Speaker Adaptation with Limited Data using Regression-Tree based Spectral Peak Alignment

Shizhen Wang, Xiaodong Cui, *Member, IEEE,* and Abeer Alwan, *Senior Member, IEEE*

***Abstract*— Spectral mismatch between training and testing utterances can cause significant degradation in the performance of automatic speech recognition (ASR) systems. Speaker adaptation and speaker normalization techniques are usually applied to address this issue. One way to reduce spectral mismatch is to reshape the spectrum by aligning corresponding formant peaks. There are various levels of mismatch in formant structures. In this paper, regression-tree based phoneme- and state-level spectral peak alignment is proposed for rapid speaker adaptation using linearization of the vocal tract length normalization (VTLN) technique. This method is investigated in a maximum likelihood linear regression (MLLR)-like framework, taking advantage of both the efficiency of frequency warping (VTLN) and the reliability of statistical estimations (MLLR). Two different regression classes are investigated: one based on phonetic classes (using combined knowledge and data-driven techniques) and the other based on Gaussian mixture classes. Compared to MLLR, VTLN and global peak alignment [24], improved performance can be obtained for both supervised and unsupervised adaptations for both medium vocabulary (the RM1 database) and connected digits recognition (the TIDIGITS database) tasks. Performance improvements are largest with limited adaptation data which is often the case for ASR applications and these improvements are shown to be statistically significant.**

***Index Terms*— Speaker adaptation, regression tree, peak alignment, speech recognition, VTLN.**

I. Introduction

INTER-SPEAKER acoustic variations, which result in spectral mismatch, are a major cause of performance degradation in automatic speech recognition (ASR) systems [1]–[3]. These variations are mostly caused by differences in the vocal tract and vocal fold apparati. Typically, adult females have shorter vocal tract lengths (VTL) and smaller vocal cords than adult males, while children have shorter VTLs and smaller vocal cords than adults [4]. This implies, according to the linear speech production theory [5], that children have higher formant and fundamental (F_0) frequencies than adults, and female adults have higher formants and F_0 than male adult speakers. Consequently, the performance of speech recognition systems may be significantly different from speaker to speaker.

To maintain robust recognition accuracy, speaker adaptation and speaker normalization techniques are usually applied to reduce spectral mismatch between training and testing utterances [6]–[14]. Speaker adaptation attempts to compensate for spectral mismatch in the back-end acoustic model domain by statistically tuning the acoustic models to a specific speaker [6]–[9]. Speaker normalization, or vocal tract length normalization (VTLN), on the other hand, aims at reducing the effects of vocal tract variability in the front-end feature domain via linear, piece-wise linear or bilinear frequency warping [10], [11]. Other frequency warping functions have also been studied [12]–[14]. A class of transforms, known as all-pass transforms (APTs), was proposed to perform VTLN in [13] and studied in detail in [14] for two classes of conformal maps, namely rational all-pass transforms (RAPTs) and sine-log all-pass transforms (SLAPTs). It was demonstrated that using multiple-parameter warping functions is more effective than single-parameter ones [14]. In speaker adaptation, parameters are speaker-specific transformation matrices and biases estimated using the maximum likelihood (ML) or maximum a posterior (MAP) criterion [7], [15]. In VTLN, the parameters to be estimated are the frequency warping factors. Hence, to make reliable statistical estimation of adaptation parameters, speaker adaptation methods generally require more adaptation data than VTLN.

In recent years, considerable research efforts have been devoted to the relationship between frequency warping in the feature domain and the corresponding transformations in the model domain [16]–[24]. For computational efficiency, several studies have proposed the possibility of directly performing VTLN in the back-end model domain. In [16], vocal tract length normalization was implemented in an MLLR framework. Claes et al. in [17] proposed a linear approximation of VTLN for reasonably small warping factors using Taylor expansion. In [18] and [19], McDonough et al. derived the linearity of VTLN in cepstral space for two all-pass transforms (rational all-pass transforms and sine-log all-pass transforms) and conducted in [20] a detailed performance comparison with MLLR on a large vocabulary database. In [21], Pitz and Ney showed that, in the continuous frequency space, VTLN is equivalent to a linear transformation in the cepstral domain for MFCCs with Mel-frequency warping (instead of Mel-frequency filter banks). Umesh et al. in [22] showed that this VTLN linearization also holds in the discrete frequency space under the assumption of strictly limited quefrency range in the cepstral domain. Cui and Alwan in [23] and [24] discussed in detail the linearization of frequency warping for several different feature extraction schemes. Under certain approximations, they showed that frequency warping of MFCC features with Mel-frequency filter banks equals a linear transformation in

S. Wang and A. Alwan are with Department of Electrical Engineering, University of California, Los Angeles, CA, 90095, USA (email: szwang@ee.ucla.edu, alwan@ee.ucla.edu).

X. Cui is now with IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598, USA (email: cuix@us.ibm.com). This work was initiated during his doctoral studies at UCLA

Portions of this article were presented at Interspeech 2006, Pittsburgh, Pennsylvania, USA.

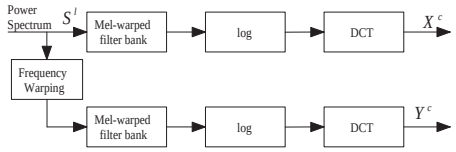


Fig. 1. Diagram of MFCC features extraction with and without frequency warping

the model domain.

In this paper, we focus on the linearization of frequency warping in [24], and develop it into a rapid MLLR-like speaker adaptation algorithm using regression-tree based phoneme- and state-level spectral peak alignment. With the proposed approach, the transformation matrices of the Gaussian mixture means are generated deterministically by aligning formant-like peaks in the spectrum through the linearization of frequency warping; the adaptation of biases and covariances are estimated statistically using the expectation maximization (EM) algorithm [25].

Recognition performance is tested on two databases, the DARPA Resource Management RM1 database [26] and the connected digits TIDIGITS database [27]. Experimental results show that the proposed approach leads to a significant improvement over MLLR, VTLN and global peak alignment for both supervised and unsupervised adaptation, especially with limited adaptation data.

The remainder of this paper is arranged as follows: in Section II, we briefly review the spectral peak alignment method and then illustrate various levels of mismatch in formant structures; in Section III, regression tree based phoneme- and state-level spectral peak alignment is discussed; and experimental setup and results are presented in Section IV. Summary and conclusions are made in Section V.

II. Alignment of Spectral Peaks

A. Linearization of frequency warping

Most state-of-the-art ASR systems utilize MFCC features. Fig. 1 shows the extraction of MFCC features with and without frequency warping. Let X^c denote the cepstral coefficients (MFCC features) of a speech signal, and Y^c be the cepstral coefficients after frequency warping in the spectral space. Strictly speaking, there is no simple linear relationship between Y^c and X^c due to the non-invertibility imposed by the Mel-frequency filter banks, but some reasonable approximations exist [17], [22], [23]. These approximations produce acceptable ASR performances with less computational cost than performing the warping directly in the spectral space. The approach proposed in [24] is a good example. In this subsection we briefly review that approach for MFCC features with Mel-frequency filter banks.

