Speaker Adaptation Based on Nonlinear Spectral Transform for Speech Recognition

Toyohiro Hayashi, Yoshihiko Nankaku, Akinobu Lee and Keiichi Tokuda

Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan

Abstract

This paper proposes a speaker adaptation technique using a nonlinear spectral transform based on GMMs. One of the most popular forms of speaker adaptation is based on linear transforms, e.g., MLLR. Although MLLR uses multiple transforms according to regression classes, only a single linear transform is applied to each state. The proposed method performs nonlinear speaker adaptation based on a new likelihood function combining HMMs for recognition with GMMs for spectral transform. Moreover, the dependency of transforms on context can also be estimated in an integrated ML fashion. The proposed technique outperformed conventional approaches in phoneme-recognition experiments.

Index Terms: Speech Recognition, Speaker Adaptation, Nonlinear Spectral Transformation

1. Introduction

A speaker adaptation technique is a powerful approach to handling mismatches between training and testing speeches. One of typical, Maximum Likelihood Linear Regression (MLLR) [1] is generally used. In this method, regression matrices are learned using a small amount of adaptation data and the linear transforms of the model parameters are estimated. Although the regression matrices are given each regression classes which classified the states of HMMs, this method can only represent a single linear transform for each state.

On the other hand, spectral transform techniques based on GMMs [2][3] in voice conversion can represent the nonlinear transforms of spectral features. This is because the transform matrix changes being dependent on input form. In the proposed technique, we considered applying a spectral transform based on GMMs to the speaker adaptation framework. Speaker adaptation can be carried out by applying a spectral transform to the observation sequences in the stage before speech recognition. However, this method could not improve the accuracy of recognition in preliminary experiments. This is because it did not take account of HMM parameters for recognition at the spectral transform. Furthermore, these spectral transform techniques cannot represent the dependency of transforms on context. MLLR makes use of a regression class that cluster the Gaussians in HMM states, and the same transform is applied to each cluster. This means that MLLR can represent the dependency of transforms on context, because transforms change being dependent on the HMM states corresponding to the context information. Therefore, ignoring the dependency of transform on context can be a disadvantage in spectral-transform-based adaptation against MLLR.

To address these issues, we propose a speaker adaptation technique based on a newly defined likelihood function combining HMMs for recognition with GMMs for spectral transform. Nonlinear speaker adaptation is performed by updating the parameters corresponding to GMM in the ML fashion. Furthermore, by introducing state dependency into the mixture weights of GMM, the dependency of transforms on context can also be represented as well as the regression class in MLLR. Moreover, the proposed method can perform not only hard classification but soft clustering of the regression class by estimating the weight parameters based on a consistent ML criterion.

This paper is organized as follows. Section 2 explains speaker adaptation based on MLLR. Speaker adaptation based on nonlinear spectral transform is presented in Section 3. and the experimental results are reported in Section 4. Finally, conclusions are drawn and future work is discussed in Section 5.

2. Speaker Adaptation Based on MLLR

MLLR computes a set of transformations that will reduce the mismatch between an initial model set and the adaptation data. More specifically, MLLR is a model adaptation technique that estimates a set of linear transformations for the mean and variance parameters of HMM systems. It generally uses regression classes that classified the states of HMMs to give different linear transforms for each states.

The transform matrices used to give a new estimate of the adapted mean and variance in the constrained MLLR (CMLLR) [4] are given by

$$\bar{\boldsymbol{\mu}}_{q_t} = \boldsymbol{H}^{(r)}\boldsymbol{\mu}_{q_t} + \tilde{\boldsymbol{b}}^{(r)} \tag{1}$$

$$\bar{\boldsymbol{\Sigma}}_{q_t} = \boldsymbol{H}^{(r)}\boldsymbol{\Sigma}_{q_t}\boldsymbol{H}^{(r)\top} \tag{2}$$

and the likelihood function is written as follows:

$$P(\boldsymbol{o}\,|\,\lambda) = \sum_{\boldsymbol{q}}\prod_{t} a_{q_{t-1}q_t}\mathcal{N}(\boldsymbol{o}_t|\bar{\boldsymbol{\mu}}_{q_t},\bar{\boldsymbol{\Sigma}}_{q_t}) \tag{3}$$

where $\boldsymbol{o} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ is a feature vector of adaptation data, $\boldsymbol{\mu}_{q_t}$ and $\boldsymbol{\Sigma}_{q_t}$ correspond to the mean vector and covariance matrix at state q_t of the model that learned for recognition, and $a_{q_{t-1}q_t}$ is the state transition probability. The state $\boldsymbol{q}$ is assumed to belong to regression class r in this expression.

