MLLR-LIKE SPEAKER ADAPTATION BASED ON LINEARIZATION OF VTLN WITH MFCC FEATURES

Xiaodong Cui and Abeer Alwan

Department of Electrical Engineering
University of California, Los Angeles, CA 90095
Email: xdcui@icsl.ucla.edu, alwan@icsl.ucla.edu

Abstract

In this paper, an MLLR-like adaptation approach is proposed whereby the transformation of the means is performed deterministically based on linearization of VTLN. Biases and adaptation of the variances are estimated statistically by the EM algorithm. In the discrete frequency domain, we show that under certain approximations, frequency warping with Mel-£lter-bank-based MFCCs equals a linear transformation in the cepstral domain. Utilizing the deduced linear relationship, the transformation matrix is generated by formant-like peak alignment. Experimental results using children's speech show improvements over traditional MLLR and VTLN. The improvements occur even with limited amounts of adaptation data.

1. Introduction

Vocal tract length normalization (VTLN) [1][2] has been extensively used in speaker adaptation to reduce the spectral mismatch via frequency warping. Recently, the relationship between the front-end feature domain and the back-end model domain, in terms of VTLN, has drawn increasing attention [3][4]. In [3], a linear relationship is investigated in the $\mathcal{Z}$ space in the form of an all-pass transformation. In [4], the authors prove, in the continuous frequency ω space, that frequency warping equals a linear transform in the cepstral domain. However, since invertiblity is required, the MFCC features studied in [4] are computed by Mel-scaling instead of Mel-scaled £lter banks. In this paper, we investigate the above-mentioned relationships in the discrete frequency space. We show that for MFCCs computed with Mel-scaled triangular £lter banks, a linear relationship can be obtained if certain approximations are made. Utilizing that relationship as a special case of MLLR, an adaptation approach based on formant-like peak alignment is proposed where the transformation of the means is performed deterministically based on the linearization of VTLN. Biases and adaptation of the variances are estimated statistically by the EM algorithm.

The remainder of the paper is organized as follows: in Section 2, the linearization conditions of VTLN in the discrete frequency domain are discussed; in Section 3, spectral peak alignment by piece-wise linear functions is described; the adaptation scheme is described in Section 4; experimental results are presented in Section 5; and conclusions are made in Section 6.

2. Linearization of VTLN

MFCCs are the most widely-used speech features for recognition. Fig.1 illustrates how MFCCs are computed with a uniformly spaced triangular Mel-scale £lter bank. Let S^l denote a linear spectrum magnitude and S^c the corresponding MFCC features. According to the scheme shown in Fig. 1, we have:

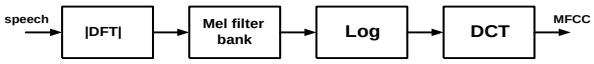

Figure 1: Diagram of MFCC feature extraction.

$$S^c = \mathbf{C} \cdot \mathbf{log}(\mathbf{M} \cdot S^l) \tag{1}$$

where $\mathbf{C}$ is the DCT matrix, **log** is the component-wise logarithm function applied to a matrix.

$$\mathbf{M} = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,K_1} & & \\ & \theta_{2,1} & \cdots & \theta_{2,K_2} & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \theta_{N,1} & \cdots & \theta_{N,K_N} \end{bmatrix}_{N \times L}$$

is the matrix form of the Mel £lter bank where L and N are the number of samples in the linear and Mel-frequency domains, respectively. $\theta_{i,j}$ is the jth weight of the ith triangular £lter and K_i represents the number of non-zero coef£cients in each £lter. In other words, each row represents the components in one £lter bin. In $\mathbf{M}$, all elements other than $\theta_{i,j}$ are zeros. Typically, N is much smaller than L.

Suppose there exists a warping function in the discrete linear frequency domain $k = \phi(l)$, where k and l are the discrete frequency sample indices. This can be represented as a frequency warping matrix $\mathbf{R}_{L \times L}$ whose components are de£ned as:

$$r_{ij} = \begin{cases} 1, & \text{if } i = \text{round}(\phi(j)) \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Let $\mathbf{X}$ be a cepstral feature vector and $\mathbf{Y}$ be the cepstral feature vector after applying the linear frequency warping, then the relationship between $\mathbf{X}$ and $\mathbf{Y}$ can be described by:

$$\mathbf{Y} = \mathbf{C} \cdot \mathbf{log}\left(\mathbf{T} \cdot \mathbf{exp}\left(\mathbf{C}^{-1} \cdot \mathbf{X}\right)\right) \tag{3}$$

where $\mathbf{C}$ and $\mathbf{C}^{-1}$ are the DCT and inverse DCT matrices, respectively, and $\mathbf{log}(\cdot)$ and $\mathbf{exp}(\cdot)$ are component-wise logarithm and exponential functions of a matrix. $\mathbf{T} = \mathbf{M} \cdot \mathbf{R} \cdot \mathbf{M}^*$ where $\mathbf{M}^*$ is the matrix that transforms features from the Mel-frequency domain to the linear frequency domain. Eq.3 is equivalent to the one presented in [2].

