Adaptation of children's speech with limited data based on formant-like peak alignment

Xiaodong Cui *, Abeer Alwan

Department of Electrical Engineering, The Henry Samuli 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

Automatic recognition of children's speech using acoustic models trained by adults results in poor performance due to differences in speech acoustics. These acoustical differences are a consequence of children having shorter vocal tracts and smaller vocal cords than adults. Hence, speaker adaptation needs to be performed. However, in real-world applications, the amount of adaptation data available may be less than what is needed by common speaker adaptation techniques to yield reasonable performance. In this paper, we first study, in the discrete frequency domain, the relationship between frequency warping in the front-end and corresponding transformations in the back-end. Three common feature extraction schemes are investigated and their transformation linearity in the back-end are discussed. In particular, we show that under certain approximations, frequency warping of MFCC features with Mel-warped triangular filter banks equals a linear transformation in the cepstral space. Based on that linear transformation, a formant-like peak alignment algorithm is proposed to adapt adult acoustic models to children's speech. The peaks are estimated by Gaussian mixtures using the Expectation-Maximization (EM) algorithm (Zolfaghari & Robinson, 1996). For limited adaptation data, the algorithm

outperforms traditional vocal tract length normalization (VTLN) and maximum likelihood linear regression (MLLR) techniques.

Key words: automatic speech recognition, speaker adaptation, children's speech, VTLN, peak alignment, limited data, MLLR.

1 Introduction

It is well known that speech characteristics of adults and children differ due to differences between their vocal apparatus. Children have higher formant and fundamental frequencies in their spectra than adults because of their shorter vocal tracts and smaller vocal cords. Since most of the current automatic speech recognition systems are trained on adult speech, such systems suffer from dramatically degraded performance for child speakers (Li & Russell, 2001; Wilpon & Jacobsen, 1996). To reduce spectral mismatch between adult and children's speech, various vocal tract length normalization (VTLN) and speaker adaptation techniques have been used (Burnett & Fanty, 1996; Das et al., 1998; Potamianos & Narayanan, 2003).

VTLN algorithms are typically applied to the front-end feature domain whereby the linear frequency or Mel-frequency axis is scaled by a warping factor which is obtained via a grid search. In (Lee & Rose, 1998), a frequency warping approach is investigated. The linear frequency warping factor is first estimated from the input speech based on the maximum likelihood criterion and then

* Corresponding author. Tel: 1-310-206-2231; Fax: 1-310-206-4685.
Email addresses: xdcui@icsl.ucla.edu (Xiaodong Cui),
alwan@icsl.ucla.edu (Abeer Alwan).

used to re-scale the filter banks when computing Mel-frequency cepstral coefficients (MFCC) features. An efficient algorithm based on a generic voiced speech model is studied in (Wegmann et al., 1998) to simplify the procedure of selecting the frequency warping factor; the technique achieves excellent results using conversational telephone speech. In addition, a variety of frequency axis re-scaling strategies are discussed in (Burnett & Fanty, 1996; Das et al., 1998; Eide & Gish, 1996; Gouvea & Stern, 1997) to address vocal tract shape variation between children and adults.

Speaker adaptation schemes, on the other hand, are adopted in the back-end acoustic model domain (Gales, 1998; Gauvain & Lee, 1994; Leggetter & Woodland, 1995). These algorithms try to tune the acoustic models towards a specific speaker utilizing adaptation data. In general, unlike VTLN which attempts to compensate for physical (vocal tract) differences, adaptation techniques are statistically driven using the maximum likelihood or maximum a posterior probability criteria. Most often, the computational complexity of speaker adaptation algorithms are higher than VTLN and require more adaptation data. If the amount of adaptation data is adequate for reliable estimation, adaptation techniques may achieve better performance than VTLN.

Among the adaptation algorithms, the maximum likelihood linear regression (MLLR) technique, introduced in (Leggetter & Woodland, 1995), is the most widely used approach. In MLLR, the relationship between the means of the Gaussian mixtures of an acoustic hidden Markov model (HMM) and that of a new speaker can be described using linear regression:

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\ \boldsymbol{\mu} + \mathbf{b} \tag{1}$$

The transformation matrix $\mathbf{A}$ and bias vector $\mathbf{b}$ are estimated using the

Expectation-Maximization (EM) algorithm with the maximum likelihood criterion. In real-world applications, one often encounters the situation where only a limited amount of adaptation data is available for the new speaker. This may be because adaptation data are difficult to obtain and/or time limitations do not permit collecting enough data. Under these conditions, MLLR performance is unsatisfactory due to unreliable parameter estimation, especially for the transformation matrix $\mathbf{A}$ since it has more parameters to estimate than the bias vector $\mathbf{b}$. To reduce the number of parameters, a 3-block diagonal matrix is usually employed where the static, delta and delta-delta (first and second derivatives, respectively) parts of the features have their own full sub-matrices and independence is assumed among the three parts (Gales et al., 1996; Young et al., 2001). A diagonal form of the transformation matrix $\mathbf{A}$ is also studied in (Leggetter & Woodland, 1995) and is shown to have limited performance improvements. In (Digalakis et al., 1999), several rapid speaker adaptation methods are summarized for large vocabulary speech recognizers. These methods explore the dependencies between speech units and efficiently make use of small amounts of data by only utilizing the biases in MLLR transforms.

In recent years, the relationship between frequency warping in the front-end feature domain and the corresponding transformation in the cepstral domain has drawn increasing attention in the speaker adaptation area (Claes et al., 1998; Ding et al., 2002; McDonough et al., 2004; Pitz et al., 2001; Pitz & Ney, 2003). Conclusions are made in (McDonough et al., 2004) and (Pitz & Ney, 2003) that VTLN equals a linear transform in the cepstral space. Perceptual linear prediction (PLP) cepstral coefficients features and MFCCs with Mel-frequency warping, instead of Mel-warped filter banks, are used in

(McDonough et al., 2004) and (Pitz & Ney, 2003), respectively. For both features, the frequency warping is invertible and the derivation is performed in continuous frequency ω or in the $\mathcal{Z}$ space.

In this paper, we first discuss, in the discrete frequency domain, the relationship between frequency warping in the front-end domain and the corresponding transformation linearity in the back-end domain of a variety of feature extraction schemes. In particular, we show that under certain approximations, the frequency warping of MFCC features with Mel-warped triangular filter banks equals a linear transformation in the model domain. The linear transform can be considered as a special case of the traditional MLLR and serves as a basis to cope with the sparse adaptation data problem. Utilizing the linear transformation, a fast adaptation approach based on formant-like peak alignment is proposed. In this proposed approach, the transformation matrix $\mathbf{A}$ is computed deterministically after which the bias vector $\mathbf{b}$ is estimated statistically within the EM framework. As mentioned earlier, MLLR needs more data to reliably estimate $\mathbf{A}$ than that needed to estimate $\mathbf{b}$. By generating $\mathbf{A}$ based on re-mapping of the formant-like peaks, the proposed approach can ameliorate the spectral mismatch between adults and children's speech while reducing the number of parameters to be estimated; this makes robust estimation of the bias $\mathbf{b}$ possible.

The remainder of the paper is organized as follows. In Section 2, the relationship between frequency warping in the front-end and corresponding transformations in the back-end is investigated for three common features, and the conditions under which this relationship is equivalent to a linear transformation are discussed. The focus is on the discrete frequency domain. In Section 3, the estimation of formant-like peaks in speech spectra using Gaussian mixtures

is described. Recognition results are presented in Section 4 and conclusions are made in Section 5.

