TEMPORAL STRUCTURE CONSTRAINED TRANSFORMATION FOR SPEAKER ADAPTATION

Eric H. C. Choi, Trym Holter, Julien Epps, Arun Gopalakrishnan

Motorola Australian Research Centre, Motorola Labs
{Eric.Choi, Trym.Holter, Julien.Epps, Arun.Gopal}@motorola.com

ABSTRACT

In this paper we suggest that rather than modeling speaker mismatch as an affine transform of the entire feature vector, it can be modeled by an affine transform of the static coefficients with additional constraints imposed by the temporal relationships of the streams of coefficients. This results in the different streams sharing the same rotation matrix, and thus reduces the complexity and memory requirements for speaker adaptation, as well as minimizes the adaptation data requirements. We present the solution for the case where temporal structure constrained transforms (TSCT) are optimized using the maximum likelihood criterion. The experiments presented in the paper show that with the proposed approach, the same accuracy after adaptation for the Wall Street Journal (WSJ) task can be achieved by using only 60% of the total number of transformation parameters that it would require if conventional block-diagonal transformation is used. In addition, TSCT provides better recognition accuracy when there is only a very limited amount of adaptation data.

1. INTRODUCTION

It is well known that speaker-independent (SI) automatic speech recognition (ASR) systems, despite the steady improvements generated over recent years, still have error rates that are much larger than corresponding speaker-dependent (SD) systems [1]. However, it takes large amounts of data to properly estimate the model parameters of an ASR system, and it is therefore in most cases not possible to collect sufficient SD data to create such models. More commonly, data from a large pool of different speakers are used to generate SI acoustic models. Even though some speakers experience very good performance with such models, a large variation in accuracy can be expected in the population, depending on how well a particular user's voice characteristics (including both physiological and sociological aspects) are represented in the training data set.

A common solution to this speaker mismatch problem is to employ speaker adaptation techniques. These techniques fall mainly into three categories which include maximum a posteriori (MAP) adaptation [2] [3], transformation based adaptation [4] [5] [6] and speaker clustering [7] [8]. Such methods modify the parameters of the initial acoustic models, using only small amounts of speaker specific data, to generate the speaker-adapted (SA) models. The goal is to approach the accuracy that can be achieved using SD ASR systems, while at the same time minimizing the training load on a user.

One of the most successful approaches to speaker adaptation in the hidden Markov model (HMM) framework is the maximum likelihood linear regression (MLLR) [6]. The most important feature of an HMM is the probability density functions (pdfs) that specify the state output distributions. These are typically Gaussian mixture models. In MLLR, the mean vectors of the Gaussians are grouped into clusters, and each mean vector ($\boldsymbol{\mu}$) in a cluster is adapted using an affine transformation of the form:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\Gamma}\boldsymbol{\mu} + \boldsymbol{\beta} \qquad (1)$$

By carefully selecting the number of clusters to be used, MLLR can not only help to create good SA models when a sufficient amount of adaptation data has been collected, but also yield improvements when only small amounts of data are available. In this case, the clustering approach will help adapting mean vectors for which there are no data available in the adaptation set. However, with small amounts of data, the number of transformation parameters that can be reliably estimated will ultimately limit the extent of the performance improvement. In this paper we present a procedure that exploits the temporal relationship that is typically used in ASR feature vectors. The objective of this approach is to reduce the equivalent dimensionality of the affine transformations. We will show how this can help in reducing the computational complexity, as well as increasing the adaptation rate. The latter is possible because fewer parameters will be required to specify each transformation.

In the next section we will develop the solution for this temporal structure constrained transformation (TSCT) under the maximum likelihood (ML) criterion. In section 3, we will then report some experimental results with this novel method for the Wall Street Journal (WSJ) database, and compare them to a standard MLLR approach. The results are discussed in section 4 while our conclusions are found in section 5.

2. TEMPORAL STRUCTURE CONSTRAINED TRANSFORMATION

The feature vectors typically used in ASR consist of a stream of static coefficients, augmented by their individual 1st order and 2nd order time derivatives. It is often assumed that there is no cross-correlation among these streams of coefficients, and this is commonly exploited to reduce the complexity in transformation-based adaptation. The implication of this assumption is that the

rotation matrix is reduced to a block-diagonal form, and thus a reduced number of adaptation parameters need to be estimated.

