Discriminative Fuzzy Clustering Maximum a Posterior Linear Regression for Speaker Adaptation

Ting-yao Hu[1], *Yu Tsao*[2], *Lin-shan Lee*[1]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
r99942144@ntu.edu.tw, yu.tsao@citi.sinica.edu.tw, lslee@gate.sinica.edu.tw

Abstract

We propose a discriminative fuzzy clustering maximum a posterior linear regression (DFCMAPLR) model adaptation approach to compensate the acoustic mismatch due to speaker variability. The DFCMAPLR approach adopts the MAP criterion and a discriminative objective function to estimate shared affine transform and fuzzy weight sets, respectively. Then, through a linear combination of the calculated fuzzy weights and shared affine transforms, more specific affine transforms are formed for model adaptation. By incorporating the MAP criterion and the discriminative information, DFCMAPLR can calculate shared affine transforms reliably and enhance the discriminative power of the adapted acoustic model. Based on the experimental results on the ASTTEL200 Mandarin corpus, we verified that DFCMAPLR outperforms not only the conventional maximum likelihood linear regression (MLLR) but also the fuzzy clustering MLLR(FCMLLR), which estimates the shared affine transform and fuzzy weight sets both based on the maximum likelihood criterion. Moreover, when compared to the baseline result, DFCMAPLR provides a clear improvement of 9.86% (24.04% to 21.67%) relative average phone error rate (PER) reduction.

Index Terms: speech recognition, speaker adaptation, FCMLLR

1. Introduction

In automatic speech recognition (ASR), a critical issue that can seriously degrade the achievable performance is the mismatch between acoustic model and testing data. One major mismatch source is speaker variability, which can be divided into inter- and intra- speaker variability. The former one represents the variations between different speakers, and the later one indicates the pronunciation changes within a same speaker. It is difficult to handle this variability mismatch by calculating a perfectly-matched acoustic model for each testing speaker because the available amount of training data from each particular speaker is usually limited. Therefore, a variety of speaker adaptation methods has been proposed. Among these methods, maximum likelihood linear regression (MLLR) has been proven effective and widely used in the ASR applications [1]. The MLLR approach uses affine transforms to project the parameters in the original acoustic model to form a new speaker-specific acoustic model. Because MLLR only calculates the affine transforms instead of re-estimating the entire set of parameters, the new acoustic model can be estimated efficiently with a small amount of data.

Many approaches have been proposed to extend the conventional MLLR approach. Some extensions incorporate the prior knowledge to avoid possible over-fittings in estimating the affined transforms. Maximum a posteriori linear regression (MAPLR) and structural MAPLR are successful examples [2-4]. Some other extensions increase the discriminative capability of the acoustic model through the transformation process. Notable examples include minimum classification error linear regression (MCELR) [5, 6], minimum phone error linear regression (MPELR) [7] and soft margin estimation linear regression (SMELR) [8].

Another category of extensions combines the MLLR process with additional mechanisms. The fuzzy clustering MLLR (FCMLLR) is an effective approach belonging to this category [9, 10]. For FCMLLR, a regression tree is first established to cluster the entire set of Gaussian mixtures. In addition to affine transforms, a set of fuzzy weights is computed for each mixture cluster. With the corresponding set of fuzzy weights and the affine transforms, FCMLLR forms a new affine transform for each cluster for adaptation. Both the shared affine transforms and fuzzy weights are estimated based on the ML criterion. Because the fuzzy weights enable FCMLLR to form more precise transformations for adaptation, it is reported that FCMLLR can provide better performance than the conventional MLLR [9, 10].

In this study, we propose a new speaker adaptation approach—discriminative fuzzy clustering maximum a posterior linear regression (DFCMAPLR), to refine the conventional FCMLLR approach. For DFCMAPLR, we use a discriminative objective function to estimate the fuzzy weights for increasing the separations among different parameters in the acoustic model. Moreover, we derive the MAP-based criterion to calculate the shared affine transforms for improving the estimation reliability.

2. Review of Fuzzy Clustering MLLR

In this section, we first review the FCMLLR framework. Then, the two online estimations, namely, the fuzzy weight estimation (FWE) and the affine transform estimation (ATE), in the FCMLLR framework are elaborated.

