Rapid Speaker Adaptation using Regression-Tree based Spectral Peak Alignment

Shizhen Wang[1], *Xiaodong Cui*[2] *and Abeer Alwan*[1]

[1]Department of Electrical Engineering, UCLA, CA, 90095
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
szwang@icsl.ucla.edu, cuix@us.ibm.com and alwan@ee.ucla.edu

Abstract

In this paper, regression-tree based spectral peak alignment is proposed for rapid speaker adaptation using the linearization of VTLN. Two different regression classes are investigated: phonetic classes (using combined knowledge and data-driven techniques) and mixture classes. Compared to MLLR and VTLN, improved performance can be obtained for both supervised and unsupervised adaptations on both medium vocabulary and connected digits recognition tasks. To further improve the performance, MLLR was integrated into this regression-tree based peak alignment. Experimental results show that the performance improvements can be achieved even with limited adaptation data.

Index Terms: speaker adaptation, peak alignment, regression tree

1. Introduction

Spectral mismatch caused by inter-speaker variation of vocal tract length is a major cause of performance degradation in automatic speech recognition. To maintain robust recognition accuracy, vocal tract length normalization (VTLN) is usually applied in speaker adaptation. For computational efficiency, several studies have proposed the possibility of directly performing VTLN in the back-end model space. In [1] and [2], the authors show that VTLN is equivalent to linear transformation in the cepstral domain in the continuous frequency space; in [3], [4] and [5], this VTLN linearization is shown to also hold in the discrete domain under certain approximations. This paper focuses on the method proposed by Cui and Alwan in [5], and develops it into a rapid speaker adaptation algorithm using regression-tree based spectral peaks alignment.

Here we briefly review the approach in [5]: Under the approximation that only central peak values are used to represent each triangular Mel-filter, VTLN warping can be implemented as a linear transformation in the cepstral domain, i.e.,

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

where

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

$\tilde{X}^c$ are the warped cepstral coefficients, X^c are the unwarped ones, $\mathbf{C}$ 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. This linearity can be used to perform rapid speaker adaptations in an MLLR-like manner [6]:

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \tag{3}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{L}\mathbf{H}\mathbf{L}^T \tag{4}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the transformed mean vector and variance matrix, $\mathbf{L}$ is the Cholesky factorization of the original variance $\boldsymbol{\Sigma}$. The mean transform $\mathbf{A}$ is determined based on Eq. 2. The bias vector $\mathbf{b}$ and the covariance transform $\mathbf{H}$ are statistically estimated from the adaptation data under the maximum likelihood criterion:

$$\mathbf{b} = \left\{\sum_{i,k}\sum_{t=1}^{T}\gamma_{ik}(t)\boldsymbol{\Sigma}_{i,k}^{-1}\right\}^{-1}\left\{\sum_{i,k}\sum_{t=1}^{T}\gamma_{ik}(t)\boldsymbol{\Sigma}_{ik}^{-1}(\mathbf{o}(t)-\mathbf{A}\boldsymbol{\mu}_{ik})\right\} \tag{5}$$

$$\mathbf{H} = \frac{\sum_{i,k}\left\{(\mathbf{L}_{ik}^{-1})^T\left[\sum_{t=1}^{T}\gamma_{ik}(t)(\mathbf{o}(t)-\hat{\boldsymbol{\mu}}_{ik})(\mathbf{o}(t)-\hat{\boldsymbol{\mu}}_{ik})^T\right](\mathbf{L}_{ik}^{-1})\right\}}{\sum_{i,k}\sum_{t=1}^{T}\gamma_{ik}(t)} \tag{6}$$

where T is the number of frames of the adaptation data, and i and k are the indices of state and mixture sets, respectively. $\gamma_{ik}(t)$ is the posterior probability of being at state i mixture k at time t given the observation $\mathbf{o}(t)$.

2. Speaker adaptation algorithm

2.1. Regression-tree based peak alignment

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

$$w_{ij} = \begin{cases} 1, & \text{if } i\text{=round}(g_\alpha(j)) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $g_\alpha(\cdot)$ is a linear or piece-wise linear warping function to align the formant-like peaks [7] in the spectrum space. It was shown in [5] that aligning only the third formant (F_3) offers the best performance, that is

$$g_\alpha(j) = \alpha \cdot j \tag{8}$$

$$\alpha = \frac{F_{3,\text{new speaker}}}{F_{3,\text{standard speaker}}} \tag{9}$$

The 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.