Our approach is different, in that rather than modeling speaker mismatch as an affine transform of the entire feature vector, it is modeled by an affine transform of the static coefficients. The additional constraints are then imposed by the temporal relationships of the streams of coefficients. This results in the different streams sharing the same rotation matrix, and therefore we can further reduce the number of non-zero elements in a transformation to one-third of that of a block-diagonal transformation. This reduces the complexity and memory requirements for the adaptation, as well as minimizes the adaptation data requirements.

2.1 Definition of TSCT

For an *n*-dimensional input feature vector $\underline{\mathbf{X}}$, its output vector ($\underline{\mathbf{Y}}$) after transformation is given by:

$$\underline{\mathbf{Y}} = \boldsymbol{\Gamma}\underline{\mathbf{X}} + \boldsymbol{\beta} \tag{2}$$

The transformation is thus defined by the *n x n* rotation matrix $\boldsymbol{\Gamma}$ and the *n* x 1 bias vector $\boldsymbol{\beta}$.

We now assume that the feature vector consists of the static coefficients ($\mathbf{x}$), augmented by their individual 1st order ($\dot{\mathbf{x}}$) and 2nd order ($\ddot{\mathbf{x}}$) time derivatives. We can write this as $\underline{\mathbf{X}} = [\mathbf{x}^T \ \dot{\mathbf{x}}^T \ \ddot{\mathbf{x}}^T]^T$. Assuming that speaker mismatch can be modeled by an affine transform of the static coefficients, these coefficients are transformed to:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{3}$$

where $\mathbf{A}$ is a *n*/3 x *n*/3 rotation matrix and $\mathbf{b}$ is a *n*/3 x 1 bias vector. We can now introduce the constraints imposed by the temporal relationships between the streams of coefficients. It follows from equation (3) that:

$$d\mathbf{y}/dt = \mathbf{A}\,d\mathbf{x}/dt = \mathbf{A}\dot{\mathbf{x}} \tag{4}$$

$$d^2\mathbf{y}/dt^2 = \mathbf{A}\,d^2\mathbf{x}/dt^2 = \mathbf{A}\ddot{\mathbf{x}} \tag{5}$$

The TSCT transformation can thus be written:

$$\underline{\mathbf{Y}} = \begin{bmatrix}\mathbf{y}\\ \dot{\mathbf{y}}\\ \ddot{\mathbf{y}}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\mathbf{x}+\mathbf{b}\\ \mathbf{A}\dot{\mathbf{x}}\\ \mathbf{A}\ddot{\mathbf{x}}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \dot{\mathbf{x}}\\ \ddot{\mathbf{x}}\end{bmatrix} + \begin{bmatrix}\mathbf{b}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} \tag{6}$$

and its rotation matrix is block-diagonal. By contrast, the standard block-diagonal transformation is reached by assuming that speaker mismatch can be modeled by an affine transformation of the entire feature vector and at the same time assuming that the cross-correlation between the streams of the feature vector is zero. For this conventional case, the transformation can be written:

$$\underline{\mathbf{Y}} = \begin{bmatrix}\mathbf{y}\\ \dot{\mathbf{y}}\\ \ddot{\mathbf{y}}\end{bmatrix} = \begin{bmatrix}\mathbf{A_1} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{A_2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{A_3}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \dot{\mathbf{x}}\\ \ddot{\mathbf{x}}\end{bmatrix} + \begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\\ \mathbf{b}_3\end{bmatrix} \tag{7}$$

In this case, each $\mathbf{A}_i$ is of dimension *n*/3 x *n*/3, while each of the $\mathbf{b}_i$ is of dimension *n*/3 x 1.