The approximation made on the Mel-frequency filter banks in [24] is to use only the central peak value to represent each triangular Mel-filter, as shown in Fig. 2. That is, to retain only the nonzero central value for each row in the Mel-frequency filter-bank matrix, and to set all other entries of the row to zero. Under such an approximation, frequency warping can be

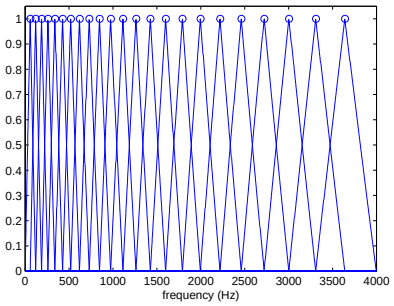


Fig. 2. Mel-frequency filter banks and the approximation made in the linearization of frequency warping in [24]: each triangular filter is represented only with its central peak value (the circle point)

implemented as a linear transformation in the cepstral domain, i.e.,

$$Y^c = \mathbf{A} \cdot X^c \tag{1}$$

where

$$\mathbf{A} = \mathbf{C} \cdot \mathbf{F_B} \cdot \mathbf{W} \cdot \mathbf{F_B^*} \cdot \mathbf{C^{-1}} \tag{2}$$

$\mathbf{C}$ is the DCT matrix, $\mathbf{F_B}$ is the approximated Mel-frequency filter-bank matrix, $\mathbf{W}$ is the frequency warping matrix, and $\mathbf{F_B^*}$ is the transformation matrix from Mel-frequency space to the linear frequency space such that $\mathbf{F_B^*} \cdot \mathbf{F_B} = \mathbf{I}$, and $\mathbf{C^{-1}}$ is the IDCT matrix. A more detailed derivation can be found in [24].

In ASR systems, both static and dynamic features are used. From Eq. (1), it is straightforward to show that dynamic features also hold this linearity, i.e.,

$$\Delta Y^c = \mathbf{A} \cdot \Delta X^c \tag{3}$$

$$\Delta^2 Y^c = \mathbf{A} \cdot \Delta^2 X^c \tag{4}$$

where Δ and Δ^2 represent the first and second order derivatives, respectively.

B. Definition of the frequency warping matrix

The frequency warping matrix $\mathbf{W}$ in Eq. (2) is defined as

$$w_{ij} = \begin{cases} 1, & \text{if } i\text{=round}(g_\alpha(j)) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where i and j are the frequency sample indices, $g_\alpha(\cdot)$ is a warping function in the discrete frequency space. In VTLN, some mathematically tractable warping functions such as linear, piece-wise linear, bilinear or quadratic functions are generally applied to perform frequency scaling. In the derivation of the transformation matrix $\mathbf{A}$ (Eq. (2)), however, no assumptions are made on the warping function, i.e., $g_\alpha(\cdot)$ can be any reasonable mapping function.

It was shown in [23] and [24] that aligning only the third formant (F_3) offers best ASR performance. In other words, $g_\alpha(j)$ is a linear warping function to align formant-like peaks in the spectrum space:

$$g_\alpha(j) = \alpha \cdot j \tag{6}$$

where

$$\alpha = \frac{F_{3,\text{ new speaker}}}{F_{3,\text{ standard speaker}}} \tag{7}$$

The reference standard speaker, chosen to represent the acoustic characteristics of the entire training set, is one of

the training speakers who yields the highest likelihood in the training stage. Since formant frequencies are gradually changing from frame to frame, the median values of F_3 over all voiced segments are used for each speaker in Eq. (7). Since F_3 has been shown to highly correlate to speaker's vocal tract length [5], this F_3 peak alignment is related to vocal tract length normalization.

Another choice for the reference standard speaker is to choose the speaker who has a neutral warping factor (α closest to 1). That is, to define the reference standard speaker as the one with F_3 closest to the mean F_3 value over the training set, or the one with the median F_3 value. Experiments with such a choice for the standard speaker resulted in slightly worse performance, partially due to the fact that by using the speaker with the highest training likelihood, we explicitly transform the acoustic parameters of each speaker toward a higher likelihood space.

C. Levels of mismatch in formant structure

As mentioned before, spectral mismatch is a major reason for performance degradation. There are various levels of mismatch in the formant structures, e.g. global average, phoneme level and state level. Fig. 3 illustrates formant estimation at these three levels. Global average formants are estimated using all the voiced segments (including vowels and voiced consonants) from the speech data; phoneme-level formants are estimated using speech segments for each phoneme; and state-level formants are estimated using segments in that state.

To illustrate different mismatch levels, we calculated the global average and phoneme-level F_3 warping factors for each test speaker from the RM1 and TIDIGITS database (see Section IV-A for detailed experimental settings). F_3 values were estimated using 10 adaptation utterances (digits) for each speaker. For the phoneme-level F_3 warping factors, we compared the three vowels in the classic vowel triangle: front vowel /IY/, mid vowel /AA/ (/AH/ for TIDIGITS, since there was no /AA/ in the data) and back vowel /UW/. The reference standard speaker for RM1 was a male adult with an average F_3 of 2524Hz and the F_3 values for /IY/, /AA/ and /UW/ were 2951Hz, 2354Hz and 2143Hz, respectively. For TIDIGITS, the reference speaker was a male with an average F_3 of 2537Hz and phoneme-level F_3 of 2968Hz, 2457Hz and 2268Hz for /IY/, /AH/ and /UW/, respectively. The global average F_3 warping factor was calculated according to Eq. (7), and the phoneme-level F_3 warping factor was defined in a similar way:

$$\alpha = \frac{F_{3,\ \text{/ph/ from new speaker}}}{F_{3,\ \text{/ph/ from standard speaker}}} \tag{8}$$

where $F_{3,\ \text{/ph/}}$ is the phoneme-level F_3 value for phoneme /ph/.

Figures 4 and 5 show the global average and phoneme-level F_3 warping factors. From these figures, we can see that the adult-to-adult warping factors (as in RM1, Fig. 4) are in the range of [0.96, 1.08], while the child-to-adult warping factors (as in TIDIGITS, Fig. 5) are in the range of [1.12, 1.26]. This is consistent with the fact that children's formant frequencies are higher than adults' [5]. Compared to Fig. 4, the warping factors in Fig. 5 show more dramatic changes from speaker to speaker. This agrees with the observation in [2] that children's speech demonstrate larger inter- and intra- speaker spectral variations than adults' speech.

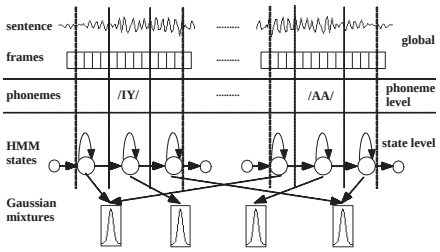


Fig. 3. Illustration of three levels of formant estimations. Boundaries are obtained through force alignment: dashed lines mark the boundaries of phonemes and dotted lines mark the boundaries for states

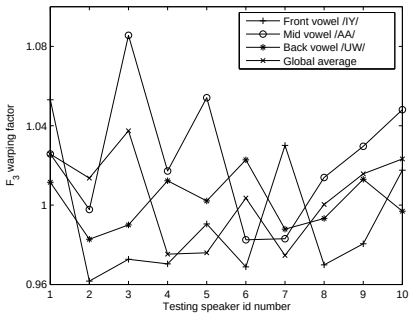


Fig. 4. F_3 warping factors for /IY/, /AA/, /UW/, and the global average for 10 test speakers (6 male and 4 female adults) from RM1

More importantly, these figures illustrate that phoneme-level warping factors may be very different from the global average and different phonemes may have different warping factors. For example, warping factors for /UW/ are around 1.0 for the adult-to-adult case and around 1.15 for the child-to-adult case; while the warping factors for /IY/ have a larger dynamic range. Thus, if phoneme-level or even lower state-level (instead of global average) warping factors are used to reduce spectral mismatch, we can expect better performance. This is the motivation for our proposed regression-tree based phoneme- and state-level spectral peak alignment methods in Section III-B[1].