Furthermore, this transform can be written as a feature space transformation as follows:

$$\hat{\boldsymbol{o}}_t^{(r)} = \boldsymbol{A}^{(r)}\boldsymbol{o}_t + \boldsymbol{b}^{(r)} = \boldsymbol{W}^{(r)}\boldsymbol{\zeta}(t) \tag{4}$$

where $\boldsymbol{\zeta}$ is an extended feature vector, and the relation between model space transformation and feature space transformation as follows:

$$\mathcal{N}(\boldsymbol{o}_t|\bar{\boldsymbol{\mu}}_{q_t},\bar{\boldsymbol{\Sigma}}_{q_t}) = |\boldsymbol{A}^{(r)}|\mathcal{N}(\hat{\boldsymbol{o}}_t^{(r)}|\boldsymbol{\mu}_{q_t},\boldsymbol{\Sigma}_{q_t}) \tag{5}$$

where the transform matrices $\boldsymbol{W}^{(r)}$ are estimated via EM algorithm.

3. Speaker Adaptation Based on Nonlinear Spectral Transform

3.1. Speaker Adaptation Using Spectral Transform Based on GMM

In the spectral transform based on GMMs, to convert spectral feature sequences of a source speaker to that of a target speaker, the joint probability of two features are modeled by GMMs. Let a vector $\boldsymbol{o}'_t = (\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)})$ be a joint feature vector of that of the source $\boldsymbol{o}_t^{(1)}$ and that of the target $\boldsymbol{o}_t^{(2)}$ at time t. Alignment between two feature sequences is obtained with Dynamic Programming (DP) matching. where the vector sequence $\boldsymbol{o}' = (\boldsymbol{o}'_1, \boldsymbol{o}'_2, \ldots, \boldsymbol{o}'_T)$ is modeled with GMM to learn the relation between source and target features. The output probability of $\boldsymbol{o}'$ given GMM $\lambda^{(G)}$ for a spectral transform can be written as follows:

$$\begin{aligned} & P(\boldsymbol{o}' \,|\, \lambda^{(G)}) \\ = \; & \sum_{\boldsymbol{m}} P(\boldsymbol{o}', \boldsymbol{m} \,|\, \lambda^{(G)}) \\ = \; & \prod_{t=1}^{T} \sum_{m_t=1}^{M} P(m_t \,|\, \lambda^{(G)}) P(\boldsymbol{o}'_t \,|\, m_t, \lambda^{(G)}) \\ = \; & \prod_{t=1}^{T} \sum_{m_t=1}^{M} w_{m_t} \mathcal{N}\left(\boldsymbol{o}'_t \,|\, \boldsymbol{\mu}'_{m_t}, \boldsymbol{\Sigma}'_{m_t}\right) \qquad (6) \end{aligned}$$

where

$$\boldsymbol{\mu}'_{m_t} = \begin{bmatrix} \boldsymbol{\mu}_{m_t}^{(1)} \\ \boldsymbol{\mu}_{m_t}^{(2)} \end{bmatrix}, \; \boldsymbol{\Sigma}'_{m_t} = \begin{bmatrix} \boldsymbol{\Sigma}_{m_t}^{(1,\,1)} & \boldsymbol{\Sigma}_{m_t}^{(1,\,2)} \\ \boldsymbol{\Sigma}_{m_t}^{(2,\,1)} & \boldsymbol{\Sigma}_{m_t}^{(2,\,2)} \end{bmatrix} \qquad (7)$$