Several different approaches can be applied to perform this formant-like peak alignment. In [5], 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 scaling factor α for all phonemes. 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. It can not, however, take advantage of increasing adaptation data.

Another approach is the regression-tree based peak alignment, i.e. to align model parameters within the same class in a similar way. This extension from global to regression-tree based peak alignment is very similar to the expansion of MLLR from a global transform to many transforms especially when adaptation data increase.

[1]Dr. Xiaodong Cui is now working with IBM T. J. Watson Research Center. The work was initiated during his doctoral studies at UCLA.

In this paper, two methods were considered to define regression classes: phoneme-based and mixture-based. In the first method, phonemic units are classified based on phonetic knowledge and/or data-driven methods. For example, phonemes can first be categorized into vowels and consonants, and then consonants can be further classified as voiced or unvoiced, and vowels can be clustered according to their F_3 values. 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. 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.

In the second method, the mixture-based classification approach, Gaussian mixture components are clustered into classes based on a measure of likelihood. In each class, F_3 is estimated and averaged, and peaks are aligned with the same warping factor.

In the following sections, we will evaluate and compare the performances of these different approaches of peak alignment adaptation (PAA).

2.2. Integration of peak alignment with MLLR

As we will show in the next section, when adaptation data are very limited, both approaches of PAA, phoneme-class and mixture-class based, work very well. With few parameters to be estimated, PAA can handle one of the limitations of MLLR: the unreliable parameter estimation for limited data. The performance of PAA, however, tends to saturate when more adaptation data become available. 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 try 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 the speaker specific models are re-estimated using the EM algorithm [9]. The auxiliary function is defined as

$$Q_{\mathcal{N}}(\lambda, \bar{\lambda}) = \sum_{i,k} \sum_{t=1}^{T} \gamma_{ik}(t) \log \mathcal{N}(\mathbf{o}(t); \mathbf{A}\bar{\boldsymbol{\mu}}_{ik} + \mathbf{b}; \bar{\boldsymbol{\Sigma}}_{ik}) \quad (10)$$

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

$$\frac{\partial Q_{\mathcal{N}}(\lambda, \bar{\lambda})}{\partial \bar{\boldsymbol{\mu}}_{ik}} = 0 \quad (11)$$

$$\frac{\partial Q_{\mathcal{N}}(\lambda, \bar{\lambda})}{\partial \bar{\boldsymbol{\Sigma}}_{ik}} = 0 \quad (12)$$

respectively, which give

$$\bar{\boldsymbol{\mu}}_{ik} = \left\{ \sum_{t=1}^{T} \gamma_{ik}(t) \mathbf{A}^T \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{A} \right\}^{-1} \left\{ \sum_{t=1}^{T} \gamma_{ik}(t) \mathbf{A}^T \boldsymbol{\Sigma}_{ik}^{-1} (\mathbf{o}(t) - \mathbf{b}) \right\} \quad (13)$$

$$\bar{\boldsymbol{\Sigma}}_{ik} = \frac{\sum_{t=1}^{T} \gamma_{ik}(t) (\mathbf{o}(t) - \tilde{\boldsymbol{\mu}}_{ik})(\mathbf{o}(t) - \tilde{\boldsymbol{\mu}}_{ik})^T}{\sum_{t=1}^{T} \gamma_{ik}(t)} \quad (14)$$

where

$$\tilde{\boldsymbol{\mu}}_{ik} = \mathbf{A}\bar{\boldsymbol{\mu}}_{ik} + \mathbf{b} \quad (15)$$

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

The integration with MLLR, denoted as PSAT in the following experiments, can be applied to global or regression-tree based peak alignment.

3. Experiments

3.1. Experimental set-up

Two different recognition tasks were carried out to evaluate the performance of the proposed adaptation algorithm. One task is tested on the DARPA Resource Management RM1 database, and the other is on the connected digits TIDIGITS database. For the two databases, speech signals were firstly downsampled to 8kHz, and then segmented into 25ms frames, with 15ms overlap. Each frame was parameterized by a 39-dimensional feature vector consisting of 12 static MFCCs plus log energy, and their first-order and second-order derivatives.