III. Speaker adaptation algorithm

A. Adaptation using spectral peak alignment

The linearity in Eqs. (1), (3) and (4) bridges the gap between the front-end feature domain and the back-end model domain techniques and thus provides an efficient way of frequency warping. It can be used to perform rapid speaker adaptations on HMM Gaussian mixtures with mean $\boldsymbol{\mu}$ and diagonal covariance $\boldsymbol{\Sigma}$ in an MLLR-like manner [7]:

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \tag{9}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{L}\mathbf{H}\mathbf{L}^T \tag{10}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the transformed mean vector and covariance matrix, $\mathbf{L}$ is the Cholesky factor of the original covariance

[1]The amount of available adaptation data is another issue and is addressed in Section III-B and III-C

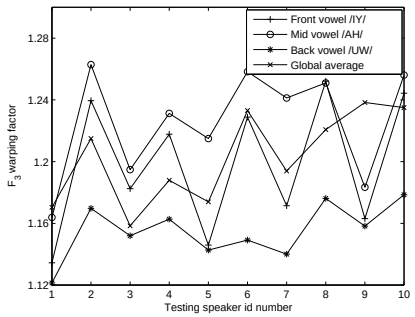


Fig. 5. F_3 warping factors for /IY/, /AH/, /UW/, and the global average for 10 test speakers (5 boys and 5 girls) from TIDIGITS

$\mathbf{\Sigma}$. The bias vector $\mathbf{b}$ in Eq. (9) and the covariance transformation matrix $\mathbf{H}$ in Eq. (10) are statistically estimated from the adaptation data under the maximum likelihood criterion [24],

$$\mathbf{b} = \left\{\sum_{m,k}\sum_{t=1}^{T}\gamma_{mk}(t)\mathbf{\Sigma}_{mk}^{-1}\right\}^{-1}\left\{\sum_{m,k}\sum_{t=1}^{T}\gamma_{mk}(t)\mathbf{\Sigma}_{mk}^{-1}(\mathbf{o}(t)-\mathbf{A}\boldsymbol{\mu}_{mk})\right\} \tag{11}$$

$$\mathbf{H} = \frac{\sum_{m,k}\left\{(\mathbf{L}_{mk}^{-1})^T\left[\sum_{t=1}^{T}\gamma_{mk}(t)(\mathbf{o}(t)-\hat{\boldsymbol{\mu}}_{mk})(\mathbf{o}(t)-\hat{\boldsymbol{\mu}}_{mk})^T\right](\mathbf{L}_{mk}^{-1})\right\}}{\sum_{m,k}\sum_{t=1}^{T}\gamma_{mk}(t)} \tag{12}$$

where T is the number of frames of the adaptation data, and m and k are the indices of state and mixture sets, respectively. $\gamma_{mk}(t)$ is the posterior probability of being at state m mixture k at time t given the observation $\mathbf{o}(t)$. By setting the off-diagonal terms of $\mathbf{H}$ to zero, the adapted covariance $\hat{\mathbf{\Sigma}}$ is also diagonal.

Unlike the statistical estimation in MLLR, the transformation matrix $\mathbf{A}$ here is generated deterministically based on Eq. (2), which depends only on the warping factors; while in MLLR a full or block-diagonal $\mathbf{A}$ needs to be statistically estimated. This would result in many more parameters than the deterministically generated $\mathbf{A}$ in Eq. (2). Though more parameters are powerful to capture slight differences among speakers, they may also lead to unreliable estimations (and thus unsatisfactory performance) with limited adaptation data. The $\mathbf{A}$ matrix generated using Eq. (2), however, can be more reliable than in MLLR when the amount of adaptation data is small; while the statistically estimated bias $\mathbf{b}$ and covariance transformation matrix $\mathbf{H}$ can benefit from increasing the amount of adaptation data. Hence, this peak alignment adaptation method performs well for varying amounts of adaptation data.

B. Regression-tree based spectral peak alignment adaptation

Several different approaches can be applied to perform formant-like peak alignment adaptation. In [24], speaker adaptation was employed as a global peak alignment, i.e. to estimate the average F_3 over all the adaptation data and generate the transformation matrix $\mathbf{A}$ according to Eq. (2) with the same warping factor (Eq. (7)) for all model units. When performing adaptation, all means of the HMM parameters share the same transformation $\mathbf{A}$. Since there is only one parameter (α) to be estimated, this global method has the potential of good performance for limited adaptation data.

As shown in Section II-C, there are various levels of mismatch in formant structures. Using only global average warping factors may not reduce the spectral mismatch uniformly for all phonemes. Since different phonemes may have different warping factors, we can use phoneme- or state-level warping factors to perform adaptation. This is the basic idea for the regression-tree based spectral peak alignment adaptation, i.e. to align similar (close in acoustic space) components in a similar way. This extension from global to regression-tree based peak alignment is similar to the expansion of MLLR from a global transform to many transforms especially when the adaptation data increase.

In this paper, two methods are considered to define regression classes: phoneme-based (using phoneme-level formants) and Gaussian mixture-based (using state-level formants). In the first method, units are classified based on phonetic knowledge and/or data-driven methods. For example, according to phonetic knowledge, phonemes can first be categorized into vowels and consonants, and then consonants can be further classified as voiced or unvoiced; vowels can further be clustered according to their phoneme-level F_3 values using data-driven methods. All model parameters for phoneme units with similar acoustic characteristics (phoneme-level formants) are placed together in the same regression class. Preliminary experiments showed that phonetic knowledge offers better performance when adaptation data are limited to less than 5 utterances, while the data-driven approach is superior when more data are available. Therefore, we chose to combine the two techniques.

Figure 6 shows an example of a regression tree based on tied phonetic knowledge and data-driven methods with eight base classes (terminal nodes) denoted as $\{2, 3, 4, 6, 7, 8, 9, 10\}$. Each phoneme belongs to one specific base class. During adaptation, the number of base classes is dynamically created depending on the amount of adaptation data. Since unvoiced consonants have no clear formant structure in their spectra, the transformation matrix $\mathbf{A}$ for unvoiced consonants is determined by the average F_3 over all voiced consonants in the adaptation data.

Since formant frequencies are gradually changing from frame to frame, it may be helpful to use further lower level formants in adaptations. In HMM models, each phoneme unit has several states, and states are represented with Gaussian mixtures. Hence, we can consider state-level formants, and define the regression tree based on Gaussian mixtures of the states. In this method, Gaussian mixture components (means and covariances) are clustered based on a measure of similarity. In each class, the state-level F_3 is estimated and averaged, and spectral peaks are then aligned with the same warping factor. Similar to global average and phoneme-level F_3 warping factors (Eq. (7) and (8)), state-level warping factor is defined as

$$\alpha = \frac{F_{3,\text{ state } m \text{ of /ph/ from new speaker}}}{F_{3,\text{ state } m \text{ of /ph/ from standard speaker}}} \tag{13}$$

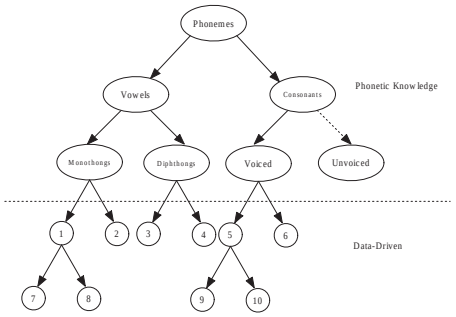


Fig. 6. An example of regression tree using combined phonetic knowledge and data-driven techniques for the phoneme-based approach. Phonemes are firstly categorized based on phonetic knowledge, and then further clustered according to their estimated F_3 values.

where $F_{3,\text{ state } m \text{ of /ph/}}$ is the state-level F_3 value of state m in phoneme /ph/.