2 Relationship between frequency warping and linear transformations

2.1 Feature Schemes

We study three kinds of speech features in this paper: cepstra without Mel-scale warping (CEP), cepstra with Mel-scale warping (MFCC1) and Cepstra computed using Mel-warped triangular filter banks (MFCC2). While MFCC2 is the most widely used front-end feature in the state-of-the-art automatic speech recognition systems, we include CEP and MFCC1 for comparison with the work in (Pitz et al., 2001) and (Pitz & Ney, 2003). In the discrete frequency domain, we will show that for the first two feature extraction strategies (CEP and MFCC1), frequency warping is indeed equivalent to a linear transform in the cepstral space as stated in (Pitz et al., 2001) and (Pitz & Ney, 2003), and the corresponding discrete frequency transformation matrices are given. However, this conclusion is not true for MFCC2 features computed using Mel-warped filter banks unless certain approximations are made.

Fig. 1 illustrates the three feature extraction schemes. The input speech signal is first pre-emphasized and framed by Hamming windows. For each speech frame, the magnitude of its Discrete Fourier Transform (DFT) is obtained and then converted into the Mel-frequency domain by certain mappings. The logarithm is then computed on the Mel-spectra to compress the dynamic range, and the output is further decorrelated by the Discrete Cosine Transform

(DCT) to obtain the final cepstral coefficients.

Let $\mathbf{S}^l$ denote the linear spectrum and $\mathbf{S}^c$ the cepstrum, according to Fig. 1, we have

$$\mathbf{S}^c = \mathbf{C} \cdot \mathbf{log}(\mathbf{M} \cdot \mathbf{S}^l) \tag{2}$$

where $\mathbf{M}$ is the Mel-mapping matrix, $\mathbf{C}$ the DCT matrix and $\mathbf{log}$ the component-wise logarithm function applied to the matrix. The three different features have different Mel-mapping matrices $\mathbf{M}$:

- For CEP, since there is no Mel-scale mapping, $\mathbf{M}$ simply equals the identity matrix. That is,

$$\mathbf{M} = \mathbf{I} \tag{3}$$

- For MFCC1, the Mel-scale warping from the linear frequency f to the Mel-frequency $\varphi(f)$ is defined by (Young et al., 2001)

$$\varphi(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

  In the discrete frequency domain, the relationship between the linear frequency index l and the Mel-frequency index k is

$$k = \text{round}\left[\varphi\left(\frac{F_{max} l}{L}\right) \cdot \frac{N}{F_{max}}\right] \tag{5}$$

  where F_{max} is the maximum frequency in the spectrum, L and N are the sample numbers in the linear and Mel-frequency domains, respectively. Define

$$\psi(l) = \varphi\left(\frac{F_{max} l}{L}\right) \cdot \frac{N}{F_{max}} \tag{6}$$

The Mel-mapping matrix can be expressed as

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0\,0\,0 \cdots & 0 \\ 0 & 1 & 0\,0\,0 \cdots & 0 \\ \vdots & \vdots & \vdots\,\vdots\,\vdots\ \ddots & \vdots \\ 0 \cdots & 0 & 1\,0 \cdots & 0 \end{bmatrix}_{N \times L} \tag{7}$$

where $\mathbf{M}$'s components are defined as:

$$m_{ij} = \begin{cases} 1, & \text{if } i = \text{round}(\psi(j)) \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Typically, in order to perform the Mel-scale mapping in the discrete frequency domain, an "oversampling" strategy is used in the linear spectral domain (Deller et al., 1987) and a smaller number of samples in the Mel-spectral domain are generated by selecting appropriate frequency components from the linear spectral domain. Therefore, L is larger than N in Eq. 7.

- For MFCC2, triangular filter banks are employed in Mel-mapping whose central frequencies are equally spaced in the Mel-frequency axis (Young et al., 2001) as shown in Fig. 2.

The corresponding Mel-mapping matrix $\mathbf{M}$ can be written as:

$$\mathbf{M} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,K_1} & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \theta_{2,1} & \cdots & \theta_{2,K_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & \theta_{N,1} & \cdots & \theta_{N,K_N} \end{bmatrix}_{N \times L} \tag{9}$$

where L and N are the sample numbers in the linear and Mel-frequency domains, respectively, $\theta_{i,j}$ are weights of the triangular filters and $K_1, \ldots, K_N$ are the numbers of non-zero weights of each triangular filter. Typically, N is much smaller than L.

2.2 Derivation of the transformation matrix $\mathbf{A}$

Suppose there exists a warping function in the discrete linear frequency domain $l = g(k)$, where k and l are the discrete frequency sample indices. This can be presented as a warping matrix R whose components are defined as:

$$r_{ij} = \begin{cases} 1, & \text{if } i = \text{round}(g(j)) \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

in case the index j, computed using the warping function, is located outside the sample number interval, e.g. $j < 0$ or $j > L-1$ where L is the total discrete sample number, 0 or $L-1$ is set for the index. Let $\mathbf{X}$ be one speech feature vector and $\mathbf{Y}$ be the feature after applying the linear frequency warping, then

$\mathbf{X}$ and $\mathbf{Y}$ have a relationship described by Eq.11.

$$\mathbf{Y} = \mathbf{C} \cdot \mathbf{log}\left(\mathbf{M} \cdot \mathbf{R} \cdot \mathbf{M}^* \cdot \mathbf{exp}\left(\mathbf{C}^{-1} \cdot \mathbf{X}\right)\right) \tag{11}$$

where $\mathbf{C}$ and $\mathbf{C}^{-1}$ are the DCT and inverse DCT matrices, respectively. $\mathbf{R}$ is the linear frequency warping matrix. $\mathbf{M}$ is the Mel-mapping matrix and $\mathbf{M}^*$ is the matrix that transforms features from the Mel-frequency domain to the linear frequency domain. $\mathbf{log}(\cdot)$ and $\mathbf{exp}(\cdot)$ are component-wise logarithm and exponential functions of the matrix.

Let $\mathbf{T} = \mathbf{M} \cdot \mathbf{R} \cdot \mathbf{M}^*$, then Eq.11 can be written as:

$$\mathbf{Y} = \mathbf{C} \cdot \mathbf{log}\left(\mathbf{T} \cdot \mathbf{exp}\left(\mathbf{C}^{-1} \cdot \mathbf{X}\right)\right) \tag{12}$$

This equation is equivalent to the one presented in (Claes et al., 1998).

Before we discuss the properties of the transform in Eq.12, let us first define 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. There is no requirement on the dimension of an IM matrix. It is not necessarily a square matrix. For example, the Mel-mapping matrix $\mathbf{M}$ in Eq.7 and the warping matrix $\mathbf{R}$ mentioned above are all IM matrices. Furthermore, it is obvious that the product of IM matrices is still an IM matrix.