2.1. FCMLLR system overview

Figure 1 shows the FCMLLR framework. With a set of speaker independent (SI) model, Ω, and a prepared set of initial affine transformations, W^0, FCMLLR calculates a speaker-adaptive (SA) acoustic model, Ω', by a two-stage operation, namely, fuzzy weight estimation (FWE) and affine transform estimation (ATE). The initial affine transforms, W^0, are usually prepared by the conventional class-based MLLR approach [1]. With the adaptation data, O, and the initial affine transforms, FCMLLR first performs FWE to calculate a fuzzy weight set, V. Then, with the calculated V, FCMLLR performs ATE to estimate an affine transform set, W. With several iterations of FWE and ATE, we can obtain optimal sets of V and W. Then, V and W are combined to form new affine transforms to adapt Ω to Ω'.

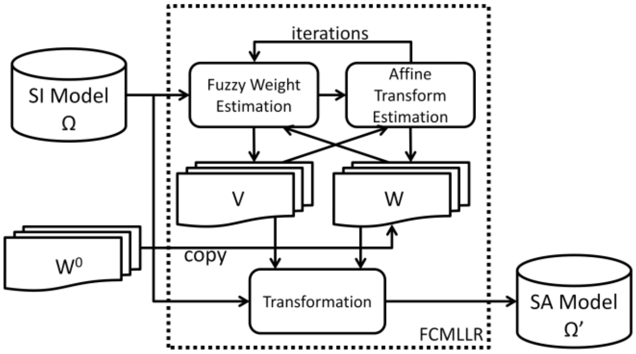

Figure 1: Block diagram of the FCMLLR framework.

2.2. ML-based FWE and ATE

The fuzzy weight estimation (FWE) and the affine transform estimation (ATE), in the FCMLLR framework are derived based on the ML criterion. Assume that Ω consists of I Gaussian components, to perform FCMLLR, we first group all the Gaussian components in Ω into J clusters. For the *i*-th Gaussian component belonging to the *j*-th cluster, FCMLLR transforms the original mean vector, μ_i, to adapted mean vector, $\hat{\mu}_i$ by:

$$\hat{\mu}_i = \sum_k v_{jk}\,(A_k\mu_i + b_k), \tag{1}$$

where A_k and b_k are the rotation matrix and bias vector of the *k*-th affine transforms, respectively; v_{jk} is the *k*-th element of the fuzzy weight vector for the *j*-th cluster. We rewrite (1) by:

$$\hat{\mu}_i = \sum_k (v_{jk}\,W_k)\xi_i, \tag{2}$$

where $W_k = [A_k\ b_k]$ and $\xi_i = [\mu_i^T, 1]^T$.

To perform adaptation on the entire set of mean vectors in Ω, FCMLLR estimates the complete fuzzy weight set, $\mathrm{V} = [v_{11}, \dots v_{jk}, \dots v_{JK}]$, and the affine transform set, $\mathrm{W} = [W_1, \dots W_k, \dots W_K]$ by a likelihood based objective function:

$$F_{\mathrm{ML}}(\mathrm{V}, \mathrm{W}, \Omega) = \log P(O|\mathrm{V}, \mathrm{W}, \Omega, \mathrm{U}_r). \tag{3}$$

where U_r is the correct label sequence corresponding to data, O.

By fixing W, FWE calculates V by maximizing (3) [9, 10]. The fuzzy weight vector for the *j*-th cluster, v_j, is obtained by:

$$\mathrm{v}_j = G_j^{-1} z_j, \tag{4}$$

where

$$G_j = \sum_{i \in C_j} \sum_t r_i(t) M_i^T \Sigma_i^{-1} M_i, \tag{5}$$

$$z_j = \sum_{i \in C_j} \sum_t r_i(t) M_i^T \Sigma_i^{-1} o_t, \tag{6}$$

where $\mathrm{v}_j = [v_{j1}, \dots v_{jk}, \dots v_{jK}]^T$ and $M_i = [W_1\xi_i, \dots W_k\xi_i, \dots W_K\xi_i]$. By performing (4) J times, we obtain the complete fuzzy weight set, V, where $\mathrm{V} = [\mathrm{v}_1, \dots \mathrm{v}_j, \dots \mathrm{v}_J]$.