For both phoneme-based and Gaussian mixture-based methods, regression trees are constructed based on the speaker independent training data and is independent of new speakers. The tree is constructed with a centroid splitting algorithm using a Euclidean distance measure. Each terminal node (base class) of the tree specifies a particular component groupings: phonemes for the phoneme-based regression tree and states for the Gaussian mixture-based regression tree. In the following sections, we will evaluate and compare the performance of these different approaches of peak alignment adaptation (PAA).

C. Integration of peak alignment with MLLR

As we will show in the next section, when adaptation data are limited, both approaches of PAA, namely phoneme-class and Gaussian mixture-class based, work well. With few parameters to estimate, PAA can handle one of the limitations of MLLR: unreliable parameter estimation for limited data. The performance of PAA, however, tends to saturate when more adaptation data become available, which is most obvious for global PAA. To some extent, this problem can be alleviated by increasing the number of regression classes. Since MLLR is able to offer better performance when more data are available, we attempt to integrate peak alignment with MLLR, i.e. to perform peak alignment first, followed by standard MLLR.

Given the peak alignment matrix $\mathbf{A}$ and the additive bias vector $\mathbf{b}$, the Gaussian mixture components of speaker specific models are re-estimated using the EM algorithm [25]. The auxiliary function is defined as

$$Q_{\mathcal{N}}(\lambda, \bar{\lambda}) = \sum_{m,k} \sum_{t=1}^{T} \gamma_{mk}(t) \log \mathcal{N}(\mathbf{o}(t); \mathbf{A}\bar{\boldsymbol{\mu}}_{mk} + \mathbf{b}; \bar{\boldsymbol{\Sigma}}_{mk}) \tag{14}$$

where $\mathcal{N}(\mathbf{o}(t); \mathbf{A}\bar{\boldsymbol{\mu}}_{mk} + \mathbf{b}; \bar{\boldsymbol{\Sigma}}_{mk})$ is the k th Gaussian mixture of state m. The maximum likelihood estimation of $\bar{\boldsymbol{\mu}}_{mk}$ and $\bar{\boldsymbol{\Sigma}}_{mk}$ can be derived from

$$\frac{\partial Q_{\mathcal{N}}(\lambda, \bar{\lambda})}{\partial \bar{\boldsymbol{\mu}}_{mk}} = 0 \tag{15}$$

$$\frac{\partial Q_{\mathcal{N}}(\lambda, \bar{\lambda})}{\partial \bar{\boldsymbol{\Sigma}}_{mk}} = 0 \tag{16}$$

respectively, which give

$$\bar{\boldsymbol{\mu}}_{\boldsymbol{mk}} = \left\{ \sum_{t=1}^{T} \gamma_{mk}(t) \mathbf{A}^T \boldsymbol{\Sigma}_{mk}^{-1} \mathbf{A} \right\}^{-1} \left\{ \sum_{t=1}^{T} \gamma_{mk}(t) \mathbf{A}^T \boldsymbol{\Sigma}_{mk}^{-1} (\mathbf{o}(t) - \mathbf{b}) \right\} \tag{17}$$

$$\bar{\boldsymbol{\Sigma}}_{mk} = \frac{\sum_{t=1}^{T} \gamma_{mk}(t) (\mathbf{o}(t) - \tilde{\boldsymbol{\mu}}_{mk})(\mathbf{o}(t) - \tilde{\boldsymbol{\mu}}_{mk})^T}{\sum_{t=1}^{T} \gamma_{mk}(t)} \tag{18}$$

where

$$\tilde{\boldsymbol{\mu}}_{mk} = \mathbf{A}\bar{\boldsymbol{\mu}}_{mk} + \mathbf{b} \tag{19}$$

$\tilde{\boldsymbol{\mu}}_{mk}$ represents the adapted speaker-specific Gaussian means.

This now can be viewed as a special case of standard speaker adaptive training (SAT) with only one speaker-dependent model [7], [28]. However, unlike the statistical estimation of the transforms as in SAT, which requires more adaptation data, the transformation matrix $\mathbf{A}$ is generated deterministically. Therefore, peak alignment has the potential for better performance than SAT with limited adaptation data. The integration with MLLR, denoted as PSAT in the following experiments, can be applied to global or regression-tree based peak alignment.

IV. Experiments

A. Experimental setup

Two different recognition tasks were carried out to evaluate the performance of the proposed algorithm. One was a medium vocabulary recognition task using the DARPA Resource Management RM1 continuous speech database [26], and another was a connected digits recognition task using the TIDIGITS database [27]. For the two databases, speech signals were firstly downsampled to 8kHz, and then segmented into 25ms frames, with a 10ms shift. Each frame was parameterized with a 39-dimensional feature vector consisting of 12 static MFCCs plus log energy, and their first-order and second-order derivatives.

For the RM1 database, triphone acoustic models were trained on the speaker independent (SI) portion of the database (72 speakers, 40 utterances from each speaker). Each triphone model had 3 states with 6 Gaussian mixtures per state. This set of SI models produced a baseline performance of 89.2% word recognition accuracy on the test set (10 speakers, 300 utterances from each speaker). Since the focus here is on rapid adaptation, for each speaker the adaptation data were limited to no more than 30 utterances for RM1 (or 35 digits for TIDIGTS), which corresponds to less than 2 minutes for RM1 (or 30 seconds for TIDIGITS). Adaptation data consisted of 1, 4, 7, 10, 15, 20, 25 or 30 utterances for each speaker, and they were randomly chosen from the speaker dependent portion of the database.

For the TIDIGITS task, acoustic models were trained on 55 adult male speakers and then tested on 10 children (5 boys and 5 girls) with 77 utterances consisting of 1, 2, 3, 4, 5, or 7 digits for each speaker. Acoustic HMMs were monophone-based with 4 states for vowels and 2 states for consonants, and 6 Gaussian mixtures per state. The baseline word recognition

accuracy was 38.9%. For each child, the adaptation data, which consisted of 1, 5, 10, 15, 20, 25, 30 or 35 digits, were randomly chosen from the test set and not used in the test.

In all adaptation experiments, a forward-backward alignment of the adaptation data was first implemented to assign each frame to a regression class (global adaptation can be considered as a special case of regression classes, with only one class.) For each class, formant-like peaks were then estimated. Depending on the amount of the adaptation data, different numbers of regression classes were experimentally tested, and the best performances were selected for comparison. Fig. 7 described the steps for both supervised (steps 2-4) and unsupervised (steps 1-4) peak alignment adaptation.

Gaussian mixture models were used to estimate formant-like peaks [29]. In the 4k Hz frequency range, adult speakers were observed to typically have four formants, while children had only three. Therefore, in the peak alignment procedure, four Gaussian mixtures were used for adults and three for children.