On the RM1 database, triphone acoustic models were trained on the speaker independent (SI) portion of the database (72 speakers, 2880 utterances).This set of SI models produced a baseline performance of 83.8% word recognition rate on the SI test set (8 speakers, 320 utterances). The 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 (SD) portion of the database.

For the TIDIGITS task, monophone acoustic models were trained on 55 adult male speakers and then tested on 10 children (5 boys and 5 girls). The baseline performance on TIDIGITS was a 38.9% word recognition rate. 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. Depending on the amount of the adaptation data available, different numbers of regression classes were experimentally tested, and the best performances were selected for comparison. Formant-like peaks were estimated using Gaussian mixture models [7]. In the 4 kHz 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 is implemented based on a grid search over [0.7, 1.2] with a stepsize of 0.05. The scaling factor producing maximal average likelihood was used to warp the frequency axis [10].

3.2. Comparison of global and regression-classes based PAA

First, experiments were conducted to compare the performance of global (GPAA), phoneme-class (PPAA) and mixture-class (MPAA) based PAA with different numbers of adaptation utterances (or digits). The block-diagonal MLLR adaptation with the optimal number of transforms was also done for comparison. Figures 1 and 2 illustrate the performance of PAA, VTLN and MLLR on the RM1 database and the TIDIGITS database, respectively.

From Figure 1, we can see that PAA can greatly improve the performance over the baseline and VTLN in all cases even with only one adaptation utterance; MLLR, however, may produce worse performance than the baseline when only a small amount of adaptation data is available. Compared to MLLR, PAA performs significantly better for limited adaptation data, with 17.0% reduction of word error rate (WER) over MLLR for 1 or 4 adaptation utterances. With increasing adaptation data, MLLR offers better results than GPAA when the adaptation data is more than 15 utterances, while MPAA can outperform MLLR for 1-25 adaptation utterances.

Among the three PAA methods, MPAA performs the best, and great improvements can be achieved when using regression-tree based over global PAA. As to the two kinds of regression-tree

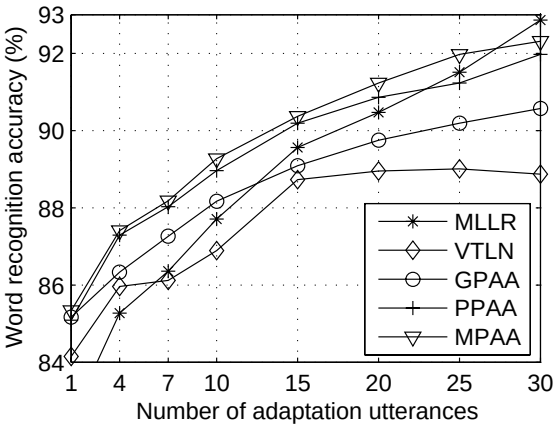

Figure 1: *Performances of VTLN, MLLR and PAA on RM1 (baseline accuracy: 83.8%)*

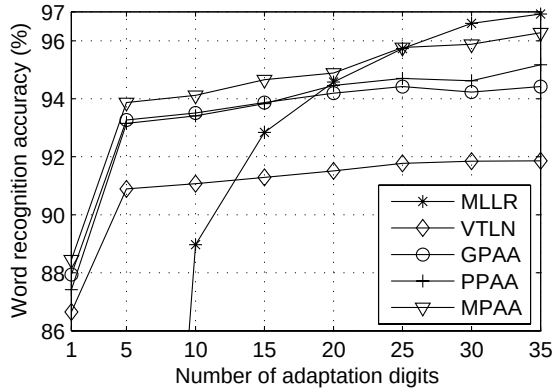

Figure 2: *Performances of VTLN, MLLR and PAA on TIDIGITS (baseline accuracy: 38.9%)*

based PAA, it can be found that MPAA performs slightly better than the PPAA in all cases. In the discussion below, MPAA will be taken as representatives for regression-tree based PAA.