Before we discuss the properties of the transform in Eq.3, let us £rst de£ne an index mapping (IM) matrix. A matrix is called an index mapping matrix if there is one and only one "1" in each row and all the other components are zeros. Obviously, the product of IM matrices is still an IM matrix. It is easy to show that the frequency warping matrix $\mathbf{R}$ is an IM matrix.

Next, we show that if the matrix $\mathbf{T}$ in Eq.3 is an IM matrix, then $\mathbf{X}$ and $\mathbf{Y}$ are related by a linear transformation. Since

$\mathbf{T}$ is an IM matrix, it only re-maps the index order of vector components and does not alter the value of it. Hence, we can exchange the order of $\mathbf{T}$ and $\mathbf{log}(\cdot)$:

$$\begin{aligned}\mathbf{Y} &= \mathbf{C}\cdot\mathbf{log}\left(\mathbf{T}\cdot\mathbf{exp}\left(\mathbf{C}^{-1}\cdot\mathbf{X}\right)\right)\\ &= \mathbf{C}\cdot\mathbf{T}\cdot\left(\mathbf{log}\cdot\mathbf{exp}\left(\mathbf{C}^{-1}\cdot\mathbf{X}\right)\right)\\ &= \mathbf{C}\cdot\mathbf{T}\cdot\mathbf{C}^{-1}\cdot\mathbf{X}\\ &= \mathbf{A}\cdot\mathbf{X}\end{aligned} \tag{4}$$

where

$$\mathbf{A}=\mathbf{C}\cdot\mathbf{T}\cdot\mathbf{C}^{-1}=\mathbf{C}\cdot\mathbf{M}\cdot\mathbf{R}\cdot\mathbf{M}^{*}\cdot\mathbf{C}^{-1} \tag{5}$$

Consequently, the means of $\mathbf{X}$ and $\mathbf{Y}$ also satisfy the same linear relation:

$$\mu_Y = E\{\mathbf{Y}\}=E\{\mathbf{A}\cdot\mathbf{X}\}=\mathbf{A}\cdot E\{\mathbf{X}\}=\mathbf{A}\cdot\mu_X \tag{6}$$

In most cases, speech features employed in automatic speech recognizers are a concatenation of static MFCCs with their £rst (delta) and second (delta-delta) order derivatives. In this paper, the derivatives are computed using £rst order difference:

$$\triangle\mathbf{X}_t=\mathbf{X}_t-\mathbf{X}_{t-1}\ ,\quad \triangle^2\mathbf{X}_t=\triangle\mathbf{X}_t-\triangle\mathbf{X}_{t-1} \tag{7}$$

It is straightforward that if Eq.4 holds, then we have:

$$\mu_{\triangle Y}=\mathbf{A}\cdot\mu_{\triangle X}\ ,\quad \mu_{\triangle^2 Y}=\mathbf{A}\cdot\mu_{\triangle^2 X} \tag{8}$$

The Mel £lter bank computation of MFCCs involves the summation of spectra samples within the frequency range of each triangular £lter. Therefore, $\mathbf{M}$ is not an IM matrix. So $\mathbf{T}$ is generally not an IM matrix either and Eq.3 can not be expressed as a linear transformation. However, suppose we substitute the output of each triangular £lter in the £lterbank with the value of the center frequency sample (peak) of that £lter, we are able to approximate $\mathbf{M}$ with an IM matrix $\tilde{\mathbf{M}}$:

$$\tilde{\mathbf{M}}=\begin{bmatrix}\tilde\theta_{1,1} & \cdots & \tilde\theta_{1,K_1} & & \\ & \tilde\theta_{2,1} & & \tilde\theta_{2,K_2} & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \tilde\theta_{N,1} & \cdots & \tilde\theta_{N,K_N}\end{bmatrix}_{N\times L}$$

where

$$\tilde\theta_{ij}=\begin{cases}1, & \text{if } \theta_{ij} \text{ is the central frequency of the } i\text{th £lter}\\ 0, & \text{otherwise.}\end{cases}$$