Next, we show that if the matrix $\mathbf{T}$ in Eq.12 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 order of vector component indices and does not alter the value of them. Therefore, we can exchange the order of $\mathbf{T}$ and $\mathbf{log}(\cdot)$:

$$\begin{aligned}\mathbf{Y} &= \mathbf{C} \cdot \mathbf{log}\left(\mathbf{T} \cdot \mathbf{exp}\left(\mathbf{C}^{-1} \cdot \mathbf{X}\right)\right) \\ &= \mathbf{C} \cdot \mathbf{T} \cdot \left(\mathbf{log} \cdot \mathbf{exp}\left(\mathbf{C}^{-1} \cdot \mathbf{X}\right)\right) \\ &= \mathbf{C} \cdot \mathbf{T} \cdot \mathbf{C}^{-1} \cdot \mathbf{X} \\ &= \mathbf{A} \cdot \mathbf{X} \end{aligned} \tag{13}$$

where

$$\mathbf{A} = \mathbf{C} \cdot \mathbf{T} \cdot \mathbf{C}^{-1} \tag{14}$$

Or, by substituting $\mathbf{T} = \mathbf{M} \cdot \mathbf{R} \cdot \mathbf{M}^*$ into Eq.14, we obtain

$$\mathbf{A} = \mathbf{C} \cdot \mathbf{M} \cdot \mathbf{R} \cdot \mathbf{M}^* \cdot \mathbf{C}^{-1} \tag{15}$$

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

$$E\{\mathbf{Y}\} = E\{\mathbf{A} \cdot \mathbf{X}\} = \mathbf{A} \cdot E\{\mathbf{X}\} \tag{16}$$

In other words,

$$\boldsymbol{\mu}_Y = \mathbf{A} \cdot \boldsymbol{\mu}_X \tag{17}$$

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

$$\triangle \mathbf{X}_t = \mathbf{X}_t - \mathbf{X}_{t-1} \tag{18}$$
$$\triangle^2 \mathbf{X}_t = \triangle \mathbf{X}_t - \triangle \mathbf{X}_{t-1} \tag{19}$$

It is straightforward that if Eq.13 holds, then we have

$$\triangle \mathbf{Y} = \mathbf{A} \cdot \triangle \mathbf{X} \tag{20}$$
$$\triangle^2 \mathbf{Y} = \mathbf{A} \cdot \triangle^2 \mathbf{X} \tag{21}$$

Thus,

$$\boldsymbol{\mu}_{\triangle Y} = \mathbf{A} \cdot \boldsymbol{\mu}_{\triangle X} \tag{22}$$
$$\boldsymbol{\mu}_{\triangle^2 Y} = \mathbf{A} \cdot \boldsymbol{\mu}_{\triangle^2 X} \tag{23}$$

As long as $\mathbf{T} = \mathbf{M} \cdot \mathbf{R} \cdot \mathbf{M}^*$ is an IM matrix, the expectations of the original feature $\mathbf{X}$ and warped feature $\mathbf{Y}$ are linearly related. Next, we will investigate the properties of the mean transformation of the three feature extraction schemes discussed in Section 2.1.

- For CEP,

$$\mathbf{M} = \mathbf{M}^* = \mathbf{I} \tag{24}$$

  both of which are IM matrices and the warping matrix $\mathbf{R}$ is also an IM matrix. Hence, $\mathbf{T}$ which is the product of three IM matrices, is also an IM matrix. According to the discussion above,

$$\boldsymbol{\mu}_Y = \mathbf{A} \cdot \boldsymbol{\mu}_X \tag{25}$$

  where

$$\mathbf{A} = \mathbf{C} \cdot \mathbf{R} \cdot \mathbf{C}^{-1} \tag{26}$$

- For MFCC1, $\mathbf{M}$ and $\mathbf{R}$ are both IM matrices. Since the Mel-mapping in Eq.7 is performed by first "oversampling" in the linear frequency domain and then selecting the desired frequency components in the Mel-frequency, the number of rows N is smaller than the number of columns L. Therefore, in order to recover the linear frequency samples from Mel-frequency,

interpolation is needed in matrix $\mathbf{M}^*$:

$$\mathbf{M}^* = \begin{bmatrix} 1\ 0\ 0 \cdots\ 0 \\ 1\ 0\ 0 \cdots\ 0 \\ 0\ 1\ 0 \cdots\ 0 \\ 0\ 1\ 0 \cdots\ 0 \\ 0\ 1\ 0 \cdots\ 0 \\ \vdots\ \vdots\ \vdots\ \ddots\ \vdots \\ 0\ 0\ 0 \cdots\ 1 \\ 0\ 0\ 0 \cdots\ 1 \end{bmatrix}_{L \times N} \tag{27}$$

where

$$m_{ij}^* = \begin{cases} 1, & \text{if } i = \text{round}(\psi^{-1}(j)) \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

The interpolation calculated according to Eq.28 generates the unseen samples by repeating existing neighboring samples. In this way, $\mathbf{M}^*$ is an IM matrix. Hence, $\mathbf{T}$ is also an IM matrix and we have

$$\boldsymbol{\mu}_Y = \mathbf{A} \cdot \boldsymbol{\mu}_X \tag{29}$$

where

$$\mathbf{A} = \mathbf{C} \cdot \mathbf{M} \cdot \mathbf{R} \cdot \mathbf{M}^* \cdot \mathbf{C}^{-1} \tag{30}$$

- For MFCC2, the Mel-mapping involves the summation of spectra samples within each triangular filter frequency range. Therefore, $\mathbf{M}$ is not an IM matrix. So $\mathbf{T}$ is generally not an IM matrix either. Eq.12 can not be expressed

as a linear transformation. However, suppose we substitute the output of each triangular filter in the filter banks with the value of the center frequency sample (peak) of that filter, we are able to approximate $\mathbf{M}$ with an IM matrix $\tilde{\mathbf{M}}$:

$$\tilde{\mathbf{M}} = \begin{bmatrix} \tilde{\theta}_{1,1} & \tilde{\theta}_{1,2} & \cdots & \tilde{\theta}_{1,K_1} & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \tilde{\theta}_{2,1} & \cdots & \tilde{\theta}_{2,K_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & \tilde{\theta}_{N,1} & \cdots & \tilde{\theta}_{N,K_N} \end{bmatrix}_{N \times L} \tag{31}$$

where

$$\tilde{\theta}_{i,j} = \begin{cases} 1, & \text{if } \theta_{i,j} \text{ is the central frequency of filter } i \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

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 filter on the Mel-frequency axis as the sample value at 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. Thus, $\mathbf{M}^*$ is an IM matrix. Since $\tilde{\mathbf{M}}$, $\mathbf{M}^*$ and $\mathbf{R}$ are all IM matrices, linear transformation of μ_Y and μ_X is guaranteed. That is,

$$\boldsymbol{\mu}_Y \approx \mathbf{A} \cdot \boldsymbol{\mu}_X \tag{33}$$

where

$$\mathbf{A} = \mathbf{C} \cdot \tilde{\mathbf{M}} \cdot \mathbf{R} \cdot \mathbf{M}^* \cdot \mathbf{C}^{-1} \tag{34}$$

The derivation in Section 2.2 shows the relationship between μ_Y and μ_X in the discrete frequency domain for the three features. In (Pitz et al., 2001) and (Pitz & Ney, 2003), the components of linear transformation matrix for Cepstral and MFCC features are computed in the continuous frequency domain as follows:

$$A_{nk}(\alpha) = \frac{2s_k}{\pi} \int_0^{\pi} d\tilde{\omega} \cos(\tilde{\omega}n) \cos(g_{\alpha}^{(-1)}(\tilde{\omega})k) \tag{35}$$

and

$$A_{nk}^{mel}(\alpha) = \frac{2s_k}{\pi} \int_0^{\pi} d\tilde{\omega}_{mel} \cos(\tilde{\omega}_{mel}n) \cos(g_{mel} \circ g_{\alpha}^{(-1)} \circ g_{mel}^{(-1)}(\tilde{\omega}_{mel})k) \tag{36}$$

where g_α and g_{mel} are linear frequency and Mel-scale warping functions, and

$$s_k = \begin{cases} 0.5, & k = 0 \\ 1, & \text{else.} \end{cases} \tag{37}$$

Note that Eq.26 for CEP and Eq.30 for MFCC1 are the discrete forms of Eq.35 and Eq.36, respectively, where $\mathbf{R}$ is the matrix form of g_α. In Eq.36, g_{mel} and g_{mel}^{-1} are represented by $\mathbf{M}$ and $\mathbf{M}^*$ in Eq.30. One advantage of using matrices in the discrete frequency domain is that it can avoid tedious and complicated calculus to calculate matrix components in the continuous frequency domain (Eq.35 and Eq.36). Generally, the analytical expression of the transformation matrices in Eq.35 and Eq.36 for an invertible warping function g_α is not always available. Even for some relatively simple warping functions, e.g. piece-wise linear, bilinear or quadratic functions, the computational load of Eq.35 and Eq.36 is high. Matrix expression in the discrete frequency domain can simplify the above calculation into simple index mapping matrices and greatly reduce

the computational complexity.