Experimental results demonstrate similar trends on TIDIGITS as shown in Figure 2. This similarity shows that the performance improvements achieved by PAA are consistent across different tasks. Comparing the results in Figures 1 and 2, we can find, however, the performance improvements on TIDIGITS to be more significant than that on RM1 database: for only one adaptation digit (or utterance), more than 81.2% WER reduction over baseline was achieved on TIDIGITS, while on RM1 the WER reduction was about 10.5% . This difference can be explained as follows. The basic idea for PAA is to reduce spectral mismatch by aligning formant-like peaks using estimated F_3. The performance improvement will be more obvious if the F_3 difference between the new speaker and the standard speaker is great, which is the case on TIDIGITS: for adult males the typical F_3 is about 2500Hz, and for children it's 3100 Hz. On the other hand, if the F_3 of the new speaker is very close to that of the standard speaker as on RM1 database, the effect of peak alignment will be small. A limiting case is when the new speaker has exactly the same F_3 as the standard speaker. In this condition peak alignment will have no effect on spectral mismatch, resulting in marginal performance improvement.

3.3. Comparison of PAA and PSAT

The performances of PAA and PSAT are compared in Figures 3 and 4 for RM1 and TIDIGITS databases, respectively. Here we take mixture-classes based peak alignment as the reference which gives the best performance among the three PAA methods. Compared to MPAA, PSAT shows similar performance for a small amount of adaptation data, but better results as the adaptation data increase. The achievable improvements seem to be slight with about 0.5% absolute improvement. The trends of the improvements, however, are obvious and consistent in all cases especially with more adaptation data.

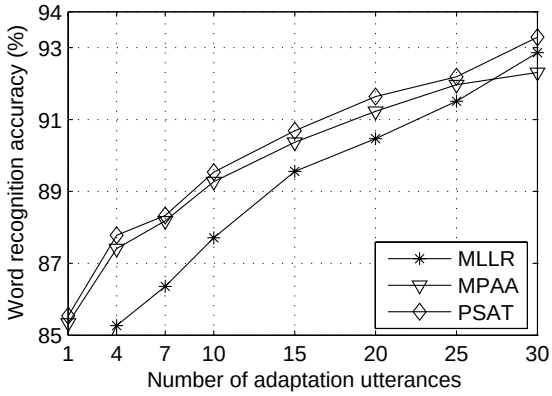

Figure 3: *Performances of PAA, MLLR and PSAT on RM1*

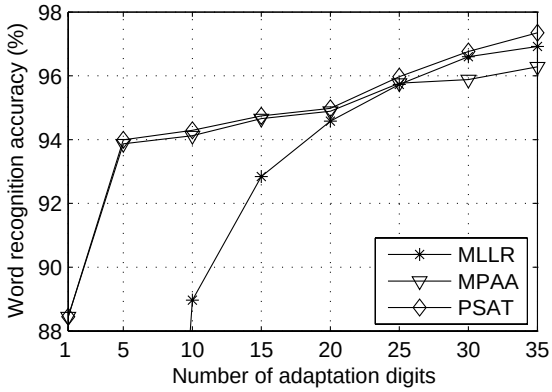

Figure 4: *Performances of PAA, MLLR and PSAT on TIDIGITS*

Compared to MLLR, the performance of PSAT is superior in all experiments, 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 over MLLR is statistically significant. This indicates that PSAT can take advantage of PAA for accurate 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 work even when there is no difference of F_3 between the new speaker and the standard speaker, in which case PSAT can become equivalent to MLLR.

3.4. Comparison of supervised and unsupervised adaptation

The previous adaptation experiments are implemented in a supervised way with the true transcription being 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 variance) is conducted to reduce spectral mismatch. For each test speaker,

formant-like peaks are estimated from the voiced segments of the adaptation utterance, which are detected using the traditional cepstrum peak analysis technique [8]. Peaks are then aligned with the average F_3, i.e. means are adapted according to the following equation:

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} \tag{16}$$

The performances of supervised and unsupervised adaptation are shown in Tables 1, 2 and 3, 4 for RM1 database and TIDIGITS, respectively. It should be noted that the performances listed here for supervised and unsupervised adaptation were 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.