and m_t is the mixture index at time t and $\boldsymbol{\mu}_{m_t}^{(1)}$ and $\boldsymbol{\mu}_{m_t}^{(2)}$ are the mean vectors at mixture component m_t of the source and target speaker, respectively. Here, $\boldsymbol{\Sigma}_{m_t}^{(1,\,1)}$ and $\boldsymbol{\Sigma}_{m_t}^{(2,\,2)}$ correspond to the covariance matrices at mixture component m_t of the source and target speaker and $\boldsymbol{\Sigma}_{m_t}^{(1,\,2)}$ and $\boldsymbol{\Sigma}_{m_t}^{(2,\,1)}$ correspond to the cross covariance matrices. These model parameters can be estimated via the EM algorithm. In the spectral conversion based on GMMs, the optimal converted feature sequence $\boldsymbol{o}^{(2)} = (\boldsymbol{o}_1^{(2)}, \boldsymbol{o}_2^{(2)}, \ldots, \boldsymbol{o}_T^{(2)})$ given a source feature sequence $\boldsymbol{o}^{(1)} = (\boldsymbol{o}_1^{(1)}, \boldsymbol{o}_2^{(1)}, \ldots, \boldsymbol{o}_T^{(1)})$ is obtained by maximizing the following conditional distribution:

$$\begin{aligned} & P(\boldsymbol{o}^{(2)} \,|\, \boldsymbol{o}^{(1)}, \lambda^{(G)}) \; = \\ & \sum_{\boldsymbol{m}} \prod_{t=1}^{T} \left[P(m_t \,|\, \boldsymbol{o}_t^{(1)}, \lambda^{(G)}) P(\boldsymbol{o}_t^{(2)} \,|\, \boldsymbol{o}_t^{(1)}, m_t, \lambda^{(G)}) \right] \; (8) \end{aligned}$$

This is because the mixture component for the transform changes depending on the input data by using the posterior probability given the observation sequence; this technique can represent the nonlinear transform.

Recognition accuracy is expected to be improved by applying the spectral transform prior to recognition. However, training GMMs for the spectral transform requires joint feature vectors. Moreover, it is impossible to prepare joint feature vectors from observed data in speaker independent models. The parameter-generation algorithm [5] can be employed to overcome this problem. Figure 1 summarizes this system. We conducted a preliminary experiment to evaluate this system. However, the accuracy of recognition could not be improved. This is because this method did not take account of HMM model parameters for recognition in the spectral transform.

3.2. Speaker Adaptation Based on Combining Acoustic Models

In this paper, we propose a speaker adaptation technique based on a new likelihood function combining HMMs for recognition with GMMs for spectral transform. The likelihood function

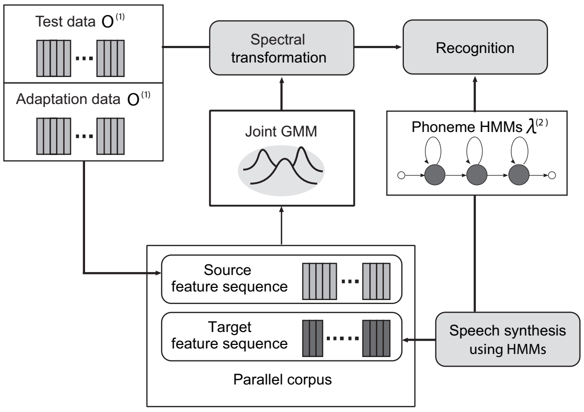

Figure 1: *speaker adaptation based on spectral transform*

combining HMMs with GMMs is defined as follows:

$$\begin{aligned} & P(\boldsymbol{o}^{(1)} \,|\, \lambda) \\ & = \int \prod_{t} P(\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)} | \lambda) d\boldsymbol{o}^{(2)} \\ & = \int \prod_{t} P(\boldsymbol{o}_t^{(1)} | \boldsymbol{o}_t^{(2)}, \lambda^{(G)}) P(\boldsymbol{o}_t^{(2)} | \lambda^{(H)}) d\boldsymbol{o}^{(2)} \\ & = \int \sum_{\boldsymbol{m}} \prod_{t} [P(m_t | \boldsymbol{o}_t^{(2)}, \lambda^{(G)}) P(\boldsymbol{o}_t^{(1)} | \boldsymbol{o}_t^{(2)}, m_t, \lambda^{(G)})] \\ & \times \sum_{\boldsymbol{q}} \prod_{t} [P(q_t | q_{t-1}, \lambda^{(H)}) P(\boldsymbol{o}_t^{(2)} | q_t, \lambda^{(H)})] d\boldsymbol{o}^{(2)} \qquad (9) \end{aligned}$$