Note that the studies in (McDonough et al., 2004) and (Pitz & Ney, 2003) use either LPC-based features or MFCC features with only the Mel-scale warping, and the discussion are in the continuous frequency domain. We showed that since Mel-warped filter bank mapping is not invertible, it will not lead to a linear transformation in the cepstral domain unless a certain approximation is made. Moreover, the approximation made in Section 2.2 with triangular Mel-warped filter bank matrices is not easy to implement in the continuous frequency domain.

2.4 Estimation of the bias vector $\mathbf{b}$

Suppose we adapt the means of Gaussian mixtures of HMMs as:

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\ \boldsymbol{\mu} + \mathbf{b} \tag{38}$$

where the transformation matrix $\mathbf{A}$ is generated using the method described in Section 2.2. We want to estimate the bias vector $\mathbf{b}$ based on the adaptation data under the maximum likelihood criterion. This can be performed using the EM algorithm (Dempster et al., 1977).

Define the EM auxiliary function we are interested in as:

$$Q_b(\lambda; \overline{\lambda}) = \sum_{u=1}^{U} \sum_{i=1}^{N} \sum_{k=1}^{M} \sum_{t=1}^{T^u} \gamma_t^u(i,k) \cdot \log \mathcal{N}(\mathbf{o}_t; \mathbf{A}\boldsymbol{\mu}_{ik} + \mathbf{b}, \boldsymbol{\Sigma}_{ik}) \tag{39}$$

where U is the number of adaptation utterances and T^u is the number 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. $\gamma_t^u(i,k) = p(s_t^u = i, \xi_t^u = k | O^u, \overline{\lambda})$ is the posterior probability of staying at state i mixture k at time t given the uth

observation sequence. $\mathcal{N}(\mathbf{o}_t; \mathbf{A}\boldsymbol{\mu}_{ik} + \mathbf{b}, \boldsymbol{\Sigma}_{ik})$ is the kth multivariate Gaussian mixture in state i with weight α_{ik} while $\boldsymbol{\mu}_{ik}$ and $\boldsymbol{\Sigma}_{ik}$ are the mean vector and covariance matrix associated with it.

Suppose the biases are tied into Q classes: $\{\omega_1, \cdots, \omega_q, \cdots, \omega_Q\}$. For a specific class ω_q, the bias $\mathbf{b}_q$ is shared across all the Gaussian mixtures $\mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k})$ with $(i,k) \in \omega_q$. The maximum likelihood estimation of $\mathbf{b}_q$ could be obtained by setting the differentiation of $Q_b(\lambda; \overline{\lambda})$ with respect to $\mathbf{b}_q$ to zero:

$$\begin{aligned}
\frac{\partial Q_b(\lambda; \overline{\lambda})}{\partial \mathbf{b}_q} &= \frac{\partial}{\partial \mathbf{b}_q} \sum_{u=1}^{U} \sum_{(i,k)\in\omega_q} \sum_{t=1}^{T^u} \gamma_t^u(i,k) \cdot \log \mathcal{N}(\mathbf{o}_t^u; \mathbf{A}\boldsymbol{\mu}_{ik} + \mathbf{b}, \boldsymbol{\Sigma}_{ik}) \\
&= \frac{\partial}{\partial \mathbf{b}_q} \sum_{u=1}^{U} \sum_{(i,k)\in\omega_q} \sum_{t=1}^{T^u} \gamma_t^u(i,k) \left[-\frac{1}{2} (\mathbf{o}_t^u - \mathbf{A}\boldsymbol{\mu}_{ik} - \mathbf{b}_q)^T \boldsymbol{\Sigma}_{ik}^{-1} (\mathbf{o}_t^u - \mathbf{A}\boldsymbol{\mu}_{ik} - \mathbf{b}_q) \right] \\
&= \sum_{u=1}^{U} \sum_{(i,k)\in\omega_q} \sum_{t=1}^{T^u} \gamma_t^u(i,k) \cdot \boldsymbol{\Sigma}_{ik}^{-1} \cdot (\mathbf{o}_t^u - \mathbf{A}\boldsymbol{\mu}_{ik} - \mathbf{b}_q) = 0
\end{aligned} \tag{40}$$

By regrouping items, Eq.40 can be rewritten as:

$$\sum_{u=1}^{U} \sum_{(i,k)\in\omega_q} \sum_{t=1}^{T^u} \gamma_t^u(i,k) \cdot \boldsymbol{\Sigma}_{ik}^{-1} \cdot (\mathbf{o}_t^u - \mathbf{A}\boldsymbol{\mu}_{ik}) = \sum_{u=1}^{U} \sum_{(i,k)\in\omega_q} \sum_{t=1}^{T^u} \gamma_t^u(i,k) \cdot \boldsymbol{\Sigma}_{ik}^{-1} \cdot \mathbf{b}_q \tag{41}$$

Therefore, the bias vector in the class ω_q can be obtained as:

$$\mathbf{b}_q = \left[\sum_{u=1}^{U} \sum_{(i,k)\in\omega_q} \sum_{t=1}^{T^u} \gamma_t^u(i,k) \cdot \boldsymbol{\Sigma}_{ik}^{-1} \right]^{-1} \cdot \left[\sum_{u=1}^{U} \sum_{(i,k)\in\omega_q} \sum_{t=1}^{T^u} \gamma_t^u(i,k) \cdot \boldsymbol{\Sigma}_{ik}^{-1} \cdot (\mathbf{o}_t^u - \mathbf{A}\boldsymbol{\mu}_{ik}) \right] \tag{42}$$

Typically, $\boldsymbol{\Sigma}_{ik}$ are diagonal covariance matrices so that Eq.42 can be solved one dimension at a time and there is no need for matrix inverse operation.

2.5 Variance Adaptation

Given the adapted Gaussian mixture means, the diagonal covariance matrices are adapted in a non-constrained manner as described in (Gales, 1996):

$$\hat{\mathbf{\Sigma}}_{ik} = \mathbf{B}_{ik}^T \mathbf{H}_q \mathbf{B}_{ik} \tag{43}$$

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 factor of $\mathbf{\Sigma}_{ik}^{-1}$. That is,

$$\mathbf{\Sigma}_{ik}^{-1} = \mathbf{C}_{ik} \mathbf{C}_{ik}^{-1} \tag{44}$$

and

$$\mathbf{B}_{ik} = \mathbf{C}_{ik}^{-1} \tag{45}$$

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

$$\mathbf{H}_q = \frac{\sum_{(i,k)\in\omega_q} \mathbf{C}_{ik}^T [\sum_{u=1}^{U} \sum_{t=1}^{T^u} \gamma_t^u(i,k)(\mathbf{o}_t^u - \boldsymbol{\mu}_{ik})(\mathbf{o}_t^u - \boldsymbol{\mu}_{ik})^T] \mathbf{C}_{ik}}{\sum_{(i,k)\in\omega_q} \sum_{u=1}^{U} \sum_{t=1}^{T^u} \gamma_t^u(i,k)} \tag{46}$$

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

3 Formant-like Peak Alignment

As mentioned earlier, speech spectra of the same sound spoken by children and adults are spectrally mismatched primarily due to physiological differences. This mismatch is the major reason for performance degradation when acoustic models trained on adult speech are used to recognize children's speech. Fig. 3 shows two spectra for one speech frame (25ms) from an adult male and a 10-

year old boy for the /uw/ sound in the digit "two". Obvious pitch and formant differences can be observed from the two figures. If the spectrum can be reshaped by aligning the corresponding formants, then the spectral mismatch would be reduced.