Similarly by fixing V, ATE calculates W by maximizing (3). We solve W in a same manner to that in the conventional MLLR. When using a diagonal matrix for the covariance of each Gaussian component, we can solve W line by line:

$$\mathrm{W}^l = (G^l)^{-1} z^l, \tag{7}$$

where

$$G^l = \sum_i \frac{1}{\sigma_i^l} \sum_t r_i(t)\varsigma_i\varsigma_i^T, \tag{8}$$

$$z^l = \sum_i \frac{1}{\sigma_i^l} \sum_t r_i(t) o_t^l \varsigma_i, \tag{9}$$

where W^l is the *l*-th row of W, o_t^l is the *l*-th element of o_t; σ_i^l, is the (*l*,*l*)-th element of covariance matrix of *i*-th mixture; $\varsigma_i = [v_{j1}\xi_i^T, v_{j2}\xi_i^T, \dots v_{jk}\xi_i^T]^T$, where $[v_{j1}, \dots v_{jk}, \dots v_{jK}]$ are obtained by the previous FWE operation.

3. Discriminative Fuzzy Clustering MAPLR (DFCMAPLR)

Compared to the conventional MLLR approach, the fuzzy weights enable FCMLLR to form a more precise transformation for model adaptation and thus provide better performance. Here, we propose the discriminative fuzzy clustering MAPLR (DFC MAPLR) approach to further improve FCMLLR by incorporating the prior knowledge and discriminative information. In this section, we first introduce the MAP-based and the discriminative objective functions used for FWE and ATE and then elaborate the DFCMAPLR framework.

3.1. MAP-based and Discriminative Objective Functions

Generally for the MAP-based model adaptation approaches, we define the following objective function:

$$F_{\mathrm{MAP}}(\Gamma, \Omega) = \log(P(O|\Gamma, \Omega, \mathrm{U}_r)P(\Gamma, \Omega)), \tag{10}$$

where Γ is a transformation.

On the other hand, we prepare the following MAP-based objective function to perform FWE and ATE:

$$F_{\mathrm{MAP}}(\mathrm{V}, \mathrm{W}, \Omega) = \log(P(O|\mathrm{V}, \mathrm{W}, \Omega, \mathrm{U}_r)P(\mathrm{V}, \mathrm{W}, \Omega)). \tag{11}$$

For the conventional methods, we can directly prepare the prior information for Γ in (10) [2-4]. However for (11), because V and W are estimated iteratively, it is not easy to prepare a suitable prior density pair for V and W. Here, we directly prepare the prior density of the adapted acoustic mean vectors, Λ', where $\Lambda' \in \{\mu_i', i = 1 \dots I\}$. Then (11) becomes:

$$F_{\mathrm{MAP}}(\mathrm{V}, \mathrm{W}, \Omega) = \log(P(O|\mathrm{V}, \mathrm{W}, \Omega, \mathrm{U}_r)P(\Lambda')), \tag{12}$$

where

$$P(\Lambda') = \textstyle\prod_i p(\mu_i'). \tag{13}$$

We prepare the prior density by the Gaussian density function:

$$p(\mu_i') = N(\eta_i, \mathrm{X}_i), \tag{14}$$

where μ_i' is the mean of *i*-th Gaussian mixture of adapted model; η_i and X_i are hyper-parameters.

As introduced in Section 1, many discriminative objective functions have been proposed for model adaptation [5-7]. In this study, we choose the likelihood ratio (LR) as the objective function, which can be formulated as:

$$F_D(\mathrm{V}, \mathrm{W}, \Omega) = \log\big(P(O|\mathrm{V}, \mathrm{W}, \Omega, \mathrm{U}_r)\big) - \lambda_n \sum_n \log\big(P(O|\mathrm{V}, \mathrm{W}, \Omega, \mathrm{U}_n)\big), \tag{15}$$

where U_n and λ_n are label sequences and the scaling factor for the *n*-th competing list.

Note that both the MAP-based and discriminative objective functions can be used to estimate V and W. Therefore, we can have several different combinations of the V and W estimations. We consider the proposed DFCMAPLR the most suitable combination based on two reasons. First, the affine transformation set, W, consists of a large amount of parameters. By using the MAP criterion for ATE, we can avoid over-fittings that might occur when calculating W. Second, the fuzzy weight set, V, characterizes affine transformation more precisely for each cluster of Gaussian components. By applying a discriminative objective function for WFE, we can effectively enhance the discriminative power of the adapted acoustic model.