2.2 Estimation of TSCT

The general criterion function to be maximised in MLLR for estimating a transformation is given by [6]:

$$\begin{aligned} J &= -\frac{1}{2}\sum_{t=1}^{\tau}\sum_{k\in\Omega}[\gamma_t(k)\left(\underline{\mathbf{o}}_t - \boldsymbol{\Gamma}\underline{\boldsymbol{\mu}}_k - \boldsymbol{\beta}\right)^T\underline{\mathbf{R}}_k\left(\underline{\mathbf{o}}_t - \boldsymbol{\Gamma}\underline{\boldsymbol{\mu}}_k - \boldsymbol{\beta}\right)]\\ &= -\frac{1}{2}\sum_{t=1}^{\tau}\sum_{k\in\Omega}[\gamma_t(k)\left(\underline{\mathbf{o}}_t - \underline{\mathbf{W}}\tilde{\underline{\boldsymbol{\mu}}}_k\right)^T\underline{\mathbf{R}}_k\left(\underline{\mathbf{o}}_t - \underline{\mathbf{W}}\tilde{\underline{\boldsymbol{\mu}}}_k\right)] \end{aligned} \tag{8}$$

where τ is the total number of feature vectors in an adaptation data set, Ω is the set of Gaussians within a regression class, $\underline{\mathbf{o}}_t$ is the feature vector at time *t*, $\gamma_t(k)$ is the posterior probability of $\underline{\mathbf{o}}_t$ occupying the *k*-th Gaussian at time *t*, $\underline{\boldsymbol{\mu}}_k$ is the mean vector of the *k*-th Gaussian and $\underline{\mathbf{R}}_k$ is the corresponding diagonal covariance matrix, $\boldsymbol{\Gamma}$ is the rotation matrix of the transform, $\boldsymbol{\beta}$ is the bias vector, $\underline{\mathbf{W}} = [\boldsymbol{\Gamma}\ \boldsymbol{\beta}]$ and $\tilde{\underline{\boldsymbol{\mu}}}^T = [\underline{\boldsymbol{\mu}}^T \quad 1]$.

In order to simplify the subsequent equations, we add a superscript (*i*) to a vector or a matrix to identify the corresponding stream of coefficients that it is referred to. For example, $\mathbf{x}^{(1)}$ refers to the static coefficients, $\mathbf{x}^{(2)}$ refers to delta coefficients (1st order time derivatives) and so on. By imposing the temporal structure constraints, equation (8) can be re-written as:

$$J = -\frac{1}{2}\sum_{t=1}^{\tau}\sum_{k\in\Omega}\gamma_t(k)\left[\sum_{i=1}^{3}\left(\mathbf{o}_t^{(i)} - \mathbf{W}\hat{\boldsymbol{\mu}}_k^{(i)}\right)^T\mathbf{R}_k^{(i)}\left(\mathbf{o}_t^{(i)} - \mathbf{W}\hat{\boldsymbol{\mu}}_k^{(i)}\right)\right] \tag{9}$$

where

$$\hat{\boldsymbol{\mu}}_k^{(1)} = \begin{bmatrix}\boldsymbol{\mu}_k^{(1)}\\ 1\end{bmatrix}, \quad \hat{\boldsymbol{\mu}}_k^{(2)} = \begin{bmatrix}\boldsymbol{\mu}_k^{(2)}\\ 0\end{bmatrix}, \quad \hat{\boldsymbol{\mu}}_k^{(3)} = \begin{bmatrix}\boldsymbol{\mu}_k^{(3)}\\ 0\end{bmatrix}, \quad \mathbf{W} = [\mathbf{A}\ \mathbf{b}].$$

To further simplify the notation, we introduce

$$\bar{\gamma}_k = \sum_{t=1}^{\tau}\gamma_t(k) \ \text{ and } \ \bar{\mathbf{o}}_k^{(i)} = \sum_{t=1}^{\tau}\gamma_t(k)\mathbf{o}_t^{(i)}$$

2.2.1 Full-rank **A** matrix

In this case, the matrix $\mathbf{A}$ is assumed to have full rank of *n*/3 and therefore the TSCT is block-diagonal. By differentiating the above criterion function in equation (9) with respect to the transform matrix $\mathbf{W}$ and equating the resultant matrix equation to 0, we obtain the following equation system:

$$\sum_{k\in\Omega}\sum_{i=1}^{3}\mathbf{R}_k^{(i)}\bar{\mathbf{o}}_k^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T} = \sum_{k\in\Omega}\bar{\gamma}_k\sum_{i=1}^{3}\mathbf{R}_k^{(i)}\mathbf{W}\hat{\boldsymbol{\mu}}_k^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T} \tag{10}$$

