. Standard MFCC computation.

periments with front end VTLN and back end adaptation of HMM means, as well as speaker adaptive modeling and training using the LT (Welling et al., 2002; Anastasakos et al., 1996). We show that in all cases, LT VTLN can give results comparable to those of regular VTLN by warping the Mel filterbank center frequencies. We also discuss optimization of the EM auxiliary function for LT estimation, and show that by estimating multiple transforms using a regression tree, results better than global VTLN can be obtained. Finally, we present the results of unsupervised VTLN and adaptation with the LT.

The rest of this paper is organized as follows. In Section 2 we consider the problem of deriving a linear transformation for FW of MFCCs, review previous work, and motivate the development of our new linear transform. The matrix for the new linear transformation is derived in Section 3, and the proposed transform is compared with previous approaches in Section 4. We then consider the estimation of FWs using MLS and EM auxiliary function as objective critera in Section 5, and also derive formulae for convex optimization of the EM auxiliary function for multiple FW parameters. Experimental results are presented in Section 6, and summary and conclusions in Section 7.

2 FW as Linear Transformation of Standard MFCC - Review of Previous Work

Standard MFCCs are computed as shown in Figure 1, and the Mel filterbank is shown in Figure 2. The filters are assumed to be triangular and half overlapping, with center frequencies spaced equally apart on the Mel scale.

During feature extraction, the speech signal is pre-emphasized and divided into frames and each frame is first windowed using the Hamming window. The short-time power spectrum vector $\mathbf{S}$ is obtained from the squared magnitude of the FFT of the windowed frame.

The log of the filterbank outputs is obtained as:

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

where H is the Mel filterbank matrix. Here, we use the notation that the log of a

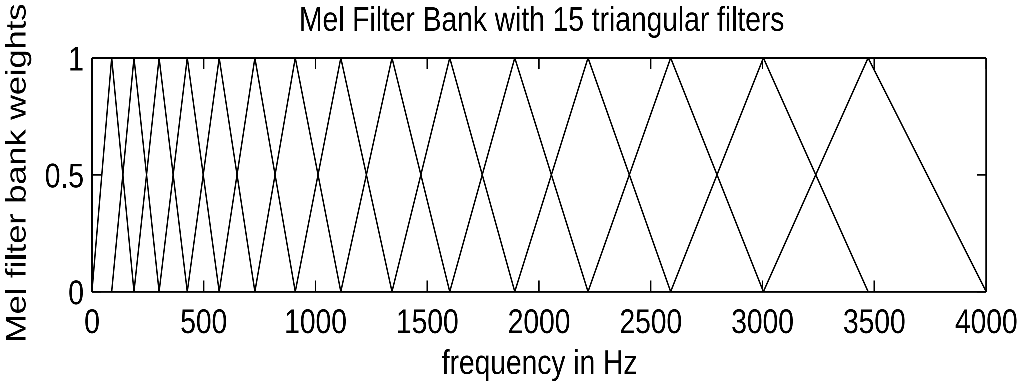

Fig. 2. The shape of the Mel filter bank shown for the case when f_s is 8kHz and the number of filters is 15.

vector is the log applied to each component.

The MFCCs are then given by

$$\begin{aligned} \mathbf{c} &= C \cdot \mathbf{L} && (2) \\ &= C \cdot \log(H \cdot \mathbf{S}) && (3) \end{aligned}$$

where C is the DCT matrix.

FW for VTLN can be applied to the linear frequency spectrum $\mathbf{S}$, or to the center frequencies of the filterbank (Lee and Rose, 1998), which is computationally more efficient since the warping only has to be performed once on the filterbank and not repeatedly for each frame of speech. For this theoretical discussion, we consider direct frequency warping of $\mathbf{S}$. Since H and C result in significant reduction of dimensionality and are non-invertible, $\mathbf{S}$ can only be approximately recovered from the Mel cepstrum $\mathbf{c}$:

$$\mathbf{S} \approx H^{-1} \cdot \exp(C^{-1}\mathbf{c})$$

where H^{-1} and C^{-1} are approximate inverses of H and C respectively. A (partial) IDCT matrix is a natural choice for C^{-1}, while different choices have been proposed for H^{-1} by Claes et al. (1998) and Cui and Alwan (2006), as discussed below.

Between the two approximate inverse operations, the application of C^{-1} is less severe since it only corresponds to a smoothing of the log filterbank output by cosine interpolation. Since the spectrum is already smoothed and warped by the Mel filterbank operation, the cepstral truncation and application of C^{-1} would result in the recovery of a reasonable Mel-warped log spectrum which can be used for further VTLN warping. The FFT spectrum recovered using an approximate filterbank inverse H^{-1}, however, would probably only be a gross approximation of the original FFT spectrum since there is large dimensionality reduction due to application of H (256×26 in our case). However, the use of a particular choice of H^{-1} to perform VTLN warping can be empirically justified by the improvement in recognition results.

By applying a warping W to the approximate linear spectrum $\mathbf{S}$ and recomputing Mel cepstra, a non-linear FW transform for MFCCs may therefore be derived as in Claes et al. (1998):

$$\hat{\mathbf{c}} = C \cdot \log\{H \cdot W \cdot H^{-1} \cdot \exp(C^{-1}\mathbf{c})\} \tag{4}$$

Claes et al. (1998) also showed that for small frequency scaling factors, the non-linear cepstral transformation of Equation 4 may be approximately linearized to:

$$\hat{\mathbf{c}} \approx (C\bar{B}C^{-1}) \cdot \mathbf{c} + C\mathbf{d} \tag{5}$$

where $\bar{B}$ is the matrix obtained from $B = H \cdot W \cdot H^{-1}$ by normalizing each row 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 $\mathbf{d}(i) = \log \sum_j B(i,j)$. For the choice of H^{-1}, Claes et al. (1998) used a special matrix M that satisfied $HM = I$, and which was found to give better results than just using the pseudo-inverse of H.

Cui and Alwan (2006) obtained a transform that has a simpler form than that in Equation 5, and was shown to give even better results, by approximating H, W and H^{-1} in Equation 4 by carefully chosen index mapping (IM) matrices, which are matrices in which each row contains only one nonzero element which is 1. Then, $B = H \cdot W \cdot H^{-1}$ is also an IM matrix, and the exponential and the logarithm in Equation 4 cancel each other out (Cui and Alwan, 2006). The cepstral transformation then becomes linear:

$$\hat{\mathbf{c}} = (CHWH^{-1}C^{-1}) \cdot \mathbf{c} \tag{6}$$

In fact, when B is an IM matrix, $\bar{B} = B$ and $\mathbf{d} = 0$ in Equation 5, and Equation 5 also reduces to Equation 6. Cui and Alwan's linear transform is therefore mathematically a special case of Claes et al's transform.

We can rewrite Equation 6 as

$$\hat{\mathbf{c}} = C \cdot \hat{\mathbf{L}} \tag{7}$$

where

$$\hat{\mathbf{L}} = HWH^{-1} \cdot \mathbf{L} = B \cdot \mathbf{L} \tag{8}$$

with

$$\mathbf{L} \approx C^{-1}\mathbf{c} \tag{9}$$

Considered from the point of view of the log Mel filterbank output $\mathbf{L}$, since B is an IM matrix, we can see from Equation 8 that Cui and Alwan's transform therefore amounts to an index mapping.

In Cui and Alwan (2006), the warping W was estimated by alignment of formant-like peaks in the linear frequency spectrum $\mathbf{S}$, and the cepstral linear transform was demonstrated to give excellent results when used for model adaptation. This raises the possibility of obtaining the same success by estimating and applying warping directly on the log Mel spectrum $\mathbf{L}$ without reconstructing the linear frequency spectrum $\mathbf{S}$ using an approximate inverse of the filterbank. Also, the discrete nature of the frequency mappings of Cui and Alwan is not conducive to estimation by efficient optimization of objective functions like maximum likelihood or discriminative criteria, which can give improved recognition accuracy, and also have other advantages over peak alignment as discussed in Section 4.

We will next discuss how to implement and estimate continuous warping on $\mathbf{L}$, the log Mel filterbank output, and show that it naturally results in a linear transformation on the MFCCs.