For comparison, speaker-specific VTLN was implemented based on a grid search over [0.8, 1.2] with a stepsize of 0.02. The scaling factor producing maximal average likelihood was used to warp the frequency axis [10]. Since VTLN is usually applied through warping the power spectrum, the Jacobian determinant is difficult to compute due to non-invertible Mel filter-bank operations. We approximated the Jacobian compensation by using the determinant of the transformation matrix $\mathbf{A}$ ($|\det \mathbf{A}|$).

B. Comparison of global and regression-tree based PAA versus MLLR and VTLN

Experiments were first conducted to compare the performance of global (GPAA), phoneme-class (PPAA) and Gaussian mixture-class (MPAA) based PAA with different numbers of adaptation utterances (or digits). In all experiments except otherwise specified, bias and diagonal covariance adaptation were performed for PAA. The block-diagonal MLLR adaptation with the optimal number of transforms was also performed for comparison. Figs. 8 and 9 illustrate the performance of GPAA, PPAA, MPAA, VTLN and MLLR. Not shown in the figures are recognition accuracies of MLLR with one adaptation utterance using RM1 (88.2%), and with one and five adaptation digits using TIDIGITS (40.5% and 57.0%, respectively.)

Fig. 8 shows that all three PAA methods can greatly improve the performance over the baseline (with no adaptation) in all cases; VTLN and GPAA provide the best performance with only one adaptation utterance, while PAA methods outperform VTLN in all other cases. MLLR, however, may produce worse performance than the baseline when only a small amount of adaptation data is available. For example, with one adaptation utterance, MLLR produces recognition accuracy of 88.2%, about one percent lower than the baseline. Compared to MLLR, PAA performs significantly better for limited adaptation data, with on average about 13.0% reduction of word error rate (WER) over MLLR for one and four adaptation utterances. With increasing adaptation data, MLLR offers better results than GPAA when the adaptation data are more than 15 utterances, while MPAA can outperform MLLR for 1-25 adaptation utterances. Since the covariance adaptation of PAA and MLLR is the same, the main differences between speakers seem to be characterized by the means of Gaussian components.

For unsupervised adaptation, perform step 1-4; for supervised adaptation, perform step 2-4

1) For unsupervised adaptation only: generating transcriptions
   - Locate voiced segments using cepstral peak analysis
   - Estimate formant-like peaks in the spectrum
   - Calculate scaling factor α (Eq. (7))
   - Generate transformation matrix $\mathbf{A}$ (Eq. (2))
   - Perform spectral peak alignment for each Gaussian mixture mean vector (without adaptation of bias and covariance) (Eq. (20))
   - Generate recognition hypotheses (with the partially adapted means) as transcriptions of the adaptation data
2) Dynamically determine the number of regression classes N based on the amount of adaptation data and cluster model parameters into N classes C_1, C_2, ..., C_N
3) Align with transcriptions to assign speech frames to regression classes
4) For each regression class C_i, $i \in \{1, 2, ..., N\}$
   - Estimate formant-like peaks in the spectrum
   - Calculate scaling factor α_i (Eq. (8) or (13))
   - Generate transformation matrix $\mathbf{A}_i$ (Eq. (2))
   - Estimate biases vector $\mathbf{b}_i$ and covariance transformation matrix $\mathbf{H}_i$ (Eq. (11), (12))
   - Adapt mean and covariance (Eq. (9), (10))

Fig. 7. The speaker adaptation algorithm using regression-tree based spectral peak alignment for both supervised and unsupervised adaptations

Among the three PAA methods, MPAA performs the best, and significant improvements can be achieved by using regression-tree based PAA over global PAA. On average, more than 11% WER reduction is obtained with MPAA over GPAA. The advantage of MPAA over GPAA becomes greater with increasing adaptation data. For the two regression-tree based PAAs, MPAA performs slightly better than PPAA in all cases. This is because these three PAA methods work at different levels to reduce spectral mismatch: MPAA at the state level, PPAA at the phoneme level and GPAA at the global level. As discussed in Section II-C and III-B, lower-level (phoneme- or state-level) alignment is expected to be more powerful than the global average to capture subtle differences between phonemes or even states, provided that the parameters are reliably estimated. Compared to global average formants used in GPAA, phoneme-level (PPAA) and state-level (MPAA) formants need to be estimated with more parameters and thus require more adaptation data. This explains the performance

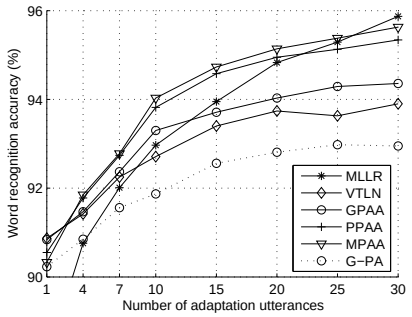


Fig. 8. Performance of VTLN, MLLR, G-PA, GPAA, PPAA and MPAA using RM1 for supervised adaptation.

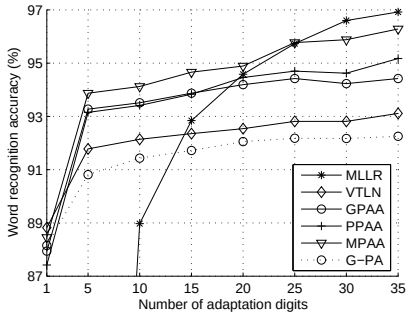


Fig. 9. Performance of VTLN, MLLR, G-PA, GPAA, PPAA and MPAA using TIDIGITS for supervised adaptation.

curves of GPAA, PPAA and MPAA in Fig. 8.

Experimental results for TIDIGITS (Fig. 9) demonstrate similar trends to Fig. 8. This similarity shows that performance improvements achieved by PAA are consistent across different tasks. Comparing Figs. 8 and 9, we notice that improvements for TIDIGITS are more significant than that for RM1 database: with only one adaptation digit (or utterance), more than 80.0% WER reduction over the baseline was obtained for TIDIGITS, while for RM1 the WER reduction over the baseline was about 10.5% .

The more significant improvements with the TIDIGITS database can be explained as follows. The basic idea for PAA is to reduce spectral mismatch by aligning formant-like peaks using estimated F_3 values. The performance improvement will be more obvious if the F_3 difference between the new speaker and the standard speaker is significant, which is the case for TIDIGITS: for adult males the typical F_3 is about 2500Hz, and for children it is 3100 Hz. On the other hand, if the F_3 of the new speaker is very close to that of the standard speaker as with the RM1 database which has only adult speakers, the effect of peak alignment will be less pronounced. An extreme case is when the new speaker has exactly the same global average F_3 value as the standard speaker. In this case, the global average warping factor α will be 1 (Eq. (7)), and the warping matrix $\mathbf{W}$ will be an identity matrix (Eq. (5)), which will result in an identity transformation matrix $\mathbf{A}$ (Eq. (2)) for global peak alignment (GPAA). [2] Thus, theoretically, in this case global peak alignment will have little effect on reducing spectral mismatch, resulting in marginal, if any, performance improvement. This is also supported by experimental results with the RM1 database using global peak alignment with only $\mathbf{A}$: the speaker with α closest to 1 shows only 1.5% average improvement, while the speaker with the largest α achieves over 10% improvement.