To allow more flexibility in this framework, we now introduce a scaling factor for the occupation count corresponding to each stream of a feature vector, i.e., $\gamma_t(k)$ is multiplied by $\lambda^{(i)}$. The equation system can then be written as:

$$\sum_{k\in\Omega}\sum_{i=1}^{3}\lambda^{(i)}\mathbf{R}_k^{(i)}\bar{\mathbf{o}}_k^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T} = \sum_{k\in\Omega}\bar{\gamma}_k\sum_{i=1}^{3}\lambda^{(i)}\mathbf{R}_k^{(i)}\mathbf{W}\hat{\boldsymbol{\mu}}_k^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T} \quad (11)$$

The above equation system can be solved row-by-row and the solution is given by:

$$\sum_{k\in\Omega}\sum_{i=1}^{3}\lambda^{(i)}r_k^{(i)}(j)\bar{o}_k^{(i)}(j)\hat{\boldsymbol{\mu}}_k^{(i)T} = \mathbf{W}(j)[\sum_{k\in\Omega}\bar{\gamma}_k\sum_{i=1}^{3}\lambda^{(i)}r_k^{(i)}(j)\hat{\boldsymbol{\mu}}_k^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T}] \quad (12)$$

where $r_k^{(i)}(j)$ is the j-th diagonal element of $\mathbf{R}_k^{(i)}$, $\bar{o}_k^{(i)}(j)$ is the j-th element of $\bar{\mathbf{o}}_k^{(i)}$ and $\mathbf{W}(j)$ is the j-th row of $\mathbf{W}$.

Note that if we choose $\lambda^{(1)} = \lambda^{(2)} = \lambda^{(3)}$, the solution to this equation system provides the maximum likelihood (ML) solution to the TSCT optimisation problem. If we choose to let these scaling factors have different values, we will deviate from this solution. However, this framework increases the flexibility of the approach by allowing the contribution from the different parameter sets to be weighted differently. It would for instance give us the opportunity to calculate the transform based only on the static coefficients by setting $\lambda^{(1)} = 1$ and $\lambda^{(2)} = \lambda^{(3)} = 0$, thus incorporating the technique suggested in [9] as a special case.

2.2.2 Diagonal **A** matrix

The solution in equation (12) can be further simplified if we assume that the **A** matrix is diagonal. Let $a(j)$ be the j-th diagonal element of **A** and $b(j)$ be the j-th element of the bias vector **b**, similar derivation from equation (9) for diagonal **A** matrix gives:

$$[\sum_{k\in\Omega}\sum_{i=1}^{3}\lambda^{(i)}r_k^{(i)}(j)\bar{o}_k^{(i)}(j)\mu_k^{(i)}(j) \quad \lambda^{(1)}\sum_{k\in\Omega}r_k^{(1)}(j)\bar{o}_k^{(1)}(j)] =$$

$$[a(j)\ b(j)]\begin{bmatrix}\sum_{k\in\Omega}\bar{\gamma}_k\sum_{i=1}^{3}\lambda^{(i)}r_k^{(i)}(j)\mu_k^{(i)}(j)^2 & \lambda^{(1)}\sum_{k\in\Omega}\bar{\gamma}_k r_k^{(1)}(j)\mu_k^{(1)}(j)\\ \lambda^{(1)}\sum_{k\in\Omega}\bar{\gamma}_k r_k^{(1)}(j)\mu_k^{(1)}(j) & \lambda^{(1)}\sum_{k\in\Omega}\bar{\gamma}_k r_k^{(1)}(j)\end{bmatrix} \quad (13)$$

From equation (13), the values of $a(j)$ and $b(j)$ can be calculated accordingly for each element index j.