3 Derivation of the Novel LT by Warping the Log Mel Filterbank Output

3.1 Linearity of the Cepstral Transformation

Equation 9 describes how the smoothed log filterbank output may be approximately recovered from the truncated cepstra using the IDCT. We use a unitary type-II DCT matrix, for which we have $C^{-1} = C^T$, with

$$C = \left[\alpha_k \cos\left(\frac{\pi(2m-1)k)}{2M} \right) \right]_{\substack{0 \le k \le N-1 \\ 1 \le m \le M}} \tag{10}$$

where M is the number of filters in the filterbank, N is the number of cepstra used in the features, and

$$\alpha_k = \begin{cases} \sqrt{\frac{1}{M}}, & k = 0 \\ \sqrt{\frac{2}{M}}, & k = 1, 2, \ldots, N-1 \end{cases} \tag{11}$$

is a factor that ensures that the DCT is unitary. Similar expressions are valid for C and C^{-1} with a non-unitary type-II DCT matrix, but then $C^{-1} \neq C^T$ and two different sets of factors α_k and β_k would be required. Note that typically $N < M$ in practice.

Equation 9 therefore becomes $\mathbf{L} = C^{-1}\mathbf{c} = C^T\mathbf{c}$ (the approximation being understood implicitly) and may be written in expanded form as

$$\mathbf{L}(m) = \sum_{k=0}^{N-1} \mathbf{c}(k) \alpha_k \cos\left(\frac{\pi(2m-1)k}{2M} \right), \; m = 1, 2, \ldots, M \tag{12}$$

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

Using the idea of cosine interpolation one can consider the IDCT approximation of Equation 12 to describe a continuous log Mel spectrum $L(u)$, where u is a continuous (scaled) Mel frequency variable:

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

with

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

We can now apply continuous warping to u. Let us take the *inverse* of the warping function to be applied, to be $\psi(u)$. The warped continuous log Mel spectrum is then:

$$\hat{L}(u) = L(\psi(u)) \tag{15}$$

The warped *discrete* log filterbank output is obtained by sampling $\hat{L}(u)$:

$$\begin{aligned}
\hat{\mathbf{L}}(m) &= \hat{L}(u)|_{u=m},\ m = 1, 2, \ldots, M & (16)\\
&= L(\psi(u))|_{u=m},\ m = 1, 2, \ldots, M & (17)\\
&= \sum_{k=0}^{N-1} \mathbf{c}(k)\alpha_k \cos\left(\frac{\pi(2\psi(m)-1)k}{2M}\right),\ m = 1, 2, \ldots, M & (18)
\end{aligned}$$

by Equations 15 and 13.

Therefore, in vector form,

$$\hat{\mathbf{L}} = \tilde{C} \cdot \mathbf{c} \tag{19}$$

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

$$\tilde{C} = \left[\alpha_k \cos\left(\frac{\pi(2\psi(m)-1)k}{2M}\right)\right]_{\substack{1\le m\le M\\ 0\le k\le N-1}} \tag{20}$$

The transformed MFCCs are given by

$$\begin{aligned}
\hat{\mathbf{c}} &= C\ \hat{\mathbf{L}} = (C\tilde{C})\ \mathbf{c}\\
&= T\ \mathbf{c} & (21)
\end{aligned}$$

Hence, the MFCCs corresponding to the warped log Mel spectrum are naturally obtained by a linear transformation of the original MFCCs, and the transformation matrix is given by

$$T = C\tilde{C} \tag{22}$$

where $\tilde{C}$ is the warped IDCT matrix given in Equation 20.

3.2 Computation of the Transform Matrix

In the above derivation, one needs to specify the warping $\psi(u)$ before the transform matrix can be computed from Equations 10, 20 and 22. The first detail is the range of values that u can take. $L(u)$ as described in Equation 13 above is periodic with a period of $2M$, and is symmetric about the points $u = \frac{1}{2}$ and $u = M + \frac{1}{2}$. Therefore, the range of u to be be warped is $\frac{1}{2} \leq u \leq M + \frac{1}{2}$.

Frequency warping functions on u may be obtained through the use of a normalized frequency variable λ 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 affine transformations:

$$u \rightarrow \lambda = \frac{u - 1/2}{M}, \quad \frac{1}{2} \leq u \leq M + \frac{1}{2} \tag{23}$$

$$\lambda \rightarrow u = \frac{1}{2} + \lambda M, \quad 0 \leq \lambda \leq 1 \tag{24}$$

Let $\theta_p(\lambda)$ be a normalized FW function controlled by parameter(s) p (see Equations 30, 31 and 32 for examples). The only practical constraint required for $\theta_p(\lambda)$ to be usable is that $0 \leq \theta_p(\lambda) \leq 1$ for $0 \leq \lambda \leq 1$. Then we can obtain a warping $\psi(u) = \psi_p(u)$ on u, using

$$\psi_p(u) = \frac{1}{2} + M \cdot \theta_p\left(\frac{u - 1/2}{M}\right) \tag{25}$$

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

By Equation 25,

$$\frac{2\psi_p(u) - 1}{2M} = \theta_p\left(\frac{2u - 1}{2M}\right) \tag{26}$$

and the warped IDCT matrix of Equation 20 can be rewritten as:

$$\tilde{C}_p = \left[\alpha_k \cos\left(\pi k \; \theta_p\left(\frac{2m - 1}{2M}\right)\right)\right]_{\substack{1 \leq m \leq M \\ 0 \leq k \leq N-1}} \tag{27}$$

Comparing Equations 21 and 22 with Equation 6, we see that the warping of the log Mel spectrum has been embedded into the IDCT matrix. In fact, if we let $\lambda_m = \frac{2m-1}{2M}$ for $1 \leq m \leq M$, then Equations 10 and 27 may be rewritten as:

$$C^T = [\ \alpha_k \cos(\pi k \ \lambda_m)\ ]_{\substack{1 \leq m \leq M \\ 0 \leq k \leq N-1}} \tag{28}$$

$$\tilde{C}_p = [\ \alpha_k \cos(\pi k \ \theta_p(\lambda_m))\ ]_{\substack{1 \leq m \leq M \\ 0 \leq k \leq N-1}} \tag{29}$$

This last equation shows clearly the simplest way of computing the warped IDCT matrix for a given normalized warping function $\theta_p(\lambda)$ and warping parameter p. We next look at some examples for $\theta_p(\lambda)$.

Examples of Normalized Frequency Warping Functions:

(1) *Piecewise Linear:* These are the type of FW functions that are commonly used in VTLN (Wegmann et al., 1996; Pitz et al., 2001).

$$\theta_p(\lambda) = \begin{cases} p\lambda, & 0 \leq \lambda \leq \lambda_0 \\ p\lambda_0 + \left(\frac{1-p\lambda_0}{1-\lambda_0}\right)(\lambda - \lambda_0), & \lambda_0 < \lambda \leq 1 \end{cases} \tag{30}$$

where λ_0 is a fixed reference frequency, around 0.7 in our experiments.

(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 (Cui and Alwan, 2006; Panchapagesan, 2006).

For $p \leq 1$,

$$\theta_p(\lambda) = p\lambda, \ 0 \leq \lambda \leq 1 \tag{31}$$

(3) *Sine-Log Allpass Transforms (SLAPT):* SLAPT frequency warping functions introduced in McDonough (2000), are capable of approximating any 1-1 arbitrary frequency warping function, and are therefore suitable for multi-class adaptation or the adaptation of individual distributions. The K-parameter SLAPT, denoted SLAPT-K, is given by:

$$\theta_p(\lambda) = \lambda + \sum_{k=1}^{K} p_k \sin(\pi k \lambda) \tag{32}$$

3.3 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 (Claes et al., 1998; McDonough and Byrne, 1999; Cui and Alwan, 2006):

$$\widehat{\Delta\mathbf{c}} = T_p \ \Delta\mathbf{c} \tag{33}$$

$$\widehat{\Delta^2\mathbf{c}} = T_p \ \Delta^2\mathbf{c} \tag{34}$$

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

$$\mathbf{x}^p = A_p\,\mathbf{x}, \quad \text{where} \quad A_p = \begin{bmatrix} T_p & 0 & 0 \\ 0 & T_p & 0 \\ 0 & 0 & T_p \end{bmatrix} \tag{35}$$

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 (Claes et al., 1998; McDonough and Byrne, 1999; Cui and Alwan, 2006):

$$\hat{\mu} = A_p\ \mu \tag{36}$$

3.3.1 Combination with MLLR Bias and Variance Adaptation

After estimating the LT (see Section 5 below), a bias vector b and an *unconstrained* variance transform matrix H may be estimated according to the Maximum Likelihood Linear Regression (MLLR) technique (Leggetter and Woodland, 1995; Gales, 1996). The adapted mean and covariance matrix $\{\hat{\mu}, \hat{\Sigma}\}$ of a Gaussian distribution $\{\mu, \Sigma\}$ are given by:

$$\hat{\mu} = A_p\ \mu + b \tag{37}$$

$$\hat{\Sigma} = B^T H B \tag{38}$$

where $\Sigma = CC^T$ and $B = C^{-1}$.