3.2. DFCMAPLR

For DFCMAPLR, we calculate the fuzzy weight set, V in Figure 1, by maximizing (15). We follow the same solution in (4) to calculate V, with G_j and z_j calculated by:

$$G_j = \sum_{i \in C_j} \sum_t r_i(t) M_i^T \Sigma_i^{-1} M_i \\ - \sum_n \lambda_n \sum_{m \in C_j} \sum_t r_m(t) M_m^T \Sigma_m^{-1} M_m, \qquad (16)$$

$$z_j = \sum_{i \in C_j} \sum_t r_i(t) M_i^T \Sigma_i^{-1} o_t \\ - \sum_n \lambda_n \sum_{m \in C_j} \sum_t r_m(t) M_m^T \Sigma_m^{-1} o_t, \qquad (17)$$

where $M_m = [W_1 \xi_m, \ldots W_k \xi_m, \ldots W_K \xi_m]$, and $\xi_m = [\mu_m^T, 1]^T$; $r_m(t)$ is the occupation probability of *m*-th Gaussian component at time t; μ_m^T and Σ_m^{-1} are the mean vector and covariance matrix of *m*-th Gaussian component, respectively. Here the *m*-th Gaussian component is the *n*-th competitor to the target *i*-th Gaussian component. We can find the competitor components from the competing list in (15). In our implementation, at each time t, we generate a correct label u_{tr} and *N* competing labels u_{tn}, *n*=1…*N*. From the correct and competing label sets, we find the target component and its *N* competitor components, respectively, to calculate G_j in (16) and z_j in (17).

DFCMAPLR calculate the affine transformation set, W in Figure 1, based on the MAP-based objective function in (12). We follow the same solution of (7), with $G^l$and z^l estimated by:

$$G^l = \sum_i \frac{1}{\sigma_i^l} \sum_t r_i(t) \varsigma_i \varsigma_i^T + \sum_i \frac{h}{x_i^l} \varsigma_i \rho_i^T, \qquad (18)$$

$$z^l = \sum_i \frac{1}{\sigma_i^l} \sum_t r_i(t) o_t^l \varsigma_i + \sum_i \frac{h}{x_i^l} \eta_i^l \varsigma_i, \qquad (19)$$

where $\rho_i = \left[v_{j1}\theta_i^T, v_{j2}\theta_i^T, \ldots, v_{jK}\theta_i^T\right]^T$, and $\theta_i = [\eta_i^T, 1]$; h is a scaling factor that controls the ratio of ML and prior distribution; x_i^l is the (*l*,*l*)-th element of X_i. In this study, we simply set $\eta_i = \mu_i$ and $\mathrm{X}_i = \alpha \Sigma_i$, where α is a scaling factor. With this setting, we can constrain the affine transforms by the prior information from the training set.

4. Experiment

The ASTTEL200 Mandarin corpus was used to prepare training and testing sets in all the experiments. This corpus consists of 200 speakers (100 male and 100 female), where each speaker pounced 200 Mandarin utterances. The average length of each utterance is 3~4 seconds. We chose 5 male and 5 female speakers and 100 utterances from each of them to form the testing set, amounting 1000 utterances in total. Speech data from the rest 190 speakers were used to form the training set, amounting 38,000 utterances in total. We performed the adaptation process in a supervised mode. For each testing speaker, we prepared 10 utterances (excluded from the 100 testing utterances) with the corresponding correct transcriptions as adaptation data.

We used hidden Markov toolkit (HTK) [9] to conduct feature extraction, training acoustic model, and testing recognition. Speech waveforms were characterized by standard 39 dimensional MFCC components, consisting of 13 static, delta, and delta-delta coefficients. We used the ML criterion to train a context-dependent triphone-based speaker independent (SI) model. Each triphone model was characterized by an HMM, which consists of 3 states, with 24 Gaussian mixtures per state. We performed free phone decoding to test recognition and reported the results by phone error rates (PERs). Each PER result directly presents the performance of a particular acoustic model adaptation approach without influences from the language model. We reported the results in average PERs on three sets, namely set F (for the 5 female testing speakers), set M (for 5 male testing speakers), and set All (for the overall testing 10 speaker).