3. EXPERIMENTS

3.1 Experimental Setup

The Wall Street Journal (WSJ) database [10] was used in the adaptation experiments. This task has a 5000-word vocabulary, and contains 10 non-native speakers of American English. The front-end process extracted 12 mel-frequency cepstral coefficients (MFCC) and a log-energy coefficient for every 10ms of speech with a 25-ms Hamming window. This static stream of coefficients was then appended with its corresponding 1st order and 2nd order time derivative coefficients to form a full size feature vector. The speaker-independent (SI) models were trained from the SI84 subset (84 speakers from WSJ0). Triphone HMMs were used and the final SI model set contained about 10K Gaussians after state tying. The adaptation and test experiments were based on the Spoke 3 evaluation data set from the 1993 ARPA CSR evaluation. The average duration of each utterance is approximately 8 seconds. Recognition was based on a Motorola proprietary speech recognizer with the standard WSJ bigram language model. A regression tree was built to cluster the individual Gaussians in order to better facilitate the use of multiple transformations when the amount of adaptation data increases.

3.2 Results

Supervised adaptation experiments with various amounts of adaptation utterances were performed for the evaluation. Preliminary experiments testing various scaling factor values revealed that the recognition accuracy obtained by using the TSCT was maximized for $\lambda^{(1)} = \lambda^{(2)} = \lambda^{(3)}$, and accordingly settings of $\lambda^{(1)} = \lambda^{(2)} = \lambda^{(3)} = 1$ were employed in the experiments reported here for using TSCT. The average word accuracy after adaptation for up to 10 utterances/speaker for each of the different types of transformations is shown in Figure 1. Also included in the figure is the SI recognition accuracy as a reference line.

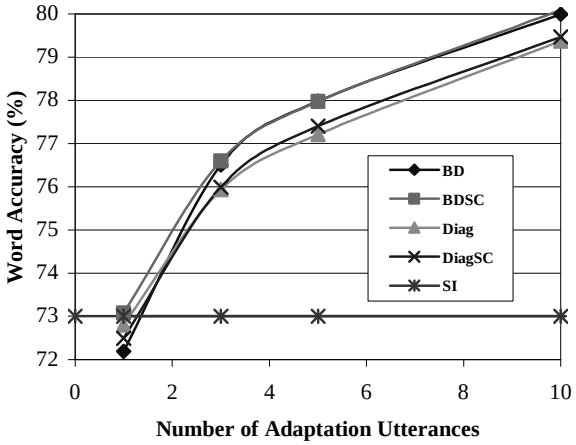

Figure 1:- Average word recognition accuracy (%) across 10 speakers for triphone models with supervised adaptation using block-diagonal (**BD**), TSCT block-diagonal (**BDSC**), diagonal (**Diag**), TSCT diagonal (**DiagSC**) transformations, and the SI result (**SI**)

From the above experimental results, it is evident that the use of TSCT is capable of producing similar accuracy to other transform types, if not better. More than that, it is achieved by using considerably fewer transformation parameters. This is illustrated in Figure 2 which shows that the use of block-diagonal TSCT's requires the least amount of total transformation parameters for a given recognition accuracy. Here the total number of transformation parameters is calculated as the number of transformations employed in an adaptation times the number of free parameters in a transformation. The potential memory savings of using TSCT's are clearly evident. Also it should be noted that the number of adaptation utterances used can be up to 40 in the figure. In the case of using block-diagonal TSCT's, the

average word accuracy is improved to 83.7% when 40 adaptation utterances from each speaker are provided. This improvement is achieved by using only 60% of the total number of transformation parameters that it would require if conventional block-diagonal transformation is used.