The MLLR formulae for estimating the bias and variance transforms are (Gales, 1996; McDonough, 2000; Cui and Alwan, 2006):

$$b = \left(\sum_g \sum_u \sum_t \gamma_{gut} \Sigma_g^{-1}\right)^{-1} \left(\sum_g \sum_u \sum_t \gamma_{gut} \Sigma_g^{-1} (\mathbf{x}_{ut} - A_p \mu_g)\right) \tag{39}$$

$$H = \frac{\sum_g C_g^T \left[\sum_u \sum_t \gamma_{gut} (\mathbf{x}_{ut} - \mu_g)(\mathbf{x}_{ut} - \mu_g)^T\right] C_g}{\sum_g \sum_u \sum_t \gamma_{gut}} \tag{40}$$

In the above equations, g is summed over the Gaussian distributions that are being transformed together, u is summed over the set of adaptation utterances and t is the time index over a given adaptation utterance u. γ_{gut} is the posterior probability that a speech frame $\mathbf{x}_{ut}$ was produced by Gaussian g, for the given transcription of the adaptation data. In the case of diagonal covariance matrices, the off-diagonal elements of H from Equation 40 above are simply ignored and zeroed out.

4 Comparison and relationships with previous transforms

As discussed in Section 1, several cepstral linear transforms have earlier been derived in the literature as equivalents of frequency warping for use in speaker normalization and adaptation. Some of them were derived for plain or PLP cepstra (McDonough et al., 1998; Pitz et al., 2001) and extended to non-standard MFCC features (Pitz and Ney, 2003; Umesh et al., 2005). Although our LT was derived for standard MFCCs by warping the log filterbank output, motivated by the work of Cui and Alwan (2006), it is closely related to the earlier transforms for cepstral features.

In fact, we have verified that for the SLAPT-1 warping function, the different cepstral LTs (McDonough's, Umesh et al.'s and ours) are numerically identical except in the first row, up to numerical accuracy in Matlab. Since this is not readily apparent from their mathematical formulations, we now wish to clarify the relationships between these different cepstral linear transforms for frequency warping. We first briefly describe the assumptions and formulae involved in the calculation of the LTs of McDonough, Pitz et al. and Umesh et. al., and then compare them with our LT.

4.1 McDonough's LT

McDonough derived his LT using the strict definition of cepstra as Laurent series coefficients of the log spectrum (see McDonough et al., 1998, , for example). With this definition, the LT can be computed for analytic transformations that preserve the unit circle in the complex plane, such as the rational and sine-log all-pass transforms (RAPT and SLAPT). If $Q(z)$ is the warping transformation, then the transformation matrix is given by:

$$a_{nm} = \begin{cases} 1 & \text{for } n = 0, m = 0 \\ 2q^{(m)}[0], & \text{for } n = 0, m > 0 \\ 0, & \text{for } n > 0, m = 0 \\ q^{(m)}[n] + q^{(m)}[-n], & \text{for } n > 0, m > 0 \end{cases} \tag{41}$$

where $q^{(m)}[n]$ are obtained from $q[n]$ using $q^{(m)}[n] = q^{(m-1)}[n] * q[n], m \geq 1$, with $q^{(0)}[n] = \delta[n]$, the unit sample sequence. This matrix differs in the first row from the one given in McDonough et al. (1998), since that was for the causal minimum-phase cepstra ($x[n]$ in McDonough et al., 1998), while this is for the plain real cepstra ($c[n]$ in McDonough et al., 1998).

Since we will later compare the computations involved in our LT with that of McDonough's, we now briefly list the steps involved in calculating McDonough's LT. For the K-parameter SLAPT FW,

$$Q(z) = zG(z) = z \exp F(z) \tag{42}$$

where

$$F(z) = \left(\frac{\pi}{2}\right) \sum_{k=1}^{K} \alpha_k (z^k - z^{-k}) \tag{43}$$

If $f^{(m)}[n]$ are defined using $f[n]$, similar to $q^{(m)}[n]$ using $q[n]$ above, then

$$g[n] = \sum_{m=0}^{\infty} \frac{1}{m!} f^{(m)}[n] \tag{44}$$

and

$$q[n] = g[n-1], n = 0, \pm 1, \pm 2, \ldots \tag{45}$$

The transformation matrix can then be calculated as shown above in Equation 41. The matrix is, in theory, doubly-infinite-dimensional.

4.2 Pitz et al.'s LT

Pitz et al. (2001) used the definition of cepstra as inverse discrete-time Fourier transform (IDTFT) coefficients of the log power spectrum to derive their cepstral LT. The transformation matrix was shown to be:

$$a_{nm} = \frac{2}{\pi} \int_0^{\pi} \cos(\omega n) \cos(\phi(\omega) m) d\omega \tag{46}$$

where $\phi(\omega)$ is a warping function on ω.

By comparing their derivation with that of McDonough's, it becomes clear that the derivations are equivalent except that in Pitz et al. (2001), all the complex integrals have been performed on the unit circle, and the assumption is made that the original unwarped cepstra are quefrency limited. For APT FW functions, Pitz et al.'s LT would therefore be identical to McDonough's LT. Note that this is theoretically true even though it may not be possible to evaluate the above integral anlytically for the APT FW function. It has been numerically verified as discussed below. Interestingly, this has not been noted in the literature.

With Pitz et al.'s treatment of cepstra as the IDTFT of the log spectrum, non-analytic FW functions like the popular piecewise-linear (PL) FW can also be used, while such functions cannot be used with McDonough's LT since they would not result in valid cepstra according to his stricter definition of cepstra as Laurent series coefficients of a function analytic in an annular region that includes the unit circle.

4.3 Umesh et al.'s LT

The integral involved in the computation of Pitz et al.'s LT (Equation 46) can be analytically evaluated only for some simple cases such as the linear and PL FWs.

Umesh et al. (2005) showed that a discrete approximation of the integral would become exact under the assumption of quefrency limitedness of cepstra. In this case, we can show that the LT matrix is given by

$$A = C_1 \tilde{C}_{1p} \tag{47}$$

where C_1 is a type-I DCT matrix, $\tilde{C}_{1p}$ is a type-I warped IDCT matrix, and p are FW parameters. Note that this specific expression was formulated by us and is equivalent to the one given in Umesh et al. (2005) where IDFT and warped DFT matrices have been used. From our formulation it is seen more clearly by comparing Equations 46 and 47 that Umesh et al.'s matrix is a discrete version of Pitz et al.'s.

Umesh et al.'s approach is still only an approximation since it involves the assumptions of quefrency limitedness of both the unwarped and warped cepstra. This assumption cannot be valid since it can be seen from McDonough's and Pitz et al.'s derivation, that even if the original cepstra were quefrency limited, the transformed cepstra would not necessarily be. However, it is a very good approximation, and we have verified that for the SLAPT-1 FW function, Umesh et al's matrix (Equation 47) is numerically identical to that of McDonough's (Equation 41) up to numerical accuracy in Matlab. This has also not been noted earlier in the literature.

Umesh et al. (2005) applied their LT derived for FW with plain cepstra, to a non-standard MFCC feature extraction scheme with a modified filterbank whose filters were uniformly spaced in the linear frequency domain, but of uniform bandwidth in the Mel domain. Their formulae for computing Mel and VTLN warped cepstral coefficients were complicated by the use of two different DCT matrices C_1 and C_2. We can show that their warping transformation matrix for MFCCs is:

$$T = C_2 C_1 \tilde{C}_{1p} C_2^{-1} \tag{48}$$

where C_2 is a type-II DCT matrix.

4.4 Our LT

We have expressed the equation for our LT in Equation 22. To be clearer, we may write it as:

$$T = C_2 \tilde{C}_{2p} \tag{49}$$

where C_2 is a type-II DCT, $\tilde{C}_{2p}$ is a type-II warped IDCT matrix, and p are FW parameters. We have given compact formulae for calculating C_2 and $\tilde{C}_{2p}$ in Equations 28 and 29.

We now see that there is a close relationship between our LT and McDonough-Umesh's LT for plain cepstra. In fact, though different types of DCT matrices have been used in our LT and Umesh's LT, because of the combination of DCT and warped IDCT matrices in both, the final transform matrices are identical in all rows except the first. This, however is only numerically true for values of M (the number of filters) that are not small. In our experiments, we used a value of $M = 26$ for computing our LT and $M = 256$ for computing Umesh et al.'s LT.

It therefore follows from the previous discussion of Umesh et al.'s transform, that for the SLAPT-1 FW, except for the first row, our LT is also an approximation of McDonough's LT. Note that the version of McDonough's LT for minimum-phase cepstra is different from both Umesh's LT and our LT in the first row.