	Number of adaptation utterances							
	1	4	7	10	15	20	25	30
MLLR	82.6	85.3	86.4	87.7	89.6	90.5	91.5	92.9
GPAA	85.2	86.3	87.3	88.2	89.1	89.8	90.2	90.6
MPAA	85.3	87.4	88.2	89.3	90.4	91.2	92.0	92.3
PSAT	85.5	87.8	88.3	89.5	90.7	91.6	92.2	93.3

Table 1: *Word recognition accuracy on RM1 (supervised)*

	Number of adaptation utterances							
	1	4	7	10	15	20	25	30
MLLR	80.9	84.0	85.1	86.3	87.9	89.9	90.6	91.6
GPAA	85.0	86.2	87.1	88.0	88.9	89.6	89.9	90.4
MPAA	83.7	85.8	86.7	88.7	89.6	90.2	91.2	91.6
PSAT	84.0	86.4	87.3	89.1	90.4	91.0	91.5	92.5

Table 2: *Word recognition accuracy on RM1 (unsupervised)*

	Number of adaptation digits							
	1	5	10	15	20	25	30	35
MLLR	40.5	57.0	88.9	92.8	94.6	95.7	96.6	96.9
GPAA	87.9	93.3	93.5	93.9	94.2	94.4	94.2	94.4
MPAA	88.5	93.9	94.1	94.7	94.9	95.8	95.9	96.3
PSAT	88.5	94.0	94.3	94.7	95.0	96.0	96.8	97.4

Table 3: *Word recognition accuracy on TIDIGITS (supervised)*

From the tables, compared to supervised adaptation, unsupervised adaptation performs a little worse in all experimental cases, but the difference is not large: 0.4% and 0.8% absolute WER increase for PSAT on RM1 with 10 and 30 adaptation utterances, respectively; 0.2% and 0.7% absolute WER increase for PSAT on 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 partly 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: the differences being less than 0.2% in all cases.

4. Summary and Conclusion

A rapid speaker adaptation method was investigated in an MLLR-like manner with the transformation for means being generated deterministically by aligning formant-like peaks. The performance of this peak alignment approach was evaluated on both medium vocabulary and connected digits recognition tasks. In both tasks,

	Number of adaptation digits							
	1	5	10	15	20	25	30	35
MLLR	38.9	55.3	88.2	92.3	94.5	95.1	95.9	96.1
GPAA	87.7	93.2	93.4	93.8	94.1	94.3	94.2	94.4
MPAA	86.4	92.3	94.0	94.1	94.5	95.3	95.1	95.2
PSAT	86.4	92.3	94.1	94.2	94.7	95.6	96.2	96.7

Table 4: *Word recognition accuracy on TIDIGITS (unsupervised)*

experimental results show that through peak alignment adaptation significant performance improvements can be achieved even for very limited adaptation data, with mixture-classes based peak alignment performing the best. When sufficient adaptation data are available, peak alignment adaptation offers results similar to or a little worse than MLLR. The PSAT method which integrates peak alignment with MLLR, however, shows better performance than MLLR 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 performance degradation is observed compared to supervised adaptation.

5. Acknowledgements

This materials is based upon work supported 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.

6. References

[1] M. Pitz and H. Ney, "Vocal Tract Normalization as Linear Transformation of MFCC", Proc. of European Conf. On Speech Communication and Technology,1445–1448, 2003

[2] 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. On Speech and Audio Proc., 11(6):549–557, 1998

[3] S. Umesh, A. Zolnay, and H. Ney, "Implementing Frequency-Warping and VTLN Through Linear Transformation of Conventional MFCC", Proc. Of Interspeech 2005,269–272, 2005

[4] Xiaodong Cui and Abeer Alwan, "MLLR-like Speaker Adaptation Based on Linearization of VTLN with MFCC Features", Interspeech 2005, 273–276, 2005

[5] Xiaodong Cui and Abeer Alwan, "Adaptation of children's speech with limited data based on formant-like peak alignment", Computer Speech & Language, in Press

[6] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", Computer Speech & Language, 12(2):75-98, 1998

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

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

[9] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, 39(1):1-38, 1977

[10] L. Lee and R. Rose, "A frequency warping approach to speaker normalization", IEEE Trans. on Speech and Audio Processing, 6(1):49C60, 1998