Regression tree based peak alignment may still perform well even in the case where global peak alignment fails to provide satisfactory improvement, since regression tree based peak alignment utilizes phoneme or state level formant information (instead of global average as in global peak alignment), and all phoneme- or state-level formant values from two different speakers may not be identical. This is another advantage of regression tree based peak alignment over global peak alignment.

Since PAA is based on an approximate linearization of VTLN, it is also of interest to study how good this approximation is. We compared the performance of VTLN and GPAA using only $\mathbf{A}$ (without bias and covariance adaptation), [3] denoted as G-PA in Figs. 8 and 9. G-PA performs a little worse than VTLN; the differences, however, are small. This means that the linearization is a good approximation to VTLN, and the adaptation of bias and covariance contributes to the better performance of GPAA.

The peak alignment technique was also compared in [30] with VTLN based on parameters estimated directly using maximum likelihood criterion, i.e., α was statistically estimated under ML criterion instead of being defined as the formant frequency ratios (Eq. (7), (8) or (13)) or being determined using a grid search. Experimental results showed that GPAA achieves similar performance to the ML-based VTLN. Peak alignment is, however, more efficient from the computational point of view. In addition, MPAA outperforms ML-based VTLN when the adaptation data are more than 5 utterances.

C. Comparison of PAA, PSAT and MLLR-SAT

In this section, we compare PAA versus PSAT which combines peak alignment followed by MLLR. Gaussian mixture-class based peak alignment (MPAA) is considered as the reference which performs the best among the three PAA methods. PSAT is applied in two ways: based on GPAA (PSAT-GPAA) and on MPAA (PSAT-MPAA).

The performance of MLLR, MPAA and PSAT are shown in Figs. 10 and 11. Compared to MPAA, PSAT (both PSAT-GPAA and PSAT-MPAA) shows better performance with improvements, on average, of about 6% with RM1 and 20% with TIDIGITS; compared to GPAA (Figs. 8 and 9), the improvements are even more significant (16% with RM1 and 24% with TIDIGITS). Improvement trends are consistent in all cases especially with more adaptation data. As to the two PSAT methods, PSAT-GPAA is a little better with a small amount of adaptation data, while PSAT-MPAA outperforms PSAT-GPAA when the adaptation data are more than 10 utterances.

[2] Strictly speaking, $\mathbf{A}$ will not be identity due to the approximation made in the linearization of VTLN. However, $\mathbf{A}$ will be very close to identity with diagonal entries very close to 1 and off-diagonal entries close to 0.

[3] This configuration of GPAA can be viewed as a direct linear approximation of VTLN

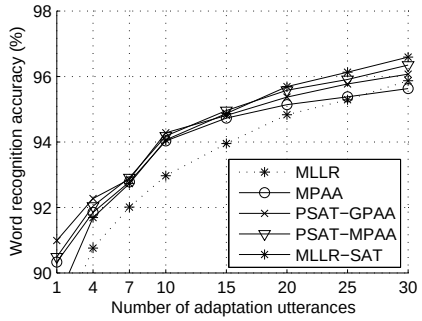


Fig. 10. Performance of MPAA, MLLR, PSAT and MLLR-SAT using RM1 for supervised adaptation.

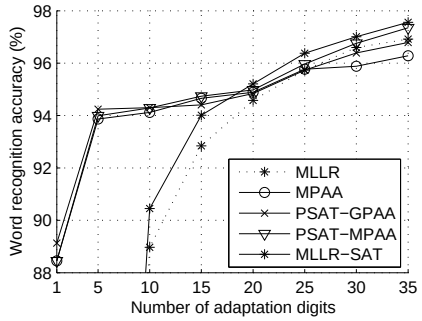


Fig. 11. Performance of MPAA, MLLR, PSAT and MLLR-SAT using TIDIGITS for supervised adaptation.

Compared to MLLR, the performance of PSAT-MPAA is superior in all experiments with on average 14% improvement for RM1 and 23% for TIDIGITS, though the difference becomes small as adaptation data increase. Significance analysis shows that for the p-level less than 0.05, the improvement of PSAT-MPAA over MLLR is statistically significant. This indicates that PSAT can take advantage of PAA for reliable parameter estimations with limited adaptation data, and of MLLR for statistical parameter estimations with sufficient adaptation data. Another advantage of PSAT is that it can still perform well even when there is no difference in global average F_3 values between the new speaker and the standard speaker, in which case PSAT-GPAA becomes equivalent to MLLR.[4]

Since PSAT can be viewed as a special case of MLLR-SAT, which is an alternative implementation of SAT through constrained MLLR transformations [7], it is interesting to compare their performance. The experiments follow the steps described in [7] and use block diagonal transforms in MLLR-SAT.

The performance of MLLR-SAT is shown in Figs. 10 and 11. MLLR-SAT provides better performance than MLLR, decreasing WER by about 10% on average. However, it performs similarly to MLLR, i.e. they both require a certain amount of adaptation data (more than 20 utterances) for robust and satisfactory performance. In contrast, PSAT is more robust for limited data, especially PSAT-GPAA, which achieves more than 17% WER reduction over MLLR-SAT for one adaptation utterance. With the increase of adaptation data, PSAT-MPAA performs better than PSAT-GPAA and provides comparable performance with MLLR-SAT. From the computational point of view, PSAT is more efficient than MLLR-SAT, with only several warping factors instead of a full or block diagonal matrix $\mathbf{A}$ to be estimated. So PSAT is more suitable for rapid adaptation where the available enrollment data for a new speaker is limited to only several utterances.

Another rapid adaptation method is Maximum A Posterior Linear Regression (MAPLR) [31]–[33]. MAPLR incorporates prior knowledge into the linear regression adaptation of means and covariances by using MAP criterion. The hyperparameters (parameters needed to describe the prior distribution) are estimated based on an empirical Bayes (EB) approach [33] and/or the structural information of the models [32]. Provided that appropriate priors are chosen, MAPLR may significantly outperform MLLR. The performance of MAPLR, however, is highly dependent on the choice of prior distributions [31]. Like MAPLR, prior knowledge can also be integrated into PAA through the MAP estimation of the bias $\mathbf{b}$ and the covariance transforms $\mathbf{H}$. In this paper, however, we focus on PAA in the MLLR framework and will explore the PAA in the MAPLR framework in future work.

D. Comparison of supervised and unsupervised adaptation

The previous adaptation experiments are implemented in a supervised way where the true transcription is known. Unsupervised adaptation can be performed by first generating the transcription through an initial recognition pass. Before this initial recognition, global peak alignment (without adaptation of bias and covariance) is conducted to reduce spectral mismatch. According to Eqs. (2), (5), (6) and (7), the generation of matrix $\mathbf{A}$ is only dependent on the warping factor α which can be estimated from voiced segments and thus requires no transcription knowledge. For each test speaker, formant-like peaks are estimated from the voiced segments of the adaptation utterance; voicing is detected using the cepstral analysis technique [34]. Spectral peaks are then aligned with the average F_3, i.e. Gaussian mixture means are adapted according to the following equation:

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} \tag{20}$$

The performance of supervised and unsupervised adaptation is shown in Tables I and II. It should be noted that the performance listed here for supervised and unsupervised adaptation was based on different numbers of regression classes: in all cases, the number of classes for unsupervised adaptation was smaller than that of the corresponding supervised case. For example, for the RM1 database, when the adaptation data consist of 20 utterances, 105 Gaussian mixture classes were found to give the best results for unsupervised adaptation, while 150 classes were optimal for supervised adaptation. The number of regression-tree base classes used in MPAA and PSAT for each testing case is given in the tables in the row labeled "# of classes". The optimal number of base classes for MLLR can be different.