In this paper, peaks are estimated using one set of Gaussian mixtures under the EM algorithm. This technique was proposed and applied in vocoder design and feature extraction in (Stuttle & Gales, 2001; Zolfaghari & Robinson, 1996, 1997). In this algorithm, the normalized magnitude of the speech spectrum for each frame is considered as a multi-mode probability density function and a Gaussian mixture model is used to fit it. The estimation is performed in an iterative manner. The estimated means, variances and mixture weights of the Gaussians correspond to the locations, bandwidths and amplitudes of the formants. Since the peaks fitted 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 are the same as in Fig. 3 and the utterances are the /uw/ sound in the digit "two" from which the speech frames in Fig. 3 are chosen. In the estimation, four Gaussian mixtures are used for the adult male and three for the child. From the figure, one can see that the estimated peaks fit the formants quite well.

To reduce the spectra mismatch, the estimated peaks are aligned by a piecewise 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 define $\omega_0^c = \omega_0^a = 1$. 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 defined as Eq.47.

$$\phi(l) = \begin{cases} \omega_i^c + \frac{\omega_{i+1}^c - \omega_i^c}{\omega_{i+1}^a - \omega_i^a} \cdot (l - \omega_i^a) & \text{for } \; l \in (\omega_i^a, \omega_{i+1}^a) \;\; \text{and} \;\; i = 0, \ldots, M-2. \\ \\ \omega_{M-2}^c + \frac{\omega_{M-1}^c - \omega_{M-2}^c}{\omega_{M-1}^a - \omega_{M-2}^a} \cdot (l - \omega_{M-2}^a) & \text{for } \; l \in (\omega_{M-1}^a, \omega_M^a). \end{cases} \tag{47}$$

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 much higher formants than adults. Therefore, in the same frequency range, they may have fewer formants than adults, as shown in Fig.4. By not requiring $\omega_M^c = \omega_M^a$, it is possible for the extra formants in adult spectra to disappear after alignment. Finally, we can generate the peak alignment matrix $\mathbf{R}$ in Eq.11 based on Eq.47 as:

$$r_{ij} = \begin{cases} 1, & \text{if } i = \text{round}(\phi(j)) \\ \\ 0, & \text{otherwise.} \end{cases} \tag{48}$$

Fig. 5 shows the piece-wise linear function computed according to Eq.47 aligning the first (F_1) and third (F_3) formant-like peaks in Fig. 4. Since formants gradually change from frame to frame, the median value for each peak is used. The two aligned peaks are marked out in the figure. In Fig. 6, the original spectrum of the child's speech (solid line) and the re-shaped 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 significantly reduced.

4 Experimental Results

Pseudo-codes which describe the implementation of training, adaptation and recognition stages of the proposed approach is shown in Fig.7.

Experiments are performed on connected digit strings from the TIDIGITS database. Acoustic models are trained on adult males and tested on children. Utterances from 55 male speakers are used in training. There are 77 utterances from each speaker with strings consisting of either 1, 2, 3, 4, 5 or 7 digits (there are no 6-digit strings in the database). Data from 5 boys and 5 girls are used in testing with 77 utterances from each speaker. In total, there are about 2500 digits in the test set. For each child, the adaptation utterances, which consists of 1, 4, 7, 10, 20 or 30 digits, are randomly chosen from the test set and not used in the testing. The speech signals are downsampled from 20 kHz to 8 kHz. Each speech frame is 25ms in length and a 10ms frame overlap is used in the analysis. Feature vectors are of 39 dimensions: 13 static features plus their first- and second-order derivatives. The features are computed using the CEP, MFCC1 and MFCC2 schemes and the derivatives are computed according to Eq.18 and Eq.19. In CEP and MFCC2, a 256-point FFT is used to obtain the magnitude spectrum, and the Mel-warped filter banks are composed of 23 triangular filters. In MFCC1, the linear frequency axis is first "oversampled" by a 1024-point FFT and then warped into 128 points on the Mel-frequency axis.

Acoustic HMMs are phoneme-based with a left-to-right topology. There are 18 monophones plus silence and inter-word short pause models. Each monophone has 2 to 4 states, depending on whether it is a vowel or consonant, with 6

Gaussian mixtures in each state.

The adaptation of the children's speech is carried out in an unsupervised manner. For each child, voiced segments are detected from the adaptation utterance via the traditional cepstrum peak analysis technique (Rabiner, 1978). Formant-like peaks are then estimated from the voiced segments by Gaussian mixtures. For a specific speaker, the median of peaks in each voiced segment is first 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.

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 the adult males and three for the children in the peak estimation procedure. 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 three features (CEP, MFCC1, MFCC2) are evaluated with the following peak alignment strategies:

- align F_1 and F_3, denoted as R(F_1,F_3)
- align F_3 only, denoted as R(F_3)
- align average F_3 which is estimated from the speech of all the children and males in the database, denoted as R($\bar{F}_3$)

Note that F_1 and F_3 refer to the formant-like peaks in the spectrum. They are not necessarily equal to the formant frequencies. We place an emphasis

on the F_3 region since F_3 has been shown to correlate with the vocal tract length (Fant, 1973). These strategies result in different alignment matrices $\mathbf{R}$ in Eq.10, and hence different transformation matrices $\mathbf{A}$ in Eq.15.

In Tables 1, 2, and 3, the performance of the formant-like peak alignment algorithm with the three alignment strategies ($R(F_1,F_3)$, $R(F_3)$ and $R(\bar{F}_3)$) is compared to the traditional VTLN with CEP, MFCC1 and MFCC2 features, respectively. VTLN is implemented in two ways, namely, speaker-dependent VTLN (VTLN1) and utterance-dependent VTLN (VTLN2). In both cases, warping factors are chosen from $[\ 0.7, 1.1\ ]$ with a stepsize of 0.05. For VTLN1, an average likelihood is first computed with the candidate warping factors across the adaptation utterances by forced alignment. The warping factor yielding the highest average likelihood is chosen as the optimal factor to scale the frequency axis in the feature extraction stage. For VTLN2, each test utterance is first recognized to obtain an initial transcription (hypothesis) and the warping factor with the highest likelihood for the utterance by forced alignment is applied to scale the frequency axis in feature extraction. The warped features are then re-recognized. In this way, adaptation data are ignored. VTLN by pooling adaptation and test data together to estimate the warping factor was also investigated but the results didn't improve over VTLN2.

Peak alignment with and without a bias are both presented in the tables. The results in parentheses are without a bias. In both cases, variance adaptation is performed. Depending on the amount of adaptation data available, the mean bias and variance transformation matrices are dynamically tied through a tree (Young et al., 2001) with 20 base classes. The threshold for node occupation is set to 50. For $R(\bar{F}_3)$, average F_3 peaks estimated from all the adult males and children' speech in the database are used in the alignment. The average

F_3 is around 2500 Hz for adult males and 3200 Hz for children.