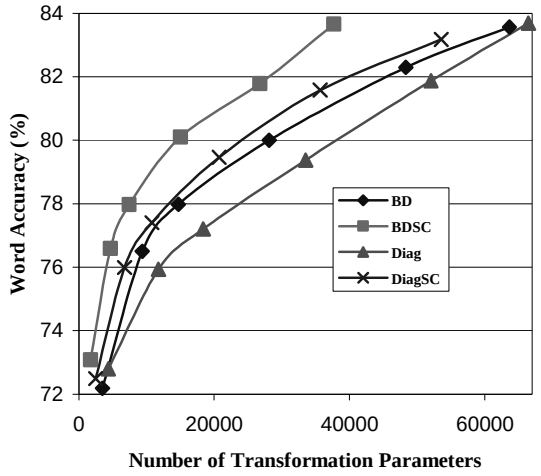

Figure 2:- Average word recognition accuracy (%) after supervised adaptation plotted against the average total number of free parameters for each transform type: block-diagonal (**BD**), TSCT block-diagonal (**BDSC**), diagonal (**Diag**), and TSCT diagonal (**DiagSC**)

4. DISCUSSION

The results from Figure 1 indicate that initial adaptation rate can be improved over the use of conventional block-diagonal transforms by adding the temporal constraints. Only block-diagonal TSCT can achieve a better word accuracy than that of the SI baseline when there is only one adaptation utterance. On the other hand, its accuracy matches that of the conventional block-diagonal transformation when more adaptation data are available. This is significant in view of the considerably smaller number of free parameters to be estimated. As expected, block-diagonal transformation performs a little better than diagonal transformation in terms of word accuracy when there are more adaptation data.

The lines in Figure 2 clearly demonstrate the superiority of TSCT in terms of reducing transformation parameters while providing almost the same or better improvement in recognition accuracy than the conventional approach. The results indicate that the use of temporal relationships among the streams of feature vector can eliminate some of the redundancy ignored in assuming that the streams are independent when using conventional block-diagonal transformation. Comparing the performance of block-diagonal TSCT with that of the diagonal TSCT, we can learn that rotation of within-stream feature space is necessary in order to better model the mismatch in different voice characteristics.

While the derivation of the TSCT was motivated by the desire to exploit the temporal relationships of different streams of feature vector to reduce redundancy in modeling speaker mismatch, it can also be viewed as a specific way of doing parameter tying for a transformation. It remains to see if we can generalize this transformation parameter tying framework to further improve the adaptation performance.

5. CONCLUSION

Rapid adaptation on very limited data is a difficult task for which there are few improvements on conventional approaches. It is during the first few utterances that a user will have their critical initial experience of a particular speech recognition system. Thus, it is claimed that the TSCT approach provides a strong advantage. In particular, we have demonstrated the performance improvements of the TSCT approach on the WSJ task, in terms of both memory storage and accuracy after adaptation. The experimental results verified that temporal structure of feature vectors can be used to reduce the complexity of a transformation and at the same time can enhance the capability of a transformation to capture different voice characteristics.

6. REFERENCES

[1] Zhang Z, Furui S. and Ohtsuki K., "On-line Incremental Speaker Adaptation for Broadcast News Transcription", *Speech Communication*, Issue 37, 2002, pp. 271-281.

[2] Jon E., Kim D. K. and Kim N. S., "EMAP-Based Speaker Adaptation with Robust Correlation Estimation", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, May 2001, pp. 321-324.

[3] Gauvain J. L. and Lee C. H., "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, April 1994, pp. 291-298.

[4] Surendran A. C. and Lee C. H., "Transformation-Based Bayesian Prediction for Adaptation of HMMs", *Speech Communication*, Issue 34, 2001, pp. 159-174.

[5] Gunawardana A. and Byrne W., "Robust Estimation for Rapid Speaker Adaptation Using Discounted Likelihood Techniques", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 2, June 2000, pp. 985-988.

[6] Leggetter C. J. and Woodland P. C., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language*, Vol. 9, No. 2, 1995, pp. 171-185.

[7] Kuhn R., Junqua J. C., Nguyen P. and Niedzielski N., "Rapid Speaker Adaptation in Eigenvoice Space", *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 6, Nov. 2000, pp. 695-707.

[8] Gales M. J., "Cluster Adaptive Training for Speech Recognition", *Proc. Int. Conf. on Spoken Language Processing*, Vol. 5, Dec. 1998, pp. 1783-1786.

[9] Choi H.C., *Spectral Transformation for Speaker Adaptation in HMM Based Speech Recognition*, Ph.D. thesis, University of Sydney, 1996.

[10] http://www.ldc.upenn.edu/