From these tables, compared to supervised adaptation, unsupervised peak alignment adaptation performs slightly worse

[4]PSAT can be considered as the combination of PAA and MLLR. As discussed in Section IV-B, in this case GPAA has little effect on reducing spectral mismatch, and MLLR is credited with the performance improvements.

TABLE I

WORD RECOGNITION ACCURACY USING RM1 FOR SUPERVISED AND UNSUPERVISED (IN PARENTHESES) ADAPTATION.

	Number of adaptation utterances							
	1	4	7	10	15	20	25	30
MLLR	88.2 (86.5)	90.8 (89.3)	92.0 (90.5)	93.0 (91.5)	94.0 (92.3)	94.8 (93.3)	95.3 (94.4)	95.9 (94.6)
GPAA	90.8 (90.7)	91.5 (91.3)	92.4 (92.2)	93.3 (93.1)	93.7 (93.6)	94.0 (93.9)	94.3 (94.0)	94.4 (94.2)
MPAA	90.3 (88.7)	91.9 (90.2)	92.8 (91.3)	94.0 (93.4)	94.7 (94.0)	95.1 (94.2)	95.4 (94.6)	95.6 (94.9)
PSAT-MPAA	90.5 (89.0)	92.0 (90.7)	92.9 (91.8)	94.2 (93.7)	95.0 (94.6)	95.6 (94.9)	95.9 (95.2)	96.4 (95.6)
# of classes	10 (5)	40 (20)	50 (35)	75 (60)	100 (80)	150 (105)	175 (135)	225 (195)

TABLE II

WORD RECOGNITION ACCURACY USING TIDIGITS FOR SUPERVISED AND UNSUPERVISED (IN PARENTHESES) ADAPTATION.

	Number of adaptation digits							
	1	5	10	15	20	25	30	35
MLLR	40.5 (38.9)	57.0 (55.3)	88.9 (88.2)	92.8 (92.3)	94.6 (94.5)	95.7 (95.1)	96.6 (95.9)	96.9 (96.1)
GPAA	87.9 (87.7)	93.3 (93.2)	93.5 (93.4)	93.9 (93.8)	94.2 (94.1)	94.4 (94.3)	94.2 (94.2)	94.4 (94.4)
MPAA	88.5 (86.4)	93.9 (92.3)	94.1 (94.0)	94.7 (94.1)	94.9 (94.5)	95.8 (95.3)	95.9 (95.1)	96.3 (95.2)
PSAT-MPAA	88.5 (86.4)	94.0 (92.3)	94.3 (94.1)	94.7 (94.2)	95.0 (94.7)	96.0 (95.6)	96.8 (96.2)	97.4 (96.7)
# of classes	5 (3)	25 (20)	30 (25)	40 (25)	55 (30)	80 (50)	100 (75)	125 (95)

in all experimental cases, but the difference is not large: 0.5% and 0.8% absolute WER increase for PSAT-MPAA using RM1 with 10 and 30 adaptation utterances, respectively; 0.2% and 0.7% absolute WER increase for PSAT-MPAA using TIDIGITS with 10 and 35 adaptation digits. There are two possible reasons for this small difference. One is that after the global peak alignment, the partially adapted models produce a high recognition accuracy and thus an acceptable labeling of the adaptation data. The other is that with a smaller number of classes, it is more likely for unsupervised adaptation to reduce the effect of misclassified frames (due to the initial recognition errors) and thus to generate robust estimation for the adaptation parameters. This explains why the unsupervised GPAA performs almost the same as the supervised case, especially for the highly mismatched TIDIGITS database with the differences being less than 0.2% in all cases. Compared to GPAA, unsupervised PSAT-MPAA achieves on average 6.8% and 12.7% WER reduction for RM1 and TIDIGITS, respectively.

E. Significance analysis

We use the matched-pair test proposed in [35] to analyze whether the performance differences between MLLR and regression-tree based peak alignment (MPAA) are statistically significant for both the supervised and unsupervised adaptations. Tables III and IV show the significance levels (p-value) of MPAA compared to MLLR for supervised speaker adaptation with various amounts of adaptation data.

These tables show that, for a given significance level $\beta = 0.05$, the average performance differences between MPAA and MLLR are statistically significant using both RM1 and TIDIGITS for supervised adaptation. Examining the significance levels for different amounts of adaptation data, we can see that the performance improvements of MPAA over MLLR are more significant for limited adaptation data (less than 20 utterances). This is due to the deterministically generated transforms $\mathbf{A}$ in MPAA versus the unreliable statistically estimated $\mathbf{A}$ in MLLR because of not enough adaptation data. Similar conclusions also hold for the unsupervised adaptations using both the RM1 database and the TIDIGITS database.

Analysis on PSAT-MPAA and MPAA doesn't show significant differences between these two algorithms. The performance improvements of PSAT-MPAA over MLLR, however, are statistically significant in all the testing cases, at significance levels less than 0.05.

V. SUMMARY AND CONCLUSION

Various levels of spectral mismatch in formant structures cause ASR systems to perform unsatisfactorily. Regression tree based spectral peak alignment is proposed as a rapid speaker adaptation to reduce phoneme- and state-level spectral mismatch. This method is investigated in an MLLR-like framework based on the linearization of VTLN. In the proposed approach, the transformation matrix for Gaussian mixture means is generated deterministically by aligning phoneme- and state-level formant-like peaks in the spectrum; adaptation of the bias and covariance is estimated using the EM algorithm. This method can be viewed as a combination of VTLN and MLLR. On the one hand, like VTLN, the transformation matrix for means has fewer parameters than MLLR to be estimated, which is advantageous for limited adaptation data. On the other hand, like MLLR, biases and covariances are adapted using the EM algorithm. Statistical estimation has an advantage when large amounts of adaptation data are available. Hence, the proposed approach has the potential of good performance for both limited and large amounts of adaptation data.

The performance of this peak alignment approach is evaluated on both medium vocabulary (the RM1 database) and connected digits recognition (the TIDIGITS database) tasks. In both tasks, experimental results show that through peak alignment adaptation significant performance improvements can be achieved even for very limited adaptation data, with state-level peak alignment (MPAA) performing the best. When sufficient adaptation data are available, peak alignment adaptation offers results similar to or slightly worse than MLLR. The PSAT method which integrates peak alignment with MLLR, however, shows better performance than MLLR and comparable performance with MLLR-SAT in all cases. Another merit of this regression-tree based spectral peak alignment is that when implementing adaptation in an unsupervised way, only a slight

TABLE III

SIGNIFICANT ANALYSIS OF PERFORMANCE IMPROVEMENTS OF MPAA OVER MLLR USING RM1 FOR SUPERVISED ADAPTATION

# of utterances	1	4	7	10	15	20	25	30
p-value	0.007	0.009	0.013	0.018	0.026	0.031	0.039	0.043

TABLE IV

SIGNIFICANT ANALYSIS OF PERFORMANCE IMPROVEMENTS OF MPAA OVER MLLR USING TIDIGITS FOR SUPERVISED ADAPTATION

# of digits	1	5	10	15	20	25	30	35
p-value	0.001	0.003	0.008	0.012	0.027	0.043	0.025	0.034

performance degradation is observed compared to supervised adaptation.