Similarly, $\mathbf{M}^*$ which maps samples from the Mel-frequency domain to the linear frequency domain can be created by setting the output of each triangular £lter on the Mel-frequency axis as the sample value of the corresponding center frequency on the linear frequency axis. The other frequency samples in the linear frequency domain are interpolated by repeating neighboring center frequencies that have already been generated as shown in Eq.9:

$$\mathbf{M}^{*}=\begin{bmatrix}1&0&0&\cdots&0\\0&1&0&\cdots&0\\0&1&0&\cdots&0\\ \vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\cdots&1\\0&0&0&\cdots&1\end{bmatrix}_{L\times N} \tag{9}$$

Thus, $\tilde{\mathbf{M}}$, $\mathbf{M}^*$ and $\mathbf{R}$ are all IM matrices. $\mathbf{T}\approx\tilde{\mathbf{M}}\cdot\mathbf{R}\cdot\mathbf{M}^{*}$ is also an IM matrix. In this way, a linear transformation of μ_Y and μ_X is guaranteed. That is,

$$\mu_Y\approx\mathbf{A}\cdot\mu_X \tag{10}$$

where

$$\mathbf{A}=\mathbf{C}\cdot\tilde{\mathbf{M}}\cdot\mathbf{R}\cdot\mathbf{M}^{*}\cdot\mathbf{C}^{-1} \tag{11}$$

The advantage of using matrices in the discrete frequency domain is that it can avoid complicated calculus computation when determining matrix components in the continuous frequency domain as shown in [4]. Since triangular Mel £lter banks are not invertible, the approximated linear relationship is dif£cult to obtain from the continuous frequency domain. Eq. 10 could be considered as a special case of maximum likelihood linear regression (MLLR), and Eq. 11 gives the £ve matrices to construct the transformation matrix $\mathbf{A}$. Among the matrices, $\mathbf{R}$ is the discretized form of a frequency warping function which could be chosen carefully, as shown in the next section, to reduce the spectral mismatch in speaker adaptation.

To investigate the effects of the linearization approximation, Fig. 2 demonstrates two MFCC features (C1 - C13) with and without linearization approximation. The approximation only results in slight differences between the two feature sets.

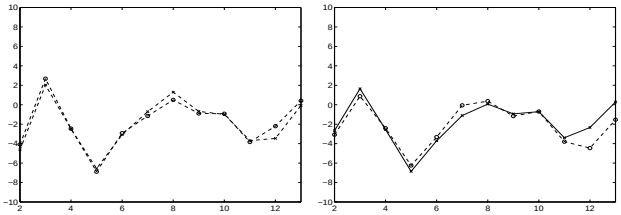

Figure 2: MFCC features with (solid line) and without (dashed line) linearization approximation. The dimensions are from C1 to C13 of an /uw/ sound.

3. Peak Alignment

Vocal tract variation, which results in spectral mismatch, is a major source for performance degradation of speech recognizers with different speakers. Fig.3 shows two /uw/ (from the digit "two") spectra of a 25ms frame from an adult male and a boy (about 10 years old). Obvious pitch and formant differences can be observed. If we can re-shape the two spectra by aligning the corresponding formants, then the spectral mismatch can be mitigated. In this paper, formant-like peaks are estimated us-

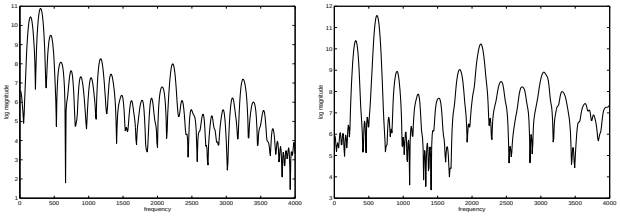

Figure 3: Spectra of the steady part of the sound /uw/ in the digit "two" from an adult male (left) and a boy (right).

ing Gaussian mixtures via the EM algorithm as proposed in [5]. In this algorithm, the magnitude of a spectrum in each frame is considered as a probability density function, and a Gaussian mixture model is used to £t it iteratively. Estimated means, variances and mixture weights of the Gaussians correspond to the

locations, bandwidths and amplitudes of the formants. Since the peaks found in this way are not necessarily the formants, they are called "formant-like" peaks.