Our approach has two advantages over McDonough's and Umesh et al.'s:

- Our LT (and Umesh et al.'s LT) can be calculated using compact closed form expressions for any FW function as in Equations 22, 28 and 29, unlike McDonough's original LT which is more complicated to calculate since it requires approximate summation of an infinite series and several iterations of discrete sequence convolution as in Equations 41 to 45. If the computation of derivatives during optimization of the objective function is also considered, the closed-form formulae would be even more convenient.
- By using a warped type-II IDCT, we have applied our LT directly to standard MFCC features, without modifying the feature extraction like Umesh et al. (2005) have done. Comparing our linear transform in Equation 49 with that of Umesh et al. in Equation 48, it is clear that our linear transform matrix for MFCCs is mathematically simpler and easier to calculate.

4.5 Other LTs for standard MFCCs

Claes et al. (1998) and Cui and Alwan (2006) derived transforms for standard MFCCs which were discussed in some detail in Section 2. As shown there, Cui and Alwan's transform is a special case of Claes et al.'s transform, but is mathematically simpler. It was also found to give better recognition results in practice. In Section 2, we motivated our proposal to perform continuous warping of the log filterbank output based on the success of the transform in Cui and Alwan (2006) which was basically a discrete mapping on the log filterbank outputs. In Cui and Alwan (2006), the FW was estimated in the linear frequency domain by alignment of formant like peaks, hence the name Peak Alignment (PA) for their method. In Section 6.3, we show that when the MLS criterion is used to estimate the FW parameter, our LT gives better performance than the LTs of Claes et al. and Cui and Alwan. This is an advantage of our method since estimatation of FW parameters directly using an MLS or other objective criterion would eliminate the need for access to the intermediate linear frequency spectrum during feature extraction, and the estimation can be performed entirely using just the previously extracted unwarped features.

Computationally, it is difficult to compare FW estimation using Peak Alignment with

MLS estimation. The most expensive part of using the MLS criterion to estimate a speaker specific warp factor, is the Viterbi forced alignment of frames and HMM states for the adaptation data, which may be performed for each warp factor, or once with unwarped features in the simplified criterion. Forced alignment with a known transcription of the adaptation data can be performed much faster than ASR decoding. Since the forced alignment and likelihood computation algorithms are part of the ASR decoder, MLS does not require any programs other than those already available, which may be useful in some applications. In PA, the EM algorithm is used to fit Gaussian mixtures to the linear frequency DFT spectrum and the formant-like peaks are estimated from these Gaussians for each frame of voiced speech. There, it is necessary to detect voicing and to specify the number of peaks used to fit the spectrum, which may depend on the age and gender of the test speaker and also the bandwidth of the speech signal used in the recognizer. With the MLS criterion, these considerations are not necessary and the FW estimation is automatic and robust for any test speaker.

One possible criticism of our approach may be that it is necessary to justify the IDCT approximation used to obtain the continuous log filterbank output that is warped. The IDCT approximation is essentially equivalent to the assumption of quefrency limitedness of the cepstra. However, this issue is common to all linear transform approaches to frequency warping when used in practice, since the linear transform needs to be applied to the available cepstra, which are necessarily truncated for practical purposes. The idea is to make the best use of the available spectral information in the truncated cepstra to perform frequency warping and adaptation. Our approach in developing the linear transform for truncated standard MFCCs by warping the smoothed log Mel filterbank output, is also similar. Also, as argued in Umesh et al. (2005), the smoothing performed by the filterbank also contributes towards quefrency limitedness. The final justification of our approach would be the ease of use of the resulting linear transform for standard MFCCs, and the successful results that are obtained with the method, which will be demonstrated in Section 6.

5 Estimation of the FW function

5.1 MLS Objective Criterion

In our work, for VTLN estimation, we used the commonly used maximum likelihood score (MLS) criterion (Lee and Rose, 1998; Pitz et al., 2001). For a feature space transform, the MLS criterion to estimate the optimal FW parameters $\hat{p}$ is:

$$\hat{p} = \arg\max_{p} \left[\log P(\mathbf{X}^p, \Theta^p | W, \Lambda) + T \log |A_p|\right] \tag{50}$$

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.

The determinant term in Equation 50 is required to properly normalize the likelihood when the feature space is transformed. For regular VTLN by Mel bin center frequency warping (Lee and Rose, 1998), the objective function only includes the first term in Equation 50 since the second term is not defined. In our experiments with the Linear Transformation, the determinant term was found to be important during training with Speaker Adaptive Modeling (SAM, see Section 6.4), but was not used in testing, since slightly better results were obtained without it.

Since the Viterbi re-alignment of utterances for each warping factor is computationally expensive, the MLS criterion is usually simplified by obtaining a frame-state alignment for the adaptation data once with unwarped features and then maximizing the likelihood with a fixed alignment to estimate the warping parameters p (Zhan and Waibel, 1997). The simplified MLS objective function is:

$$\mathcal{F}(p) = \sum_{t=1}^{T} \log\left(\sum_{m=1}^{M} c_{tm}\mathcal{N}(\mathbf{x}_t^p; \mu_{tm}, \Sigma_{tm})\right) + T\log|A_p| \tag{51}$$

where $\sum_{m=1}^{M} c_{tm} = 1$ for the mixture Gaussian state output distribution at time t. A gradient search or quasi-Newton method may be used to optimize the simplified MLS objective function for multiple FW parameters (Panchapagesan and Alwan, 2006).

The MLS criterion can also to be used to estimate LT FW to transform the means of the HMMs in the back end as in Equation 36:

$$\hat{p} = \arg\max_{p} \left[\log P(\mathbf{X}, \Theta^p | W, \Lambda^p)\right] \tag{52}$$

where the variables are as explained above for Equation 50 except that here it is not the adaptation data but the HMMs Λ that are modified to Λ^p for FW parameters p.

5.2 The EM Auxiliary Function

The FW parameters can also be estimated by maximizing the EM auxiliary function over the adaptation data (McDonough, 2000; Loof et al., 2006). This objective function is identical to the one used for MLLR and CMLLR (constrained MLLR, Gales, 1998), except the linear transformation to be estimated is constrained by the FW parametrization. Speaker Adaptive Training (SAT) also uses iterative maximization of the EM auxiliary function to alternately estimate FW parameters and HMM parameters (Anastasakos et al., 1996).

Here we consider only estimation of a feature transform, which we denote CLTFW similar to CMLLR. The basic auxiliary function to be *minimized* may be expressed as:

$$\mathcal{F}(p) = \frac{1}{2} \sum_{g} \sum_{t} \gamma_g(t) \left[(A_p \mathbf{x}_t - \mu_g)^T \Sigma_g^{-1} (A_p \mathbf{x}_t - \mu_g) - \log(|A_p|^2) \right] \tag{53}$$

where g varies over the set of Gaussian distributions for which the transform is to be estimated, t is time or frame index of the adaptation data, and $\gamma_g(t)$ is the posterior probability that feature frame $\mathbf{x}_t$ was generated by Gaussian g for the given transcription of the adaptation utterances.

For diagonal covariance models, this can be simplified to:

$$\mathcal{F}(p) = \frac{1}{2} \sum_{i=1}^{d} \left[a_i G^{(i)} a_i^T - 2 a_i k^{(i)T} \right] - \beta \log(|A_p|) \tag{54}$$

where d is the feature vector size, a_i is the ith row of A_p, and

$$G^{(i)} = \sum_{g} \frac{1}{\sigma_{gi}^2} \sum_{t} \gamma_g(t) \mathbf{x}_t \mathbf{x}_t^T \tag{55}$$

$$k^{(i)} = \sum_{g} \frac{\mu_{gi}^2}{\sigma_{gi}^2} \sum_{t} \gamma_g(t) \mathbf{x}_t^T \tag{56}$$

$$\beta = \sum_{g} \sum_{t} \gamma_g(t) \tag{57}$$

The computations involved in this approach are mostly during the accumulation of the statistics (i.e. computing $G^{(j)}$ and $k^{(j)}$). Once the statistics have been accumulated, the computational cost of optimizing the objective function is significantly smaller since it is twice differentiable and typically convex, and a few iterations of Newton's method are found to be sufficient to optimize it for a reasonably small number of FW parameters (10 or so). Different CLTFW transforms can also be estimated for different classes of distributions similar to CMLLR, without much increase in computations, since it is seen from Equation 54 that the accumulator values for a set of Gaussians is the sum over the individual Gaussians. The accumulator method of optimizing the EM auxiliary function for CLTFW may be extended in a very natural manner for the estimation of an aditive bias on top of the CLTFW transform.