ACKNOWLEDGMENT

This work was supported in part by NSF Grant No. 0326214 and by a fellowship from the Radcliffe Institute for Advanced Study to Abeer Alwan. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect views of the NSF.

REFERENCES

[1] X. Huang and K.F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 1(2), pp. 150-157, 1993.

[2] S. Lee, A. Potamianos and S. Narayanan, "Acoustic of children's speech: developmental changes of temporal and spectral parameters, " *J. Acoust. Soc. Am.*, vol. 105(3), pp. 1455-1468, 1999.

[3] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and elderly," in *Proc. ICASSP*, vol. I, pp. 349-352, 1996.

[4] H. Wakita, "Normalization of vowels by vocal tract length and its application to vowel identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, pp. 183-192, 1977.

[5] G. Fant, *Acoustic Theory of Speech Production.* The Hague, The Netherlands: Mouton, 1960.

[6] V. Digalakis, D. Rtischev and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Processing*, vol. 3(5), pp. 357-366, 1995.

[7] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12(2), pp. 75-98, 1998.

[8] V. Digalakis, S. Berkowitz, E. Bocchieri, C. Boulis and W. Byrnc, "Rapid speech recognizer adaptation to new speakers," in *Proc. ICASSP*, pp. 765-768, 1999.

[9] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[10] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Processing*, vol. 6(1), pp. 49-60, 1998.

[11] S. Wegmann, D. McAllaster, J. Orloff and B. Peskin, "Speaker normalisation on conversational telephone speech," in *Proc. ICASSP*, vol. I, pp. 339-341, 1996.

[12] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP*, pp. 346-349, 1996.

[13] J. McDonough, W. Byrne and X. Luo, "Speaker normalization with all-pass transforms," in *Proc. ICSLP*, vol. VI, pp. 2307-2310, 1998.

[14] J. McDonough, "Speaker compensation with all-pass transforms," Ph.D. dissertation, Johns Hopkins University, 2000.

[15] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2(2), pp. 291-298, 1994.

[16] G. Ding, Y. Zhu, C. Li and B. Xu, "Implementing vocal tract length normalization in the MLLR framework," in *Proc. ICSLP*, pp. 1389-1392, 2002.

[17] T. Claes, I. Dologlou, L. Bosch and D.V. Compernolle, "A novel feature transformation for vocal tract length normalization in automatic speech recognition," *IEEE Trans. Speech Audio Processsing*, vol. 11(6), pp. 549-557, 1998.

[18] J. McDonough and W. Byrne, "Speaker adaptation with all-pass transforms," in *Proc. ICASSP*, pp. 757-760, 1999.

[19] J. McDonough, T. Shaaf and A. Waibel, "Speaker adaptation with all-pass transforms," *Speech Commnication*, vol. 42, pp. 75-91, 2004.

[20] J. McDonough and A. Waibel, "Performance comparisons of all-pass transform adaptation with maximum likelihood linear regression," in *Proc. ICASSP*, pp. I313-I316, 2004.

[21] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC," in *Proc. Eur. Conf. Speech Communication and Technology*, pp. 1445-1448, 2003.

[22] S. Umesh, A. Zolnay and H. Ney, "Implementing frequency-warping and VTLN through linear transformation of conventional MFCC," in *Proc. Interspeech-2005*, pp. 269-272, 2005.

[23] X. Cui and A. Alwan, "MLLR-like speaker adaptation based on linearization of VTLN with MFCC features," in *Proc. Interspeech-2005*, pp. 273-276, 2005.

[24] X. Cui and A. Alwan, "Adaptation of children's speech with limited data based on formant-like peak alignment," *Computer Speech and Language*, vol. 20(4), pp. 400-419, 2006.

[25] A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39(1), pp. 1-38, 1977.

[26] P. Price, W.M. Fisher, J. Bernstein and D.S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. ICASSP*, vol. 1, pp. 651-654, 1998.

[27] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, vol. 9, pp. 328-331, 1984.

[28] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, pp. 1137-1140, 1996.

[29] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of Gaussians," in *Proc. ICSLP*, pp. 1229-1232, 1996.

[30] S. Panchapagesan, "Frequency Warping by Linear Transformation of Standard MFCC," in *Proc. Interspeech*, pp. 397-400, 2006.

[31] W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix variate priors", in *Proc. EuroSpeech*, pp. 1-4, 1999.

[32] W. Chou and X. He, "Maximum a posterior linear regression (MAPLR) variance adaptation for continuous density HMMs", in *Proc. EuroSpeech*, pp. 1513-1516, 2003.

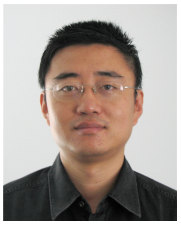
[33] C. Chesta, O. Siohan and C. Lee, "Maximum a posterior linear regression for hidden Markov model adaptation", in *Proc. EuroSpeech*, pp. 211-214, 1999.

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

[35] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithm," in *Proc. ICASSP*, pp. 532-535, 1989.

Shizhen Wang received the B.S. degree from Shandong University, Jinan, China, in 2002, the M.S. degree from Tsinghua University, Beijing, China, in 2005, both in electrical engineering. He is currently working towards the Ph.D. degree in electrical engineering at University of California, Los Angeles (UCLA).

His research interests include speech recognition, speech processing and statistical signal processing.



Xiaodong Cui (M'05) received the B.S. degree (with highest honors) from Shanghai Jiao Tong University, Shanghai, China, in 1996, the M.S. degree from Tsinghua University, Beijing, China, in 1999, and the Ph.D degree from University of California, Los Angeles, in 2005, all in electrical engineering.

From 2005 to 2006, he was a Research Staff Member at DSP Solutions R&D Center, Texas Instruments, Dallas, Texas, focusing on noise robust issues for embedded speech recognition systems. Since 2006, he has been a Research Staff Member at Human Language Technologies, IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests include multilingual speech-to-speech translation, speech recognition (particularly noise robustness and speaker adaptation), digital speech processing, statistical signal processing, machine learning, and pattern recognition.

Abeer Alwan (SM'00) received her Ph.D. in EECS from MIT in 1992. Since then, she has been with the Electrical Engineering Department at UCLA as an Assistant Professor (1992-1996), Associate Professor (1996-2000), Professor (2000-present), Vice Chair of the BME program (1999-2001), and Vice Chair of the EE Graduate Affairs (2003-2006). Prof. Alwan established and directs the Speech Processing and Auditory Perception Laboratory at UCLA (http://www.ee.ucla.edu/ spapl). Her research interests include modeling human speech production and perception mechanisms and applying these models to improve speech-processing applications such as noise-robust automatic speech recognition, compression, and synthesis. She is the recipient of the NSF Research Initiation Award (1993), the NIH FIRST Career Development Award (1994), the UCLA-TRW Excellence in Teaching Award (1994), the NSF Career Development Award (1995), and the Okawa Foundation Award in Telecommunications (1997). Dr. Alwan is an elected member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She served, as an elected member, on the Acoustical Society of America Technical Committee on Speech Communication (1993-1999, and 2005-2007), on the IEEE Signal Processing Technical Committees on Audio and Electroacoustics (1996-2000) and on Speech Processing (1996-2001, 2005-present). She is a member of the Editorial Board of Speech Communication and was an editor-in-chief of that journal (2000-2003), and is an Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing. Dr. Alwan is a Fellow of the Acoustical Society of America and a 2006-2007 Fellow of the Radcliffe Institute for Advanced Study at Harvard University.