Fig. 4 illustrates the spectrograms with peaks estimated using Gaussian mixtures. The speakers and utterances are the same as in Fig. 3. It is observed that typically in the 4 kHz frequency range, adult speakers have four formants while child speakers have only three. Hence, four Gaussian mixtures are used for adult males and three Gaussian mixtures for kids in the estimation of the experiments in this paper. Based on

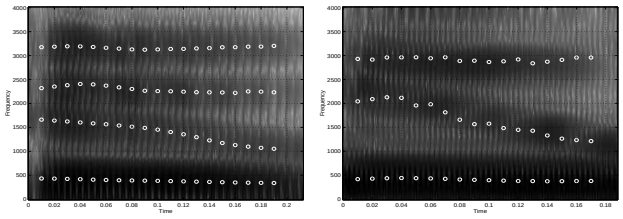

Figure 4: Formant-like peaks estimated (white circles) using Gaussian mixtures for the sound /uw/ in digit "two" from an adult male (left) and a boy (right).

the estimated peaks, we align them using a piece-wise linear function. Suppose we have $M-1$ peaks to align, they are $\{\omega_1^c,\ldots,\omega_{M-1}^c\}$ for the child speaker, and $\{\omega_1^a,\ldots,\omega_{M-1}^a\}$ for the adult speaker. Also, we de£ne $\omega_0^c=\omega_0^a=1$. Since $\{\omega_1^c,\ldots,\omega_{M-1}^c\}$ and $\{\omega_1^a,\ldots,\omega_{M-1}^a\}$ are estimated Gaussian mixture means, they are real numbers, not necessarily integers. The piece-wise linear function is described in Eq.12.

$$\phi(l)=\begin{cases}\omega_i^c+\frac{\omega_{i+1}^c-\omega_i^c}{\omega_{i+1}^a-\omega_i^a}\cdot(l-\omega_i^a)\\ \qquad\text{for } l\in(\omega_i^a,\omega_{i+1}^a) \text{ and } i=0,\ldots,M-2.\\ \omega_{M-2}^c+\frac{\omega_{M-1}^c-\omega_{M-2}^c}{\omega_{M-1}^a-\omega_{M-2}^a}\cdot(l-\omega_{M-2}^a)\\ \qquad\text{for } l\in(\omega_{M-1}^a,\omega_M^a).\end{cases} \tag{12}$$

Note that we require $\omega_0^c=\omega_0^a$ but there is no requirement that $\omega_M^c=\omega_M^a$. This is because children usually have higher formants than adults. and therefore, in the same frequency range, have fewer formants than adults. By not requiring that ω_M^c equals ω_M^a, it is possible for the extra formants in adult spectra to disappear after alignment. The left panel of Fig. 5 shows the piece-wise linear function computed according to Eq.12 aligning the £rst and the third formant-like peaks in Fig. 4. Since formants gradually change from frame to frame, the median value for each peak is used. In the right panel of Fig. 5, the original spectrum of the child's speech (solid line) and the reshaped spectrum (dotted line) of the adult's speech from Fig. 3 are illustrated. Compared with the spectra in Fig. 3, the mismatch between the two spectra is greatly reduced. On the basis

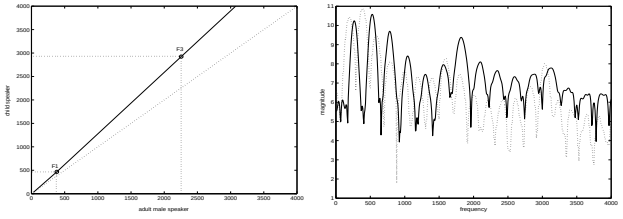

Figure 5: Piece-wise linear function (left) and reshaped adult's spectrum after peak alignment (right).

of the frequency warping function $\phi(l)$, $\mathbf{R}$ could be created and eventually $\mathbf{A}$ could be obtained by Eq.11.

4. Adaptation Scheme

We adapt the means and variances of the acoustic models separately in an unconstrained fashion. The means of Gaussian mixtures of HMMs are transformed as:

$$\boldsymbol{\mu}^{new}=\mathbf{A}\,\boldsymbol{\mu}+\mathbf{b} \tag{13}$$

where the transformation matrix $\mathbf{A}$ is generated deterministically using Eq.11 after obtaining $\mathbf{R}$ based on peak alignment described in Section 3. The estimation of the bias vector $\mathbf{b}$ from the adaptation data is carried out under the maximum likelihood criterion, which can be accomplished using the EM algorithm [6].