Loof et al. (2006) also discuss briefly how this accumulator based approach may be extended to the case with a global feature space LDA/HLDA transform. The approach can also be extended to the multi-class semi-tied covariance (STC, Gales, 1999) case, as long as all the Gaussians considered for CLTFW estimation share the same STC transformation.

For the estimation of multiple FW parameters like with the SLAPT FW using the EM auxiliary function, it is efficient to use a convex optimization method. Newton's method can be used since the auxiliary function is twice differentiable (McDonough, 2000). We consider the diagonal covariance case and derive the formulae for calculating the first derivative of the objective function as follows.

Differentiating $\mathcal{F}(p)$ in Equation 54 with respect to p, we have:

$$\frac{\partial \mathcal{F}(p)}{\partial p_k} = \sum_{i,j=1}^{d} \frac{\partial \mathcal{F}(p)}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial p_k} \tag{58}$$

If we let

$$\mathcal{F}(p) = \mathcal{F}_1(p) - \beta \log(|A_p|) \tag{59}$$

where

$$\mathcal{F}_1(p) = \frac{1}{2} \sum_{i=1}^{d} \left[a_i G^{(i)} a_i^T - 2 a_i k^{(i)T} \right] \tag{60}$$

then

$$\frac{\partial \mathcal{F}(p)}{\partial A} = \frac{\partial \mathcal{F}_1(p)}{\partial A} - \beta \frac{\partial \log(|A_p|)}{\partial A} \tag{61}$$

where for a function f, $\frac{\partial f}{\partial A}$ denotes the matrix of partial derivatives $\frac{\partial f}{\partial a_{ij}}$. It can be shown (for example McDonough, 2000, Section 5.1) that

$$\frac{\partial \log(|A|)}{\partial A} = (A^{-1})^T \tag{62}$$

We have:

$$\frac{\partial \mathcal{F}_1(p)}{\partial a_i} = a_i G^{(i)} - k^{(i)} \tag{63}$$

where $\frac{\partial \mathcal{F}_1(p)}{\partial a_i}$ is the vector of partial derivatives $\frac{\partial \mathcal{F}_1(p)}{\partial a_{ij}}$. Therefore $\frac{\partial \mathcal{F}(p)}{\partial A}$, can be computed from Equations 62, 61 and 63. We also need $\frac{\partial A_p}{\partial p_k}$ to compute $\frac{\partial \mathcal{F}(p)}{\partial p_k}$ from Equation 58. We have:

$$\frac{\partial A_p}{\partial p_k} = \begin{bmatrix} \frac{\partial T_p}{\partial p_k} & 0 & 0 \\ 0 & \frac{\partial T_p}{\partial p_k} & 0 \\ 0 & 0 & \frac{\partial T_p}{\partial p_k} \end{bmatrix} \tag{64}$$

By Equation 22,

$$\frac{\partial T_p}{\partial p_k} = C \cdot \frac{\partial \tilde{C}_p}{\partial p_k} \tag{65}$$

To compute $\dfrac{\partial \tilde{C}_p}{\partial p_k}$, recall Equation 29 by which we have:

$$\tilde{C}_p(i,j) = \alpha_j \cdot \cos\left[\pi j\ \theta_p(\lambda_i)\right] \tag{66}$$

for $1 \leq i \leq M,\ 0 \leq j \leq N-1$, and where $\lambda_i = \dfrac{2i-1}{2M}$. Then,

$$\frac{\partial \tilde{C}_p(i,j)}{\partial p_k} = -\alpha_j \cdot \pi \cdot j \cdot \sin\left[\pi \theta_p(\lambda_i) j\right] \cdot \frac{\partial \theta_p(\lambda_i)}{\partial p_k} \tag{67}$$

For the frequency warping functions used (Equations 30 to 32), the derivative with respect to the parameter is easily computed. For example, for the piecewise-linear warping (Equation 30), we have:

$$\frac{\partial \theta_p(\lambda)}{\partial p} = \begin{cases} \lambda, & 0 \leq \lambda \leq \lambda_0 \\ \lambda_0 \cdot \frac{1-\lambda}{1-\lambda_0}, & \lambda_0 < \lambda \leq 1 \end{cases} \tag{68}$$

The gradient of the objective function in Eq. 51 with respect to the FW parameters p, $\nabla_p \mathcal{F}(p)$, can therefore be calculated using Equations 58 to 68.

Formulae for the Hessian matrix of second derivatives of the objective function with respect to FW parameters were also derived, and used in Newton's method for optimizing $\mathcal{F}(p)$.

6 Experimental Results

We validated the LT by testing it on connected digit and continuous speech recognition tasks and comparing the performance with that of regular VTLN by warping the filterbank center frequencies (hereafter referred to as Regular VTLN in this paper). The main advantages of using the LT over Regular VTLN, as discussed in Section 1, are computational savings and flexibility of implementation. The spectral information available during LT parameter estimation consists only of a smoothed mel-warped log spectrum, contained in the truncated cepstra in the pre-computed recognition features and the corresponding HMM means. Considerably more spectral information is available to Regular VTLN, which has access to the linear frequency spectrum of each speech analysis frame. In the results below, we therefore mainly aim to show that VTLN and adaptation using the LT, while being computationally superior and working with less available information, can give recognition performance comparable to that of Regular VTLN.

6.1 Connected Digit Recognition Experiments

In Panchapagesan (2006), the effectiveness of both front end VTLN and back end adaptation using the LT were demonstrated on connected digit recognition of children's speech using the TIDIGITS database. We wish to summarize the most important relevant conclusions here before presenting our experimental results with continuous speech recognition.

The baseline system in Panchapagesan (2006) was the same as in Cui and Alwan (2006), with phoneme models for connected digit recognition trained from 55 male speakers in the TIDIGITS database. Testing was performed on 5 boys and 5 girls, and results were obtained for adaptation with 1, 5 and 11 digits.

In the front end, LT VTLN outperformed Regular VTLN when an optimal speaker-specific warp factor for the piecewise-linear FW was estimated from the adaptation data. For the LT VTLN, the Jacobian normalization resulted in worse performance and was therefore not used. Both VTLN methods outperformed MLLR for small amounts of data (less than 11 adaptation digits).

The performance of the LT in back end adaptation of HMM Means was comparable with that of the Peak Alignment (PA) approach of Cui and Alwan (2006), and with front end LT VTLN. All these FW-based methods again outperformed MLLR for small amounts of data (less than 11 adaptation digits). As explained in Section 4, our LT may be seen as a smooth parametrization of the discrete mapping used in PA. Both the MLS criterion and alignment of median F3 and other formants are robust methods of estimating frequency warping. Therefore it is perhaps to be expected, that the performance of our LT FW estimated using the MLS criterion is comparable to that of PA when used for adaptation for HMM means. However, as we show in Section 6.3, the PA LT does not perform as well as our LT when the MLS criterion is used to estimate the FW. We have already described the main advantages of our method over PA in Sections 1 and 4.

6.2 Continuous Speech Recognition Experiments

We also performed experiments on continuous speech recognition using the Resource Management (RM1) database. The speech data was downsampled to 8000 Hz in our experiments and context dependent triphone models were trained on speech from 72 adult speakers in the speaker independent training set. All triphone HMMs contained 3 emitting states and 6 Gaussian mixtures per state. The Mel filterbank contained 26 filters, and the features vectors consisted of the first 13 MFCCs with the corresponding first and second derivatives. Cepstral Mean Subtraction (CMS) was also performed on each utterance.

Recognition experiments were performed on 50 test utterances from each of 10 speakers from the speaker dependent test data in the database. The baseline recognition accuracy was 90.16 %.

VTLN and back-end adaptation were tested with varying amounts of adaptation data to validate the effectiveness of the new linear transform in improving accuracy in continuous speech recognition. Experiments were performed with 1, 5 and 10 adaptation utterances from each test speaker. For adaptation with a single utterance, the 10 utterances marked for rapid adaptation in the RM1 database were used. For more than one adaptation utterance, ten different combinations of utterances were randomly selected for each speaker and results were obtained for each combination of adaptation utterances using each of the adaptation techniques. The results were then averaged over the adaptation combinations and the speakers. The pool of adaptation utterances was separate from the set of test utterances for each speaker.

Table 1 shows the results of VTLN experiments comparing LT VTLN with Regular VTLN. A speaker-specific warp factor for the piecewise-linear (PL) FW was estimated from the adaptation data for each test speaker, using a grid search to optimize the MLS criterion of Section 5. The warping factor step size in the grid was 0.01. It was again observed that slightly better results were obtained without the Jacobian Normalization term in the MLS criterion during the estimation of the parameter for LT VTLN and these are the results shown. With LT VTLN, the PL FW gave slightly better results than the linear and the SLAPT-1 FWs.