where $\boldsymbol{o}_t^{(1)}$ and $\boldsymbol{o}_t^{(2)}$ correspond to the feature sequences before and after the transform and $P(\boldsymbol{o}_t^{(1)} | \boldsymbol{o}_t^{(2)}, \lambda^{(G)})$ is the posterior distribution of GMMs. Here, $P(\boldsymbol{o}_t^{(2)} | \lambda^{(H)})$ is the likelihood function of HMMs for recognition, $\boldsymbol{m} = (m_1, \ldots, m_T)$ is the mixture index sequence of GMMs, and $\boldsymbol{q} = (q_1, \ldots, q_T)$ is the state index sequence of HMMs. In the proposed method, parameters $\lambda^{(G)}$ that maximize Equation (9) are estimated.

In equation (9), the transformed feature sequence $\boldsymbol{o}^{(2)}$ is marginalized out, therefore the likelihood function is defined as a function only of the input feature sequence $\boldsymbol{o}^{(1)}$. Hence, the parameters of the proposed model can be estimated from $\boldsymbol{o}^{(2)}$ without using joint feature sequences. However, to directly optimize the proposed model involves large computational cost because a closed form solution cannot be derived due to the normalization term of the posterior distribution $P(m_t | \boldsymbol{o}_t^{(2)}, \lambda^{(G)})$. To avoid this problem, we propose the following approximation.

$$P(m_t | \boldsymbol{o}_t^{(2)}, \lambda^{(G)}) \quad \approx \quad P(m_t | q_t, \lambda^{(G)}) \qquad (10)$$

Using this approximation, likelihood-function Equation (9) is written as follows:

$$\begin{aligned} & P(\boldsymbol{o}^{(1)} \,|\, \lambda) \\ & = \int \sum_{\boldsymbol{m}} \prod_{t} [P(m_t | q_t, \lambda^{(G)}) P(\boldsymbol{o}_t^{(1)} | \boldsymbol{o}_t^{(2)}, m_t, \lambda^{(G)})] \\ & \times \sum_{\boldsymbol{q}} \prod_{t} [P(q_t | q_{t-1}, \lambda^{(H)}) P(\boldsymbol{o}_t^{(2)} | q_t, \lambda^{(H)})] d\boldsymbol{o}^{(2)} \qquad (11) \end{aligned}$$

Although the GMM part (the first two terms) in equation (11) is a linear transform due to the approximation in equation (10), the transform derived from equation (11) is still nonlinear, because the posterior probability of m_t depends on $\boldsymbol{o}_t^{(1)}$. In addition, it

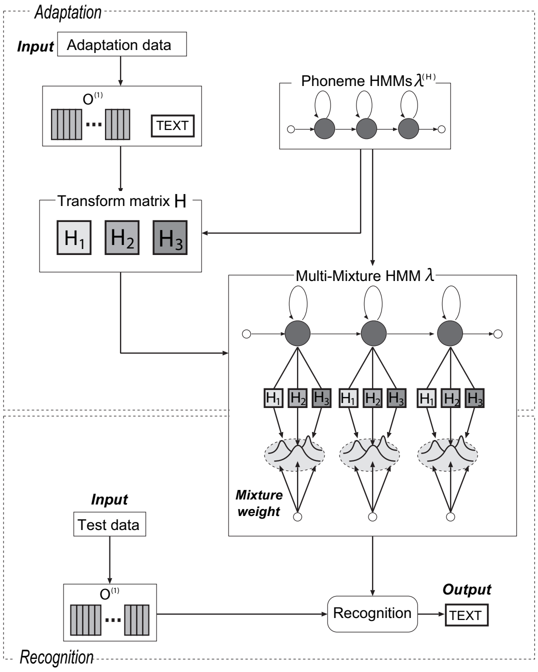

Figure 2: *summary of proposed method system*

can represent the dependence of transform parameters on state as a regression class of MLLR by yielding the dependence on mixture weight of state q_t. Here, the model parameters for each component of the likelihood function are defined as follows:

$$P\left(m_t|q_t,\lambda^{(G)}\right) = w_{m_tq_t} \tag{12}$$

$$P\left(\boldsymbol{o}_t^{(1)}|\boldsymbol{o}_t^{(2)},m_t,\lambda^{(G)}\right) = \mathcal{N}\left(\boldsymbol{o}_t^{(1)}|\boldsymbol{H}_{m_t}\boldsymbol{o}_t^{(2)}+\boldsymbol{\mu}_{m_t}^{(1)},\boldsymbol{\Sigma}_{m_t}^{(1)}\right) \tag{13}$$

$$P\left(q_t|q_{t-1},\lambda^{(H)}\right) = a_{q_{t-1}q_t} \tag{14}$$

$$P\left(\boldsymbol{o}_t^{(2)}|q_t,\lambda^{(H)}\right) = \mathcal{N}\left(\boldsymbol{o}_t^{(2)}|\boldsymbol{\mu}_{q_t}^{(2)},\boldsymbol{\Sigma}_{q_t}^{(2)}\right) \tag{15}$$

where $w_{m_tq_t}$ is the mixture weight of GMMs and $a_{q_{t-1}q_t}$ is the state transition probability of HMMs. Here, $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ and $\boldsymbol{\Sigma}^{(1)}$ and $\boldsymbol{\Sigma}^{(2)}$ are the mean vectors and covariance matrices of the adaptation speaker and model speaker, respectively, and $\boldsymbol{H}_{m_t}$ is a transform matrix at the mixture component m_t. Using these parameters, likelihood-function Equation (11) can be rewritten as follows:

$$P(\boldsymbol{o}^{(1)}\mid\lambda) = \sum_{\boldsymbol{m},\boldsymbol{q}}\prod_t a_{q_{t-1}q_t}w_{m_tq_t}\mathcal{N}(\boldsymbol{o}_t^{(1)}|\hat{\boldsymbol{\mu}}_{m_tq_t},\hat{\boldsymbol{\Sigma}}_{m_tq_t}) \tag{16}$$

$$\hat{\boldsymbol{\mu}}_{m_tq_t} = \boldsymbol{\mu}_{m_t}^{(1)} + \boldsymbol{H}_{m_t}\boldsymbol{\mu}_{q_t}^{(2)} \tag{17}$$

$$\hat{\boldsymbol{\Sigma}}_{m_tq_t} = \boldsymbol{\Sigma}_{m_t}^{(1)} + \boldsymbol{H}_{m_t}\boldsymbol{\Sigma}_{q_t}^{(2)}\boldsymbol{H}_{m_t}^{\top} \tag{18}$$

It can be seen from Equations (3) and (18) that the proposed model has a similar form to CMLLR with Multi-mixture HMMs. Figure 2 has a summary of this system. The difference from CMLLR is that the covariance matrix of the proposed model contains the bias term $\boldsymbol{\Sigma}_{m_t}^{(1)}$ and all the mean and covariance of mixture components are shared in each state. In other words, the mixture components of the proposed model are expanded by combining the original Gaussian component of HMM with the mixture components of GMM. Furthermore, applying the approximation in Equation (10), the mixture weights depend on the state index q_t. This means that the proposed model can represent the dependency of the transform matrices on context as well as the regression class of CMLLR. If the weight of the GMM mixture component corresponding to the regression class is set to 1.0 and the others are set to 0.0, the proposed method can represent the same regression class of CMLLR. Therefore, it can be seen that the proposed method includes the conventional CMLLR as the model structure. Moreover, soft clustering of regression class can be performed by estimating the weight parameters based on the ML fashion.