Suppose the biases are tied into Q classes : $\{\omega_1,\cdots,\omega_q,\cdots,\omega_Q\}$. For a speci£c class ω_q, the bias $\mathbf{b}_q$ is shared across all the Gaussian mixtures $\mathcal{N}(\mathbf{o}_t;\boldsymbol{\mu}_{ik},\boldsymbol{\Sigma}_{ik})$ with $(i,k)\in\omega_q$ and is given by:

$$\mathbf{b}_q = \left[\sum_{u=1}^{U}\sum_{(i,k)\in\omega_q}\sum_{t=1}^{T^u}\gamma_t^u(i,k)\cdot\boldsymbol{\Sigma}_{ik}^{-1}\right]^{-1}\cdot \left[\sum_{u=1}^{U}\sum_{(i,k)\in\omega_q}\sum_{t=1}^{T^u}\gamma_t^u(i,k)\cdot\boldsymbol{\Sigma}_{ik}^{-1}\cdot(\mathbf{o}_t^u-\mathbf{A}\boldsymbol{\mu}_{ik})\right] \tag{14}$$

where U is the number of utterances in the adaptation data and T^u is the number of frames in the uth utterance. $i\in\{1,2,\cdots,N\}$ and $k\in\{1,2,\cdots,M\}$ are the indices of state and mixture sets, respectively. $\gamma_t^u(i,k)=p(s_t^u=i,\xi_t^u=k|\mathbf{O}^u,\overline{\lambda})$ is the posterior probability of being at state i mixture k at time t given the uth observation sequence. $\boldsymbol{\mu}_{ik}$ and $\boldsymbol{\Sigma}_{ik}$ are the mean vector and covariance matrix associated with it, respectively. Typically, $\boldsymbol{\Sigma}_{ik}$ is a diagonal covariance matrix so that Eq.14 can be solved one dimension at a time.

Given the adapted Gaussian mixture means, the diagonal covariance matrices are adapted as described in [7]:

$$\boldsymbol{\Sigma}_{ik}^{new}=\mathbf{B}_{ik}^T\mathbf{H}_q\mathbf{B}_{ik} \tag{15}$$

where $\mathbf{H}_q$ is the linear covariance transformation shared by all Gaussian mixtures in the class ω_q, namely, $(i,k)\in\omega_q$. $\mathbf{B}_{ik}$ is the inverse of the Cholesky factorization of $\boldsymbol{\Sigma}_{ik}^{-1}$. That is,

$$\boldsymbol{\Sigma}_{ik}^{-1}=\mathbf{C}_{ik}\mathbf{C}_{ik}^{-1} \quad\text{and}\quad \mathbf{B}_{ik}=\mathbf{C}_{ik}^{-1} \tag{16}$$

The maximum likelihood estimation of the covariance linear transformation $\mathbf{H}_q$ is given by

$$\frac{\sum_{(i,k)\in\omega_q}\mathbf{C}_{ik}^T[\sum_{u=1}^{U}\sum_{t=1}^{T^u}\gamma_t^u(i,k)(\mathbf{o}_t^u-\boldsymbol{\mu}_{ik})(\mathbf{o}_t^u-\boldsymbol{\mu}_{ik})^T]\mathbf{C}_{ik}}{\sum_{(i,k)\in\omega_q}\sum_{u=1}^{U}\sum_{t=1}^{T^u}\gamma_t^u(i,k)} \tag{17}$$

By forcing the $\mathbf{H}_q$'s off-diagonal terms to zeros, a diagonal covariance matrix $\boldsymbol{\Sigma}_{ik}$ is obtained after adaptation.

To summarize, the adaptation scheme investigated above deterministically generates the transformation matrix $\mathbf{A}$ and statistically estimates the bias $\mathbf{b}$ and variance transformation $\mathbf{H}$. Therefore, the number of parameters to be estimated is reduced which can achieve better performance in case of limited adaptation data. Obviously, bias and variance adaptation can still bene£t from large amounts of adaptation data.