From the tables, significant improvements over the baseline (no adaptation) are observed for all three features when peak alignment is used. The transformation matrix **A**, generated by aligning the formant-like peaks, contributes the most to the improved performance, and the bias **b** gives further improvements. Among the three alignment strategies R(F_1,F_3), R(F_3) and R($\bar{F}_3$), R(F_3) yields the best results which achieves, on average, 86.8%, 91.0% and 88.8% word error rate (WER) reduction over the baseline for CEP, MFCC1 and MFCC2, respectively. It is also very interesting to note that, since F_3 is closely related to the speaker's vocal tract length (Fant, 1973; Claes et al., 1998), aligning F_3 peak is related to vocal tract length normalization. Both R(F_3) and R($\bar{F}_3$) outperform the traditional VTLN when the number of adaptation digits is larger than 4. In particular, R(F_3) obtains 78.8%, 32.0% and 32.7% WER reduction over VTLN for CEP, MFCC1 and MFCC2, respectively.

Since the peak alignment algorithm bridges the front-end feature domain and back-end model domain by a linear transformation in terms of a linear frequency warping function, it can be considered as a special form of traditional MLLR. Therefore, it is interesting to compare the performance of the two. Fig. 8 shows the recognition results of MLLR, VTLN and peak alignment with varying numbers of adaptation digits. MFCC2 features are used in the experiments and the peak alignment is performed using R(F_3). The MLLR transformation matrices have a block diagonal form and are estimated based on the regression tree with 5 base classes. The threshold for node occupation is set to 500. From the figure, MLLR has poor performance when the adaptation data is limited, due to the unreliable estimation of model param-

eters. Peak alignment and VTLN significantly outperform MLLR under this condition because they utilize spectral information to reduce the mismatch in adaptation. As the amount of adaptation data increases, MLLR performance improves. Therefore, MLLR has an advantage when large amounts of data are available while VTLN is advantageous for limited amounts of data. In the proposed peak alignment algorithm, we primarily generate the linear mean transformation by aligning the formant-like peaks (similar to VTLN), on the basis of which, statistical approaches such as tree-based tied variance and bias adaptation are performed. In this way, the algorithm performs well for both large and limited amounts of adaptation data.

5 Summary and Conclusions

In this paper, the relationship between linear frequency warping in the front-end feature extraction and model transformation in the back-end is investigated in the discrete frequency domain with three feature extraction schemes: cepstra without Mel-scale warping, cepstra with Mel-scale warping and cepstra with Mel-warped triangular filter banks. A linear transformation is shown for the first two schemes. The transformation is linear for the third scheme only if certain approximations are made. The linear transformation is based on a discrete frequency mapping function (R) which can be considered as the discretized form of a general mapping function in the continuous frequency domain. Therefore, the linear transformation could cover a wide range of frequency warping functions.

The linear transform can be considered as a special case of standard MLLR and serves as a basis to deal with the sparse adaptation data problem. Utilizing

the linear transformation, a fast adaptation approach based on formant-like peak alignment is proposed. In this proposed approach, the transformation matrix $\mathbf{A}$ of Gaussian mixture means is computed deterministically after which the bias vector $\mathbf{b}$ is estimated statistically within the EM framework. Non-constrained Gaussian covariance adaptation are also conducted statistically. Both the estimation of biases and transformations of covariances are dynamically tied via a tree structure.

The proposed algorithm is utilized to adapt children's speech using acoustic models trained on adult data when the adaptation data is limited. Compared to traditional VTLN and MLLR with various amounts of adaptation data, significant improvements are observed. Best results are obtained when the peak alignment scheme uses speaker-specific F_3 information for the alignment. On average, with the widely-used MFCC feature with Mel-warped triangular filter banks, speaker-specific F_3 alignment outperforms VTLN by 33%, and MLLR by 54% with limited adaptation data.

6 Acknowledgements

This material is supported in part by NSF Grant No. 0326214. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Burnett, D and Fanty, M., 1996. Rapid unsupervised adaptation to children's speech on a connected-digit task, *Proc. of Int. Conf. on Spoken Language*

Processing, 1145-1148.

Claes, T., Dologlou, I., Bosch, L. and Compernolle, D., 1998. A novel feature transformation for vocal tract length normalization in automatic speech recognition, *IEEE Trans. on Speech and Audio Processing*, Vol. 11, No. 6, 603-616.

Das, S., Nix, D. and Picheny, M., 1998. Improvements in children's speech recognition performance, *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 433-436.

Deller, J.,Proakis, J. and Hansen, J., 1987. *Discrete-Time Processing of Speech Signals*, Prentice Hall.

Dempster, A.,Laird, N. and Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Vol. 39, No. 1, 1-38.

Digalakis, V. et al., 1999. Rapid speech recognizer adaptation to new speakers, *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*.

Ding, G., Zhu, Y., Li, C. and Xu, B., 2002. Implementing vocal tract length normalization in the MLLR framework, *Proc. of Int. Conf. on Spoken Language Processing*, 1389-1392.

Eide, E. and Gish, H., 1996. A parameter approach to vocal tract length normalization, *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 346-349.

Fant, G., 1973. *Speech Sounds and Features*, The MIT Press.

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

Gales, M., Pye, D., Wooland, P., 1996. Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation, *Proc. of Int. Conf. on Spoken Language Processing*, 1832-1835.

Gales, M., 1998. Maximum likelihood linear transformations for HMM-based speech recognition, *Computer Speech and Language*, Vol. 12, 75-98.

Gauvain, J.-L. and Lee, C.-H., 1994. Maximum A posteriori estimation for multivariate Gaussian mixture observations of Markov chains, *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, 291-298.

Gouvea, E. and Stern, R., 1997. Speaker normalization through formant-based warping of the frequency scale, *Proc. of European Conf. on Speech Communication and Technology*, 1139-1142.

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

Leggetter, C. and Woodland, P., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models, *Computer Speech and Language*, Vol. 9, 171-185.

Li, Q. and Russell, M., 2001. Why is automatic recognition of children's speech difficult, *Proc. of European Conf. on Speech Communication and Technology*, 2671-2674.

McDonough, J., Schaaf, T. and Waibel, A., 2004. Speaker adaptation with all-pass transforms, *Speech Communication*, Vol. 42, 75-91.

Pitz, M., Molau, S., Schluter, R. and Ney, H., 2001. Vocal tract normalization equals linear transformation in cepstral space, *Proc. of European Conf. on Speech Communication and Technology*, 2653-2656.

Pitz, M. and Ney, H., 2003. Vocal tract normalization as linear transformation of MFCC, *Proc. of European Conf. on Speech Communication and Technology*, 1445-1448.

Potamianos, A. and Narayanan, S., 2003. Robust recognition of children's speech, *IEEE Trans. on Speech and Audio Processing*, Vol. 11, No. 6, 603-

616.

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

Stuttle, M. and Gales, M., 2001. A mixture of Gaussians front end for speech recognition, *Proc. of European Conf. on Speech Communication and Technology*", 675-678.

Wegmann, S. et al., 1996. Speaker normalization on conversational telephone speech, *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 339-341.

Wilpon, J. and Jacobsen, C., 1996. A study of speech recognition for children and the elderly, *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 349-352.

Young, S. et al, 2001. *HTK User Manual 3.1* , Cambridge University.

Zolfaghari, P. and Robinson, T., 1996. Formant analysis using mixtures of Gaussians, *Proc. of Int. Conf. on Spoken Language Processing*, 1229-1232.

Zolfaghari, P. and Robinson, T., 1997. A segmental formant vocoder based on linearly varying mixture of Gaussians, *Proc. of European Conf. on Speech Communication and Technology*, 425-428.