The performance of LT VTLN is seen to be comparable to that of Regular VTLN.

	No. of adaptation utterances		
Algorithm	1	5	10
LT VTLN	91.46	91.59	91.54
Regular VTLN	91.42	91.60	91.66

Table 1
Recognition Accuracy in VTLN Experiments using the RM1 database. FW parameters were estimated with the MLS criterion for both methods. Baseline Accuracy: 90.16 %

In Figure 3 sample discrete log filterbank outputs, before and after warping with LT and Regular VTLN are shown. The speech frame is from the triphone 'S-AH+B' in the word 'sub'. The features of the utterance were normalized with the corresponding estimated PL FW parameter for each VTLN method from the particular utterance. The warped log filterbank outputs of the two VTLN methods are seen to be very similar, which explains the very similar performance seen in Table 1. This seems to imply that most of the spectral information required for VTLN is already contained in the unwarped truncated cepstra, which is why LT VTLN may be as successful as Regular VTLN.

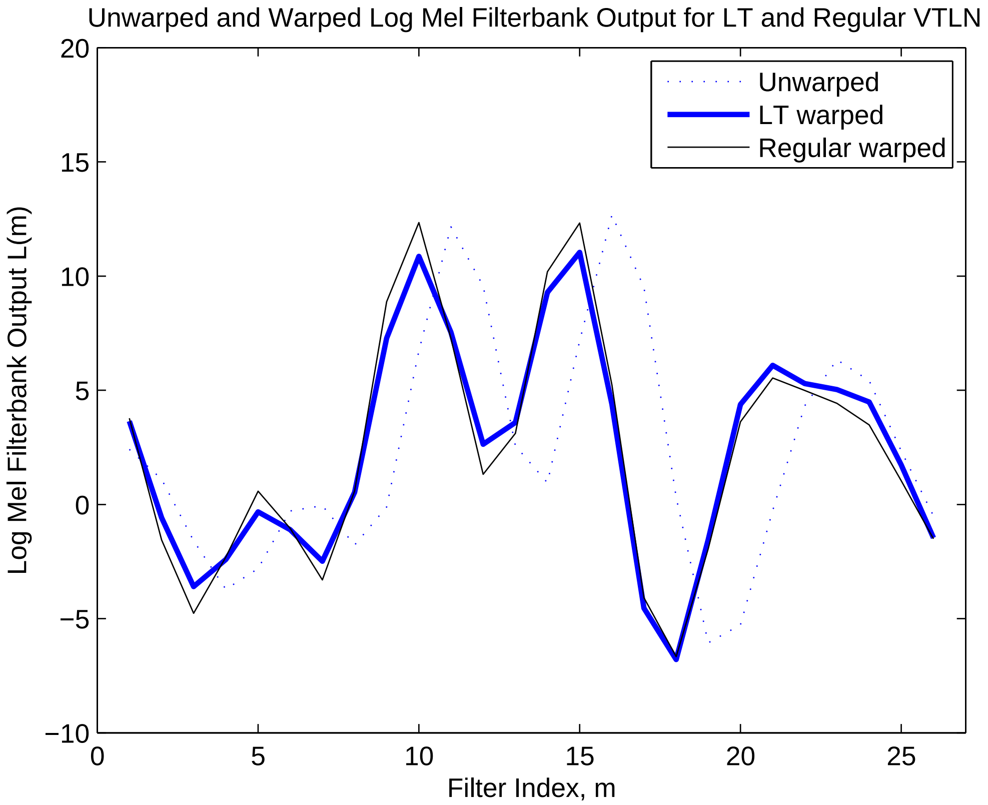

Fig. 3. Discrete log filterbank outputs, unwarped (dotted line) and warped, with LT VTLN (thick solid line) and Regular VTLN (thin solid line). The speech frame is from the triphone 'S-AH+B' in the word 'sub'.

We then performed VTLN estimation with the simplified MLS objective function as in Equation 51 of Section 5.1, with fixed frame-state alignment obtained with unwarped features. Again, the PL FW with a grid step size of 0.01 was used. The results are shown in Table 2. As can be seen, both Regular and LT VTLN have comparable results, with the results for both being slightly worse with the simplified objective function, as compared to the results in Table 1.

	No. of adaptation utterances		
Algorithm	1	5	10
LT VTLN	91.33	91.33	91.33
Regular VTLN	91.29	91.28	91.34

Table 2
Recognition Accuracy in VTLN Experiments with Fixed Frame-State Alignment. Baseline Accuracy: 90.16 %

Table 3 shows the results of global speaker adaptation experiments on the RM1 database. The LT was used to adapt HMM Means as in Equation 36, and is combined

Algorithm	**No. of adaptation utterances**		
	1	5	10
Back End LT FW + MLLR bias & var	91.58	91.74	91.76
MLLR	84.89	92.38	92.43

Table 3
Recognition Accuracy in Global Speaker Adaptation Experiments with limited data on the RM1 database: LT Applied in the back-end and 3-block MLLR. Baseline Accuracy: 90.16 %

with MLLR mean bias and unconstrained variance transforms as described in Section 3.3.1. The results of standard MLLR with a 3-block mean transformation matrix and unconstrained variance transformation are also shown for comparison. Comparing Tables 3 and 1 we see that back end HMM mean adaptation with the LT combined with unconstrained MLLR bias and variance adaptation, gives results comparable to VTLN in the front end. The results confirm earlier observed trends (Cui and Alwan, 2006; McDonough, 2000) that FW based methods are definitely superior to MLLR for very limited adaptation data (1 utterance), where MLLR actually gives worse performance than the baseline. With increased adaptation data, MLLR gives better performance.

6.3 Comparison with other LT approximations of VTLN for standard MFCCs

As discussed in Section 2, Claes et al. (1998) and Cui and Alwan (2005, 2006) have earlier proposed linear transforms for approximating VTLN with standard MFCC features. In Table 4 we show results comparing our LT with those of Cui and Alwan's Peak Alignment (PA) LT, and Claes et al's LT. The recognition results shown are on the RM database with VTLN estimated on 1 utterance, since it is desirable in practice to estimate the VTLN parameter with limited data. The MLS criterion was used to estimate the PL FW parameter for all methods. The results of Regular VTLN are also shown.

It is seen that our LT performs as well as Regular VTLN, while the PA LT and Claes et al.'s LT do not perform as well, when the FW parameter is estimated using the MLS criterion with 1 utterance. The parametrization of the transform is therefore very important since it determines the behavior of the objective function and performance of the VTLN parameter estimated using the criterion.

As we have discussed in Section 4, our LT is numerically almost identical to McDonough's and Umesh et al.'s LTs, except in the first row. Therefore, the performance of these LTs was very similar to that of our LT.

Algorithm	**Recognition Accuracy, %**
Baseline	90.16
Regular VTLN	91.42
Our LT VTLN	91.46
PA LT	90.82
Claes et al.'s LT	90.79

Table 4
Comparison of different LT approximations for VTLN with MFCC features, on the RM1 database. FW parameters were estimated on 1 utterance with the MLS criterion for all methods.

6.4 Speaker Adaptive Modeling Experiments

It is well known that the effectiveness of VTLN is greatly improved when it is performed also during training (McDonough, 2000; Welling et al., 2002). In this way, the trained models capture more of the phonetic variability and less of the inter-speaker variability in the training data. Speaker Adaptive Modeling (abbreviated as SAM here, Welling et al., 2002) and Speaker-Adapted Training (SAT, Anastasakos et al., 1996; McDonough, 2000) are two techniques for incorporating VTLN during the training process.

We first performed VTLN during training along the SAM framework. The main feature of this technique is that the optimal warping factor for each training speaker is selected iteratively using single Gaussian mixture HMMs and the MLS criterion. Initial models are trained without any warping, and then at each iteration the optimal warping factor for each speaker in the training set is obtained by MLS over the training data from that speaker, and models are retrained with the new warping factors. The use of single Gaussians mixtures during the iterative warp factor estimation is important because that gives the best results. After a certain number of iterations or when the warping factors converge, the final models are trained with the best warping factor for each speaker, and with the desired number of Gaussians per mixture.