In the adaptation process, the model parameters of GMMs $\lambda^{(G)} = \{w_{m_tq_t},\boldsymbol{\mu}_{m_t}^{(1)},\boldsymbol{\Sigma}_{m_t}^{(1)},\boldsymbol{H}_{m_t}\}$ are estimated via EM algorithm using adaptation data. The update procedure of model parameters in the proposed method is summarized as follows:

1. $\boldsymbol{\mu}_{m_t}^{(1)},\boldsymbol{\Sigma}_{m_t}^{(1)}$ and $\boldsymbol{H}_{m_t}$ that maximize likelihood function $P(\boldsymbol{o}^{(1)}\mid\lambda)$ are estimated via EM algorithm by using adaptation data $\boldsymbol{o}^{(1)} = (\boldsymbol{o}_1^{(1)},\boldsymbol{o}_2^{(1)},\ldots,\boldsymbol{o}_T^{(1)})$.
2. Each mean and covariance matrices of expanded mixture components are updated based on Equations (17) and (18) by using transform matrices $\boldsymbol{H}_{m_t}$.
3. The mixture weight of each state $w_{m_tq_t}$ that maximizes likelihood function $P(\boldsymbol{o}^{(1)}\mid\lambda)$ is estimated via the EM algorithm.

In step 1., assuming $\boldsymbol{\Sigma}_{m_t}^{(1)} = \boldsymbol{0}$, $\{\boldsymbol{\mu}_{m_t}^{(1)},\boldsymbol{H}_{m_t}\}$ can be estimated by using the same procedure of CMLLR.

4. Experiments

4.1. Speaker Dependent Experiments

For training speaker dependent HMM sets, we used an ATR Japanese speech database B-set uttered by six male and four female speakers. Two male speakers were selected; one for training and one for adaptation (training: MHT and adaptation: MTK). Each speaker uttered 503 sentences. We used 400 sentences of uttered by speaker MHT for training, and 50 different sentences by speaker MTK for adaptation, and the remaining 53 sentences by speaker MTK were used for evaluation. The speech data were down-sampled from 20 to 16 kHz and windowed at a 10-ms frame rate using a 25-ms Blackman window. Each spectral feature vector consisted of 18 mel-cepstral coefficients and their delta and delta-delta coefficients.

In this experiment, the following two methods were compared.

- MLLR: speaker adaptation based on MLLR. Only the mean vector transform was used in this experiments. The number of regression classes was 4, 8, and 16.
- NLST: speaker adaptation based on a Non-Linear Spectral Transform. Only the mean vector transform was used in this experiments. The initial parameters for the proposed model are given by MLLR. The number of mixtures was 4, 8, and 16.

When the number of regression classes in MLLR is equal to the number of mixtures in NLST, these methods have the same number of transform matrices in adaptation. Figure 3 plots the phoneme accuracy while iterating the updates of parameters. The phoneme accuracy without adaptation was 60.71%. It can be seen that large improvement was not observed in MLLR adaptation after the second iteration. In contrast, the accuracy of the proposed method gradually improved by updating GMM parameters and weights iteratively. When comparing NLST to MLLR with the same number of transform matrices, NLST outperformed MLLR under all conditions. This is because the proposed method carried out nonlinear spectral transform by using

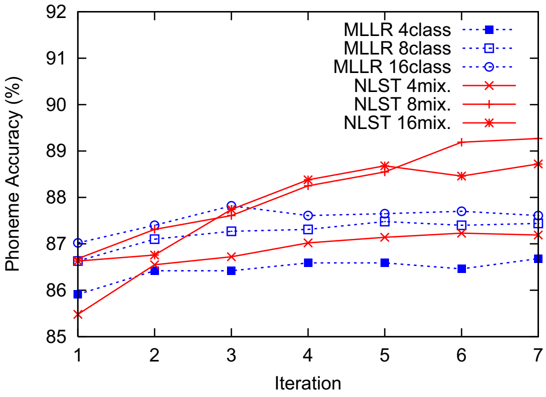

Figure 3: *Results for speaker-dependent phoneme recognition using MLLR and proposed technique with only mean transform*