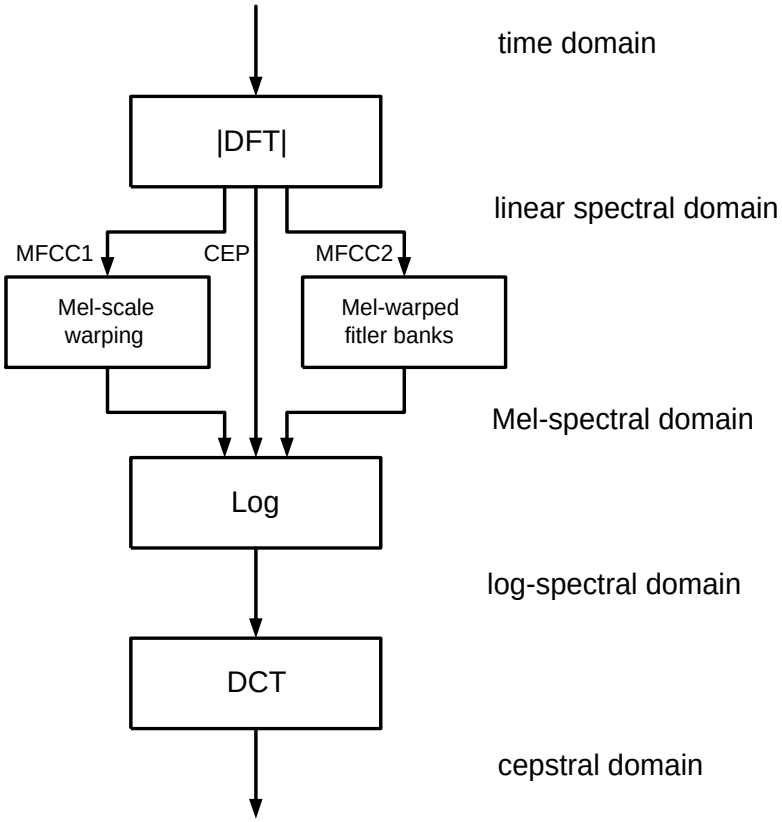

Fig. 1. Diagram of the three feature extraction schemes discussed. The feature CEP is computed with no Mel-scale warping. MFCC1 is computed with a Mel-scale warping function, and MFCC2 is computed with Mel-warped triangular filter banks.

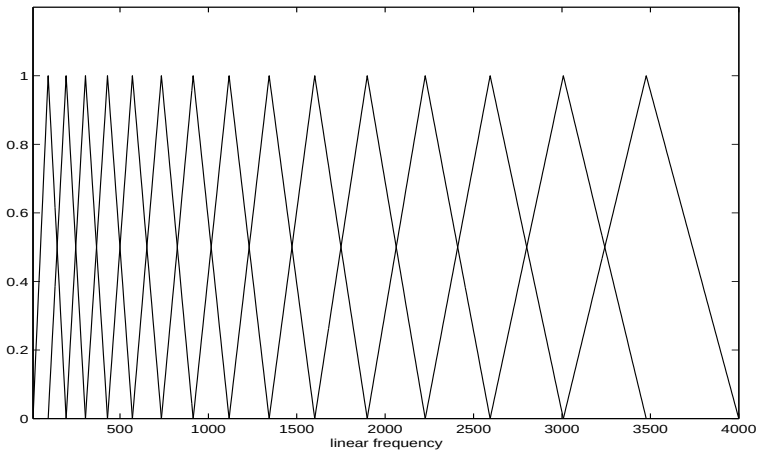

Fig. 2. Mel-scaled triangular filter bank.

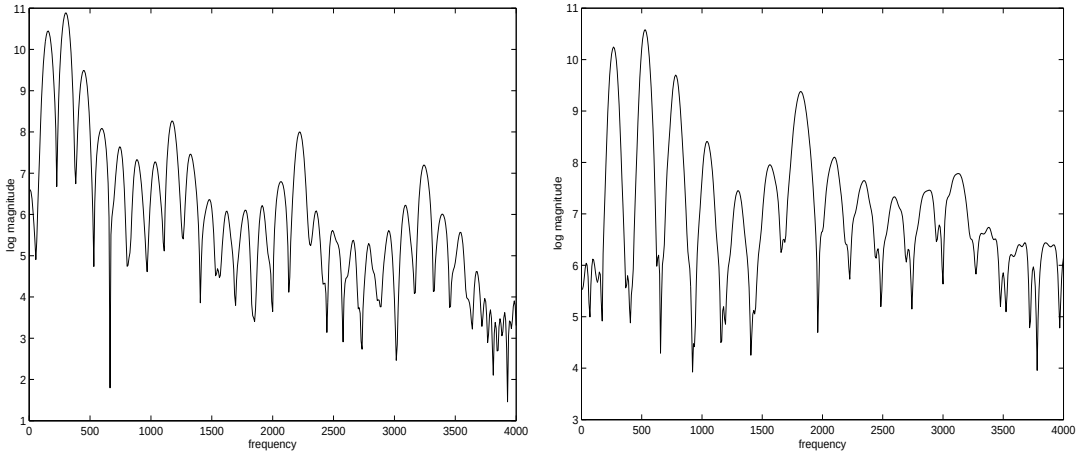

Fig. 3. Spectra for the steady part of the sound /uw/ in the digit "two" from an adult male (left) and a boy (right).

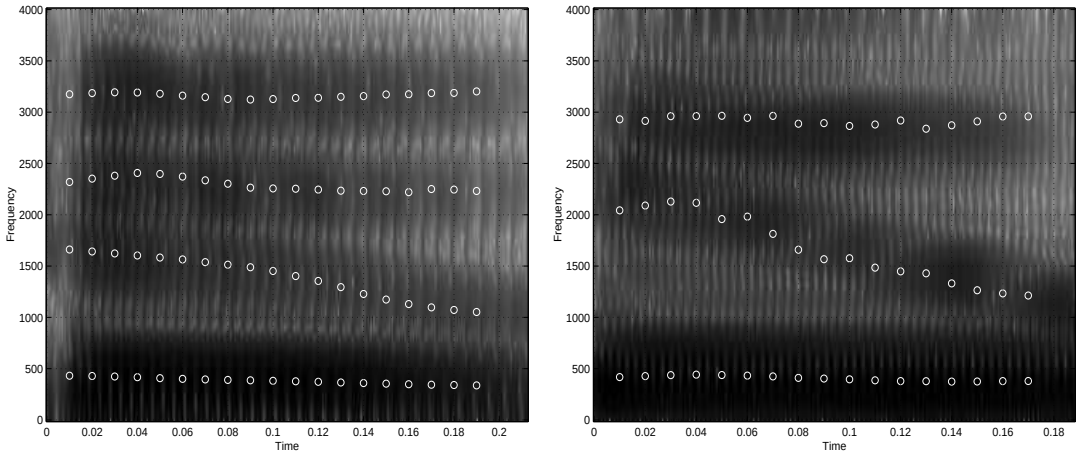

Fig. 4. Formants estimated (white circles) using Gaussian mixtures for the sound /uw/ in digit "two" from an adult male speaker (left) and a child speaker (right).