5. Experimental Results and Discussion

Experiments are performed on connected digit strings from the TIDIGITS database. Acoustic models are trained on adult males and tested on children. Speech data from 55 male speakers are used in training and data from 5 boys and 5 girls in testing. Each speaker contributes 77 utterances with 1-7 digits for each utterance. For each child, the adaptation utterances, which consists of 4, 7, 10, 20 or 30 digits, are randomly chosen from the test set and not used in the testing. The speech signals are down-sampled from 20 to 8 kHz. Each speech frame is 25ms long with a 10ms overlap. Feature vectors are of 39 dimensions: 13 static features plus their £rst- and second-order derivatives. Acoustic HMMs are phoneme-based with a left-to-right topology. There are 18 monophones plus silence and short pause models. Monophones have 2 to 4 states with 6 Gaussian mixtures in each state. In adaptation, voiced segments are detected from the speech signals via the cepstral peak analysis technique [8]. Formant-like peaks are estimated from the voiced segments by Gaussian mixtures [5]. For a speci£c speaker, the median of peaks in each voiced segment is £rst obtained and the average over all the medians serves as the estimate of the peaks and is used in the alignment. The adult male who yields the highest likelihood in the training set is selected as the "standard" adult speaker and used to represent the acoustic characteristics of the entire adult training set. The £rst and the third formants are then aligned. The Gaussian mixtures are initialized with means uniformly located on the frequency axis with equal mixture weights. For each frame, 20 EM iterations are performed. The biases and variances are tied using a regression tree with 20 base classes.

Table 1 shows the performance of the proposed adaptation approach (PA) compared to the traditional MLLR and VTLN algorithm. The MLLR transformation matrices are 3-block diagonal and tied using the same regression tree as the biases and variances. The VTLN is implemented utterance by utterance as follows: HMMs £rst provide an initial hypothesis for each utterance, and warping factors within $[0.7, 1.1]$ are then applied to the signal with a stepsize of 0.05. The optimal warping factor is chosen as the one which gives the highest likelihood score based on forced alignment and is used to scale the frequency axis in the feature extraction stage. For comparison, the performance of linear approximation of VTLN studied in [2] is also presented in the table and denoted LA_VTLN. From the table, MLLR has poor performance when the adaptation data are limited, which is due to unreliable parameter estimation. PA and VTLN signi£cantly outperform MLLR under this condition because they utilize spectral information to reduce the mismatch and have fewer parameters to estimate. As the amount of adaptation data grows, MLLR performance improves. Hence, MLLR has an advantage when large amounts of data are available while VTLN is advantageous for limited amounts of data.

	Number of adaptation digits				
Algorithm	4	7	10	20	30
no adaptation	38.9	38.9	38.9	38.9	38.9
MLLR	60.2	72.5	76.6	89.9	92.0
VTLN	89.8	89.8	89.8	89.8	89.8
LA-VTLN	85.0	85.8	86.4	89.9	91.6
PA	90.1	91.1	91.0	93.5	95.1

Table 1: Recognition accuracy of children's speech with MFCC features (TIDIGITS).

In the proposed approach, transformation of means is £rst deterministically generated by aligning the formant-like peaks, on the basis of which, statistical approaches such as tree-based tied variance and bias adaptation are performed. In this way, the algorithm can take advantage of both large and limited amounts of adaptation data. Moreover, the linear approximation investigated in this paper gives better performance than that studied in [2], which may be because a relatively large scaling factor is required for children's speech, while the Taylor expansion made in [2] based on a small factor does not hold. Experiments using adult female's acoustic models also demonstrate a similar trend although the baseline (no adaptation) performance is higher.

6. Conclusions

An MLLR-like adaptation approach is proposed where the transformation of the means is performed deterministically based on linearization of VTLN. Biases and adaptation of the variances are estimated statistically by the EM algorithm. We show that under certain approximations, frequency warping of Mel-£lter-bank-based MFCCs equals a linear transformation in the cepstral space. Based on that linear relationship, a formant-like peak alignment algorithm to adapt adult acoustic models to children's speech is proposed. Performance improvements are observed compared to traditional MLLR and VTLN. [1]

7. References

[1] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[2] T. Claes, I. Dologlou, L. Bosch, and D. Compernolle, "A novel feature transformation for vocal tract length normalization in automatic speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 1998.

[3] J. McDonough, T. Schaaf, and A. Waibel, "Speaker adaptation with all-pass transforms," *Speech Communication*, vol. 42, pp. 75–91, 2004.

[4] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC," *Proc. of European Conf. on Speech Communication and Technology*, pp. 1445–1448, 2003.

[5] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of Gaussians," *Proc. of Int. Conf. on Spoken Language Processing*, pp. 1229–1232, 1996.

[6] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[7] M. Gales, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.

[8] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[9] X. Cui and A. Alwan, "Adaptation of children's speech with limited data based on formant-like peak alignment," *Computer Speech and Language*, to appear.

[1] This material is supported in part by NSF Grant No. 0326214. Any opinions, £ndings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily re¤ect the views of the NSF.