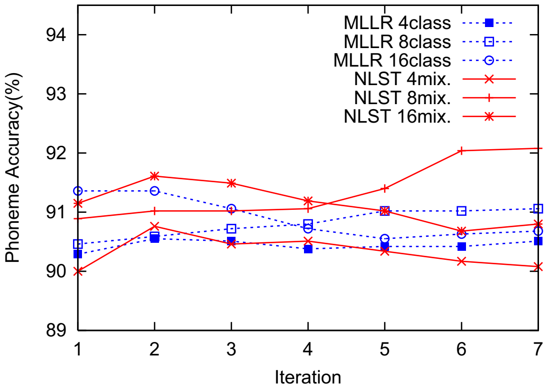

Figure 4: *Results for speaker-independent phoneme recognition using MLLR and proposed technique with only mean transform*

all the transform matrices according to the weights, even though a single transform matrix was applied to each state based on the regression class in MLLR.

4.2. Speaker Independent Experiments

For training speaker independent HMM sets, four male and female speakers were selected from the ATR Japanese speech database and 400 sentences were used for each speaker in the training, and 50 different sentences by speaker MTK were used for adaptation and the remaining 53 sentences by speaker MTK were used for evaluation. Figure 4 plots the results for speaker-independent phoneme-recognition. Phoneme accuracy without adaptation was 84.33%. The results indicated that the proposed method was also effective for the speaker-independent model.

Additionally, the following two methods with mean and variance transforms were also compared in this experiment.

- CMLLR: speaker adaptation based on CMLLR. The number of regression classes was 4, 8, and 16.
- CNLST: speaker adaptation based on a Non-Linear Spectral Transform. We assumed $\mathbf{\Sigma}_{mt}^{(1)} = \mathbf{0}$ and the initial parameters for the proposed model were given by CMLLR in this experiments. The number of mixtures was 4, 8, and 16.

Figure 5 plots the results for phoneme recognition using CMLLR and the proposed method. This figure shows that the proposed method outperforms the conventional CMLLR approach

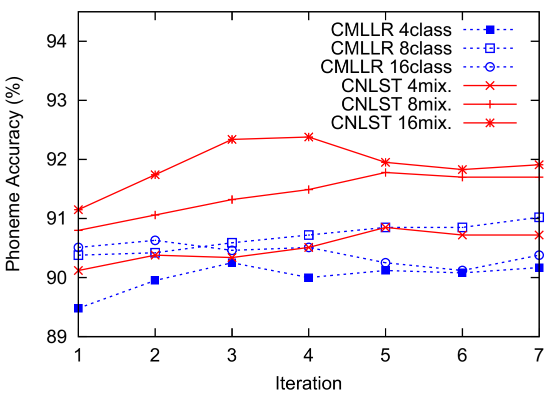

Figure 5: *Results for speaker-independent phoneme recognition using CMLLR and proposed technique with mean and variance transforms*

with mean and variance transforms. Furthermore, the effectiveness of the variance transform in the proposed method can be seen by comparing Figures 4 and 5.

5. Conclusions

This paper proposed a speaker adaptation technique based on a nonlinear spectral transform that was carried out in ML fashion by using weighted multiple transform matrices. The proposed method outperformed conventional MLLR methods in the phoneme-recognition experiments. Estimating the variance bias parameters and clustering of the weight parameters will be future work.

6. Acknowledgements

The research leading to these results was partly funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), of the Ministry of Internal Affairs and Communication, Japan.

7. References

[1] 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, No.2, pp.171–185, 1995.

[2] Y. Stylianou, O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion,"Proc. of IEEE Trans. on Speech and Audio Processing, vol. 6, pp. 131–142, 1998.

[3] T. Toda, A. W. Black and K. Tokuda, "Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory," Proc. of IEEE Trans. on ASLP, vol.15, No.8, pp.2222–2235, Nov.2007.

[4] M. J. F Gales and P. C. Woodland, "Mean and covariance adaptation within MLLR framework," Computer Speech and Language, vol.10, pp.249–264, 1996.

[5] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. of ICASSP, pp.1315–1318, 2000.