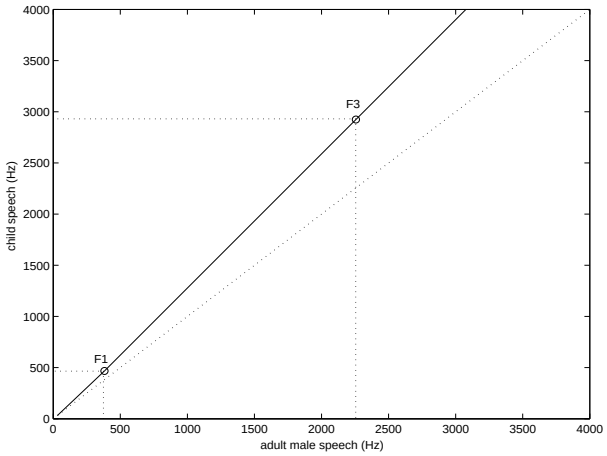

Fig. 5. Piece-wise linear function (solid line) which aligns the first and third formant-like peaks of the adult and child's speech in Fig. 4. The dotted line is the reference line for $y = x$.

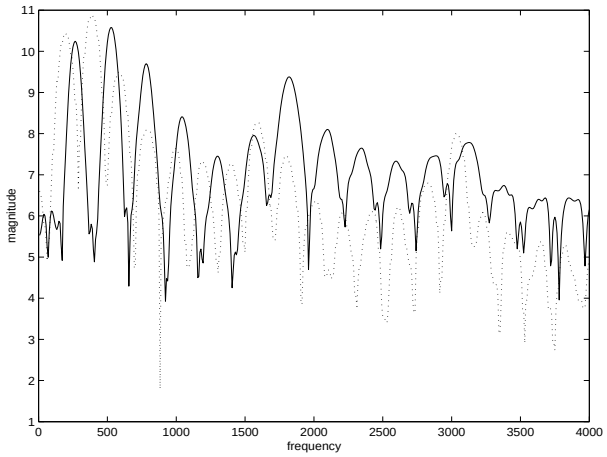

Fig. 6. Original boy's spectrum (solid line) and the re-shaped adult male's spectrum (dotted line) of Fig. 3.

	Number of adaptation digits					
algorithm	1	4	7	10	20	30
baseline	51.7	51.7	51.7	51.7	51.7	51.7
VTLN1	61.4	64.8	64.2	72.2	71.5	69.4
VTLN2	69.9	69.9	69.9	69.9	69.9	69.9
$R(F_1,F_3)$	90.3 (88.7)	89.1 (86.8)	91.5 (91.5)	92.6 (90.3)	95.4 (93.1)	95.9 (93.0)
$R(F_3)$	90.7 (89.6)	92.9 (91.9)	92.3 (91.9)	93.6 (92.2)	95.8 (93.9)	96.4 (93.7)
$R(\bar{F}_3)$	87.9 (88.6)	89.0 (87.9)	89.4 (88.5)	92.0 (91.2)	95.1 (92.1)	95.2 (92.4)

Table 1

Recognition accuracy of children's speech with CEP features for (1) baseline, or no adaptation, (2) VTLN1 (speaker-dependent), (3) VTLN2 (utterance-dependent), and (4) three peak alignment schemes ($R(F_1,F_3)$, $R(F_3)$ and $R(\bar{F}_3)$) with and without a bias vector. The results in the parentheses are without a bias. The acoustic models are trained on adult male data and tested on children's.

A. TRAINING

Train acoustic models ($\boldsymbol{\mu}, \boldsymbol{\Sigma}$) using adult speech data;

Select the standard speaker with the highest likelihood;

Locate voiced segments;

Estimate formant-like peaks of the spectrum and store as a reference;

B. ADAPTATION

Get the adaptation data from the test child speaker;

Locate voiced segments;

Estimate the formant-like peaks of the spectrum;

Align the peaks between the test and standard speakers;

Generate warping matrix $\mathbf{R}$ (Eq.47 and Eq.48);

Generate transformation matrix $\mathbf{A}$ (Eq. 26, Eq.30, or Eq.34);

Estimate bias $\mathbf{b}$ using the tree structure (Eq.42);

Adapt variance $\boldsymbol{\Sigma}$ using the tree structure (Eq.46);

C. RECOGNITION

Get the input speech signal for the test child speaker;

Perform recognition using the adapted acoustic models ($\hat{\boldsymbol{\mu}}$,$\hat{\boldsymbol{\Sigma}}$).

Fig. 7. Adaptation Algorithm.

	Number of adaptation digits					
algorithm	1	4	7	10	20	30
baseline	36.1	36.1	36.1	36.1	36.1	36.1
VTLN1	87.0	89.6	92.2	92.8	92.6	92.9
VTLN2	91.5	91.5	91.5	91.5	91.5	91.5
$R(F_1,F_3)$	89.4 (89.5)	89.0 (87.5)	91.9 (91.8)	92.8 (90.8)	94.9 (92.2)	95.9 (92.9)
$R(F_3)$	90.0 (89.8)	93.8 (93.1)	93.9 (93.8)	94.7 (93.0)	96.2 (94.3)	96.7 (94.7)
$R(\bar{F}_3)$	89.1 (89.2)	91.9 (91.6)	92.4 (90.7)	93.2 (92.0)	95.6 (93.4)	96.6 (93.2)

Table 2

Recognition accuracy of children's speech with MFCC1 features. See Table 1 caption for an explanation of the testing conditions.

	Number of adaptation digits					
algorithm	1	4	7	10	20	30
baseline	38.9	38.9	38.9	38.9	38.9	38.9
VTLN1	82.0	87.9	88.4	88.2	90.0	89.4
VTLN2	89.8	89.8	89.8	89.8	89.8	89.8
$R(F_1,F_3)$	86.1 (85.3)	90.1 (88.8)	91.1 (90.6)	91.0 (90.5)	93.5 (93.3)	95.1 (93.6)
$R(F_3)$	89.5 (87.7)	93.3 (92.8)	92.6 (92.8)	92.8 (91.9)	95.0 (94.3)	95.6 (94.3)
$R(\bar{F}_3)$	88.2 (87.1)	91.7 (91.4)	91.5 (91.7)	92.5 (91.3)	94.6 (93.4)	96.0 (93.2)

Table 3

Recognition accuracy of children's speech with MFCC2 features. See Table 1 caption for an explanation of the testing conditions.

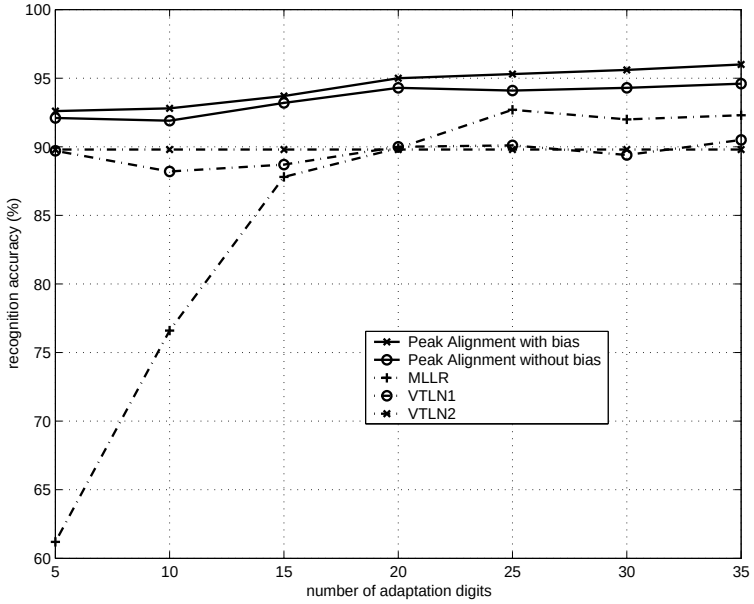

Fig. 8. Performance of MLLR, VTLN and the peak alignment algorithm using $R(F_3)$ with different numbers of adaptation digits.