Ten iterations were performed during SAM VTLN parameter estimation with the PL FW for both Regular and LT VTLN. One important observation was that when the Jacobian Normalization (JN, see Section 5.1) term was not included in the MLS objective function, the performance of the LT was very poor, even worse than without any SAM. This was investigated and it was found that the warping factor did not converge during the iterations, and the mean warping factor (which should presumably be close to 1, the initial value corresponding to no warping) continuously decreased to around 0.93 in ten iterations without the JN term. After including the JN term in the warping parameter estimation, the training speakers' warping factors

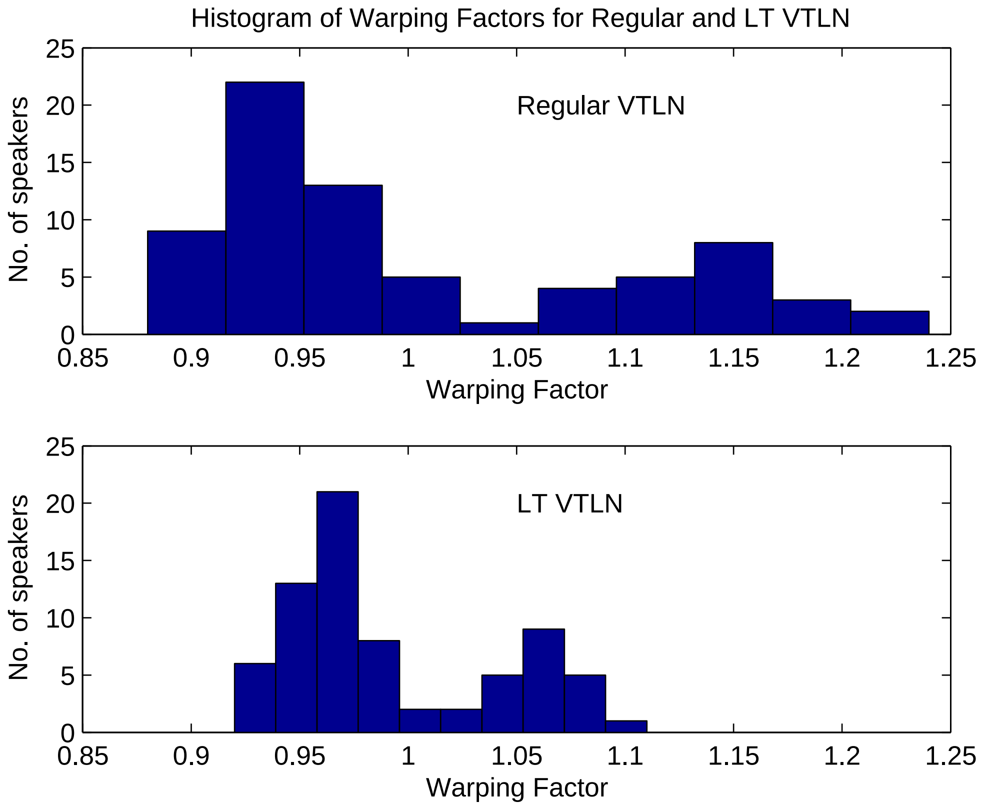

Fig. 4. Histograms of warping factors in Speaker Adaptive Modeling, for Regular and LT VTLN.

were observed to converge, and the mean value at the end of ten iterations was around 0.99. However, during testing, it was again observed that slightly better results were obtained without the JN term in the MLS estimation and these are the results that are shown.

The histograms of estimated warping factors of the 72 training speakers for both Regular VTLN and LT VTLN with the PL FW are shown in Figure 4. For each VTLN method, ten bins over the corresponding ranges of warping factor were used for calculating the histogram, but both histograms are plotted over the same range of warping factors, from 0.85 to 1.25, for comparison. It is observed that the range of the warping factors for LT VTLN is significantly smaller than that of Regular VTLN, probably due to the fact that warping in LT VTLN is being performed on an already Mel warped log spectrum.

The results of SAM VTLN experiments are shown in Table 5.

We first notice that when SAM is performed, the baseline accuracy is lower than without SAM, but once the test speaker is normalized, the accuracy is significantly better than without SAM.

The performances of the two VTLN methods are comparable when normalization is

	No. of adaptation utterances		
Algorithm	0	1	5
Regular VTLN	86.82	92.81	93.07
LT VTLN	86.82	92.64	92.79
Back End LT FW + MLLR Bias & Var.	86.82	92.87	93.31

Table 5
Recognition Accuracy in SAM VTLN Experiments using the RM1 database. 10 iterations of warping factor estimation were performed for each VTLN method for the training speakers and testing was performed with the corresponding method. The baseline with SAM models was the same (86.82 %) for both Regular and LT VTLN.

performed also during training. The important results here are those for adaptation with 1 utterance, since MLLR would be preferred when more utterances of adaptation data is available. Here, the difference in accuracies is small, around 0.17% absolute. However, better results were obtained with back end LT FW combined with MLLR bias and variance adaptation, tested on models trained with LT VTLN, which are also shown in Table 5.

Therefore, in all cases, results comparable to Regular VTLN can be obtained with the LT, by applying it in the back end instead of the front end.

We have also verified that with a global Semi-Tied Covariance (STC) matrix included, the performance of LT VTLN SAM models tested with LT VTLN is still comparable to that of Regular VTLN SAM models tested with Regular VTLN.

6.5 Speaker Adaptive Training Experiments

We also implemented SAT with feature space LT which we denoted CLTFW similar to CMLLR (constrained MLLR which is equivalent to feature space MLLR) and tested it on the RM1 database. CLTFW parameters were estimated by optimizing the EM auxiliary function as discussed in Sections 5.2 and 5.3. SAT uses the iterative maximization of the EM auxiliary function to jointly estimate speaker transforms and HMM parameters. Ten iterations of SAT were performed with global LT and the PL FW on single mixture HMMs and the final single-mixture SAT speaker transforms were used to retrain 6-mixture HMMs using the baseline models and single-pass retraining. Multiple iterations of model re-estimation were then performed keeping the transforms fixed.

We tested the CLTFW SAT models with CLTFW adaptation with 1, 5 and 10 utterances, and the recognition results are shown in Table 6.

Train/Test Conditions	**No. of adaptation utterances**		
	1	5	10
G-CLTFW PL SAT / G-CLTFW PL	92.82	92.91	92.94
G-CLTFW PL SAT / RC CLTFW (SLAPT-5)	92.82*	93.03	93.31
G-CLTFW PL SAT / RC CLTFW (SLAPT-5) +Bias	92.82*	93.30	94.07

Table 6
Recognition Accuracy in Global (G-) CLTFW SAT Experiments with the PL FW using the RM1 database. 10 iterations of SAT warping factor estimation were performed for the training speakers. * indicates insufficient data to estimate further transforms.

It is seen that when the Global (G-) CLTFW SAT models were tested with G-CLTFW, the performance was comparable to that obtained with VTLN SAM (refer Table 5), and the performance saturates for larger number of utterances. However, improved results for more adaptation data were obtained when multiple parameter SLAPT-5 CLTFW was estimated for multiple classes using a regression tree. A frame count threshold of 400 for estimating a transform at a regression node was found to be effective. During estimation, 5 iterations of CLTFW parameter estimation were performed on a single utterance to first estimate a global PL CLTFW transform (similar to VTLN estimation), and this global transform was used to obtain alignments for two iterations of multi-class SLAPT-5 CLTFW estimation. It is seen that the performance of multi-class CLTFW improves with more data. An additive bias was included in the transform, and the performance improved significantly.

Therefore, multiple parameter SLAPT-5 CLTFW-Bias transforms estimated using the EM auxiliary function and a regression tree, can give significantly better performance than global VTLN, and improving performance with increasing data.

Since Regular VTLN is not a non-invertible operation on standard MFCCs, the Jacobian determinant term required in the EM auxiliary function for SAT cannot be computed (McDonough, 2000; Sankar and Lee, 1996). Also, even if the Jacobian determinant term were neglected, the accumulator based approach (Gales, 1998) for efficient optimization of the EM auxiliary function with CLTFW cannot be used with Regular VTLN. For multiple class adaptation to be performed with Regular VTLN, features would have to be recomputed with different warping factors for different distributions. As we have shown, recomputation of features is expensive and this is not practical.

Experiments with multi-class CLTFW SAT and comparisons and combination with HMM mean adaptation (MLLR for example) and LDA/STC would be the topic of

Algorithm	No. of adaptation utterances	
	1	5
LT VTLN	92.63	92.86
Back End LT FW + MLLR Bias & Var.	92.75	93.16

Table 7
Recognition Accuracy in Unsupervised VTLN and Adaptation Experiments on the RM1 database using models trained with LT Speaker Adaptive Modeling. Baseline Recognition Accuracy is 86.82 %

future work.

6.6 Unsupervised Adaptation

We have so far given the results of supervised adaptation experiments, where the transcription of the adaptation data is known. Frequency warping methods are known to be effective in adaptation in an unsupervised mode as well (McDonough, 2000; Cui and Alwan, 2006). This was confirmed for VTLN and back end model adaptation using our LT, for the case of the speaker adaptive models trained as discussed in the previous section. The results are shown in Table 7. In these experiments, an initial recognition pass was first performed over the adaptation data, and the resulting transcriptions were then used to estimate the FW parameter using the MLS criterion and the MLLR mean bias and variance transforms.

