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-filterbank-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 $Z$ space in the form of an all-pass transformation. In [4], the authors prove, in the continuous frequency space, that frequency warping equals a linear transform in the cepstral domain. However, since invertibility is required, the MFCC features studied in [4] are computed by Mel-scaling instead of Mel-filtered speech. In this paper, we investigate the above-mentioned relationships in the discrete frequency space. We show that for MFCCs computed with Mel-scaled triangular filter 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 filter bank. Let $S'$ denote a linear spectrum magnitude and $S''$ the corresponding MFCC features. According to the scheme shown in Fig. 1, we have:

$$S'' = C \cdot \log(M \cdot S')$$ (1)

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

$$M = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,K} \\ \theta_{2,1} & \cdots & \theta_{2,K} \\ \vdots & \vdots & \vdots \\ \theta_{N,1} & \cdots & \theta_{N,K} \end{bmatrix}$$

is the matrix form of the Mel filter bank where $L$ and $N$ are the number of samples in the linear and Mel-frequency domains, respectively. $\theta_{i,j}$ is the $j$th weight of the $i$th triangular filter and $K_i$ represents the number of non-zero coefficients in each filter. In other words, each row represents the components in one filter bin. In $M$, 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 $R_{L \times L}$, whose components are defined as:

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

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

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

where $C$ and $C^{-1}$ are the DCT and inverse DCT matrices, respectively, and $\log(\cdot)$ and $\exp(\cdot)$ are component-wise logarithm and exponential functions of a matrix. $T = M \cdot R \cdot M^*$ where $M^*$ 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 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. Obviously, the product of IM matrices is still an IM matrix. It is easy to show that the frequency warping matrix $R$ 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 $\log(\cdot)$:

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

where

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

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

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

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

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

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

$$
\mu_{\Delta Y} = \mathbf{A} \cdot \mu_{\Delta X}, \quad \mu_{\Delta^2 Y} = \mathbf{A} \cdot \mu_{\Delta^2 X}
$$

The Mel filter bank computation of MFCCs involves the summation of spectra samples within the frequency range of each triangular filter. 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 filter in the filterbank with the value of the center frequency sample (peak) of that filter, we are able to approximate $\mathbf{M}$ with an IM matrix $\tilde{\mathbf{M}}$:

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

where

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

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 filter 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:

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

Thus, $\tilde{\mathbf{M}}$, $\mathbf{M}^*$ and $\mathbf{R}$ are all IM matrices. $\mathbf{T} \approx \tilde{\mathbf{M}} \cdot \mathbf{R} \cdot \mathbf{M}^* \cdot \mathbf{C}^{-1}$ is also an IM matrix. In this way, a linear transformation of $\mu_Y$ and $\mu_X$ is guaranteed. That is,

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

where

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

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 filter banks are not invertible, the approximated linear relationship is difficult 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 five 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).
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).

Therefore, the number of parameters to be estimated is reduced 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 specific class $\omega_q$, the bias $b_q$ is shared across all the Gaussian mixtures $N(\mathbf{o}^i; \mu_{ijk}, \Sigma_{ijk})$ with $(i, k) \in \omega_q$ and is given by:

$$
b_q = \left[ \sum_{s=1}^u \sum_{(i,k)\in\omega_q} \sum_{l=1}^{T^u(i,k)} \gamma_6^s(i,k) \cdot \Sigma_{6k}^{-1} \right]^{-1} \left[ \sum_{s=1}^u \sum_{(i,k)\in\omega_q} \sum_{l=1}^{T^u(i,k)} \gamma_6^s(i,k) \cdot \Sigma_{6k}^{-1} \cdot (\mathbf{o}_s^i - \mathbf{A}_{ij} \mu_{ij}) \right], \tag{14}$$

where $U$ is the number of utterances in the adaptation data and $T^u$ is the number of frames in the $u$th utterance. $i \in \{1,2,\cdots,N\}$ and $k \in \{1,2,\cdots,M\}$ are the indices of state and mixture sets, respectively. $\gamma_6^s(i,k) = p(s^i_t = i, \xi^i_t = k|\mathbf{O}^i, \mathbf{X})$ is the posterior probability of being at state $i$ mixture $k$ at time $t$ given the $u$th observation sequence. $\mu_{ij}$ and $\Sigma_{ij}$ are the mean vector and covariance matrix associated with it, respectively. Typically, $\Sigma_{ij}$ 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]:

$$
\Sigma_{ij}^{-1} = \mathbf{B}_i^j \mathbf{H}_i \mathbf{B}_k
$$

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

$$
\Sigma_{ik}^{-1} = \mathbf{C}_i^k \mathbf{C}_k^{-1} \text{ and } \mathbf{B}_k = \mathbf{C}_k^{-1}
$$

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

$$
\sum_{(i,k)\in\omega_q} \sum_{s=1}^u \sum_{l=1}^{T^u(i,k)} \gamma_6^s(i,k) (\mathbf{o}_s^i - \mu_{ijk})(\mathbf{o}_s^i - \mu_{ijk})^T \mathbf{C}_{ik}^{-1} \sum_{(i,k)\in\omega_q} \sum_{s=1}^u \sum_{l=1}^{T^u(i,k)} \gamma_6^s(i,k)
$$

By forcing the $\mathbf{H}_i$'s off-diagonal terms to zeros, a diagonal covariance matrix $\Sigma_{ik}$ 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

Figure 5: Piece-wise linear function (left) and reshaped adult’s spectrum after peak alignment (right).
which can achieve better performance in case of limited adaptation data. Obviously, bias and variance adaptation can still benefit 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 first- 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 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. The first 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 first provide an initial hypothesis for each utterance, and warping factors within each class are then aligned. The Gaussian mixtures are initialized with transformation of 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: Recognition accuracy of children’s speech with MFCC features (TIDIGITS).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of adaptation digits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>no adaptation</td>
<td>38.9</td>
</tr>
<tr>
<td>MLLR</td>
<td>60.2</td>
</tr>
<tr>
<td>VTLN</td>
<td>89.8</td>
</tr>
<tr>
<td>LA-VTLN</td>
<td>85.0</td>
</tr>
<tr>
<td>PA</td>
<td>90.1</td>
</tr>
</tbody>
</table>

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

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. In the proposed approach, transformation of means is first determinedistically 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-filter-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.

7. References


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.