Comparing Tables 5 and 7, it is seen that the results of unsupervised LT VTLN are not much different from those of supervised LT VTLN. In fact, the warping factors estimated with supervised and unsupervised adaptation were only slightly different. This is probably because of our already high baseline recognition accuracy where the transcription produced by the initial recognition pass is close to the actual transcription. With a worse baseline, one may have to use confidence measures calculated from the likelihoods obtained with the initial recognition pass, to select a subset of the adaptation data for warping factor estimation. However, since the VTLN parameter estimated with very little data also performs well (as seen from our earlier results with the TIDIGITS database), the LT would be very effective in unsupervised adaptation.

7 Summary and Conclusions

We have developed a novel linear transform (LT) to implement frequency warping for standard filterbank based MFCC features. There are a few important advantages of using a linear transform for frequency warping: VTLN estimation by optimizing an objective function is performed computationally more efficiently with a LT than with regular VTLN by warping the center frequencies of the filterbank; the transform can also be estimated and applied in the back end to HMM means; and one need not have to access to or reconstruct intermediate variables like the linear frequency spectrum in order to apply the FW, which would be useful in distributed speech recognition.

The main idea of our approach was to directly warp the smoothed log Mel spectrum obtained by cosine interpolation of the log Mel filterbank output with the IDCT. This results in a linear transformation in the Mel cepstral domain. The warping was parametrized and embedded into a warped type-II IDCT matrix, which can be easily calculated using a compact formula. Our LT for MFCCs was also shown to be closely related to the plain cepstral LTs of McDonough, Pitz et al. and Umesh et al. In fact, these LTs for FW are all found to be numerically almost identical for the sine-log all-pass transform (SLAPT) warping functions. Our formula for the transformation matrix is, however, computationally simpler and unlike other previous linear transform approaches to VTLN with MFCC features (Pitz and Ney, 2003; Umesh et al., 2005), no modification of the standard MFCC feature extraction scheme is required. The parameters of the warping are easily estimated by maximum likelihood score (MLS) or the EM auxiliary function, using the commonly used grid search or convex optimization methods for multiple FW parameters. Formulae for calculating the gradient of the EM auxiliary function with respect to the warping parameters were also therefore derived.

The Linear Transform (LT) had previously been validated on connected digit recognition of children's speech with the TIDIGITS database (Panchapagesan, 2006). In this paper, we have presented extensive results on continuous speech recognition with the Resource Management (RM1) database. In VTLN and VTLN Speaker Adaptive Modeling (SAM) experiments with the RM1 database, the performance of the new LT VTLN was comparable to that of Regular VTLN. For the LT, the inclusion of the Jacobian normalization term in the MLS criterion was found to be quite important for convergence of the FW parameters during training using SAM. During testing, however, better results were obtained without the Jacobian determinant term in the MLS criterion. Our LT was also found to perform better than the earlier proposed transforms of Cui and Alwan (2005, 2006) and Claes et al. (1998) for approximate VTLN with MFCC features, when the MLS criterion was used to estimate the FW parameter. LT adaptation of HMM means combined with MLLR mean bias and variance adaptation typically gave results that were comparable to the front end VTLN methods. The FW based methods were found to be significantly better than MLLR for limited adaptation data. We also performed Speaker Adaptive Training (SAT) with feature space LT denoted CLTFW. Global PL CLTFW SAT models tested with global PL CLTFW gave results comparable to SAM and VTLN, and the performance saturates with increasing adaptation data. By estimating multiple parameter SLAPT-5 CLTFW transforms using a regression tree, and including an additive bias, we obtained significantly better performance than global VTLN, and improving results with increasing adaptation data. Warping factors estimated in an unsupervised mode were almost identical with those from supervised estimation, and therefore the performance of unsupervised VTLN and model adaptation with the LT were almost as good as with supervised VTLN and adaptation.

Our experimental results with LT VTLN are only comparable in performance to regular VTLN. Since our aim was to obtain a linear transform equivalent for VTLN with standard MFCC features, it is important to demonstrate that the involved approximations do not lead to performance degradation. It is probably also not to be expected that an approximation would perform better than the original method. By estimating multiple transforms using the EM auxiliary function and a regression tree, we have also shown that it is possible to obtain results better than global VTLN. It would be the topic of future work to compare and/or combine multi-class CLTFW with MLLR adaptation.

Though the computations required for VTLN implementation may be small compared to the overall effort for training and testing, the computational advantage of LT VTLN over regular VTLN discussed in Section 1 becomes significant when the VTLN parameter has to be estimated in real time. For example, in DSR, the computational savings during FW parameter estimation, the ability to estimate and implement VTLN directly on the features without having access to the feature extraction modules and the flexibility of application (front-end or back-end) would be a significant advantage of LT over regular VTLN. We believe that the proposed linear transform would prove very useful in practice in embedded and distributed speech recognition applications, where resources are limited.

8 Acknowledgements

We wish to thank Xiaodong Cui for the use of his connected digit recognition and adaptation systems, Shizhen Wang for discussions on this work, and Markus Iseli for help with programming. We also thank the anonymous reviewers for their comments that helped improve the paper. This work was supported in part by the NSF.

References

T. Anastasakos, J. McDonough, and J. Makhoul, “A compact model for speaker-adaptive training,” *Proc. ICASSP 1997*, pp.1043-1046.

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.

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

X. Cui and A. Alwan, “Adaptation of Children’s Speech with Limited Data Based on Formant-like Peak Alignment,” *Computer Speech and Language*, Vol. 20, Issue 4, pp. 400-419, October 2006.

S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, pp.357-366, Aug. 1980.

A. Dempster, N. Laird, and D. Rubin, “Maximum Likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society* B, vol. 39, No.1, pp.1-38, 1977.

E. Eide and H. Gish, “A parametric approach to vocal tract length normalization,” *Proc. ICASSP*, pp.346-349, 1996.

M. J. F. Gales, “Mean and variance adaptation within the MLLR framework,” *Computer Speech and Language*, Vol. 10, pp.249-264, 1996.

M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, no. 2, pp. 75 98, Apr. 1998.

M. J. F. Gales, “Semi-tied covariance matrices for hidden Markov models,” *IEEE Transactions Speech and Audio Processing*, vol. 7, pp. 272-281, 1999.

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

B-H. Juang, W. Chou and C-H. Lee, “Minimum classification error rate methods for speech recognition,”’ IEEE Trans. Speech and Audio Processing, Vol.5, No.3, May 1997.

L. Lee and R. C. Rose, “A frequency warping approach to speaker normalization,” *IEEE Trans. Speech and Audio Processing*, vol. 6, No. 1, 49-60, 1998.

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.

J. Loof, H. Ney, and S. Umesh, “VTLN Warping Factor Estimation Using Accumulation of Sufficient Statistics,” *Proc. ICASSP 2006*, vol. 1, p.1-4.

J. McDonough, W. Byrne and Luo, X., “Speaker normalization with all-pass transforms,” *Proc. ICSLP*, Vol.6 , pp. 2307-2310, 1998.

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

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

J. McDonough, T. Schaaf, and A. Waibel, “Speaker Adaptation with All-Pass Transforms”, Speech Communication Special Issue on Adaptive Methods in Speech Recognition, January, 2004.

. McDonough, and A. Waibel, “Performance Comparisons of All-Pass Transform Adaptation with Maximum Likelihood Linear Regression”, Proc. ICASSP 2004.

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

S. Panchapagesan, “Frequency Warping by Linear Transformation of Standard

MFCC", *Proceedings of Interspeech 2006, ICSLP*, pp. 397-400.

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

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

S. Umesh, A. Zolnay and H. Ney "Implementing frequency-warping and VTLN through linear transformation of conventional MFCC," *INTERSPEECH* 2005, pp.269-272.

K. Visweswariah and R. Gopinath, "Adaptation of front end parameters in a speech recognizer," *INTERSPEECH*-04, 21-24.

S. Wegmann, D. McAllaster, J. Orloff and B. Peskin, "Speaker normalization on conversational telephone speech," *Proc ICASSP*, pp.339-341, 1996.

L. Welling, H. Ney, and S. Kanthak, "Speaker Adaptive Modeling by Vocal Tract Normalization," *IEEE Trans. Speech and Audio Processing*, Vol. 10, No. 6 pp.415-426, 2002.

P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," Technical Report, CMU-CS-97-148, May 1997.

R. Zhao and Z. Wang, "Robust speech recognition based on spectral adjusting and warping," *Proc. ICASSP 2005*, vol. 1, pp.553-556.