A tree structure was constructed based on the Gaussian mixtures of the SI model. We set two thresholds, R_{W} and R_{V}, to determine the numbers of shared affine transforms and fuzzy cluster weight vectors, respectively. With the available adaptation data, we first calculated the accumulated statistics for every node in the tree. For example, we calculated the accumulated statistics, $\mathrm{A}_q = \sum_{i \in C_q} \sum_t r_i(t)$, for the *q*-th node. If A_q is larger than the threshold R_{W}, then we prepare an affine transform for the *q*-th node. If not, we check the accumulated statistics at the parent node. To determine the number of fuzzy cluster weight vectors, the same procedure is performed but with the threshold R_{V}. Because we intend to use the fuzzy weight vectors to estimate transforms in more detail, we set a larger number for R_{W} than R_{V}.

4.1. FWE and ATE

First, we investigate the performances achieved by the FWE and ATE. We show the results of the conventional MLLR, FCMLLR with FWE along, and FCMLLR with FWE+ATE on set M as "MLLR", "FWE", and "FWE+ATE" in Figure 2. To have a fair comparison, the number of affine transoms used in "MLLR" is the same to that of shared affine transforms used in "FWE", and "FWE+ATE". This number is determined by the threshold R_{W}

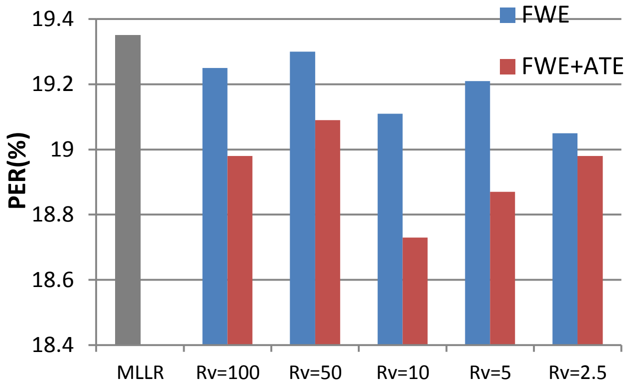

Figure 2. PERs(%) with different occupation threshold (R_{V}) for FCMLLR on set M.

and the prepared tree structure. In this set of experiments, we set R_{W}=500. Since only one iteration of FWE and ATE was performed, the affine transforms for "MLLR" were the same to the shared affine transforms in "FWE". For "FWE" and "FWE+ ATE", we varied the occupation threshold, R_{V}, from 100 to 2.5. This gave various numbers of clusters, J, and accordingly various numbers of fuzzy weight sets, ranging from 10 to 200.

From Figure 2, we first observe that for all R_{V}, "FWE" outperforms "MLLR". This verifies that the fuzzy weights enable FCMLLR to form a more precise transformation than MLLR and thus characterize the testing speaker better than the conventional MLLR. Next, we observe that "FWE+ATE" outperforms "FWE". This result confirms the improvements achieved by ATE. Since R_{V}=10 gives the best performance, we use this setting in the following experiments. From Figure 2, we also note that ATE gives more improvements than FWE. Therefore in the following, we first discuss the methods to improve the ML-based ATE and then the methods to improve the ML-based FWE.

4.2. Comparison of Different Criteria for ATE

In this set of experiments, we fix the ML-based WFE and compare the performances obtained by different objective functions for ATE. Table 1 lists the baseline (using the SI model) and the conventional MLLR results on sets F, M and All as "Baseline" and "MLLR" in the first and second rows. In the third, fourth and fifth rows, we list the results of ML-based FWE with ML-based, MAP-based, and discriminative ATE as "ML+ML", "ML+MAP", and "ML+D", respectively. Please note that "ML+ ML" also stands for the FCMLLR results.

From Table 1, we first observe that all the ATE techniques with ML-based FWE outperform "Baseline" and "MLLR" for all the three testing sets. This set of results indicates that by using ML-based FWE with ML-based, MAP-based and discriminative ATE, we can achieve better recognition performance than the baseline system and the conventional MLLR. Next, we observed that "ML+MAP" outperforms both "ML+ML" and "ML+D". This set of results suggests that MAP-based ATE can give better performance than the ML-based and discriminative ATE. Since the MAP-based ATE gives the best performance in this set of experiments, we fix the MAP-based ATE and compare different criteria for FWE in the next section.

Table 1: PERs(%) with different criteria for ATE (with fixed ML-based FWE) on sets F, M, and All.

	set F	set M	set All
Baseline	26.70	21.38	24.04
MLLR	25.98	19.35	22.67
ML+ML	25.90	18.73	22.32
ML+MAP	**25.51**	**18.64**	**22.08**
ML+D	25.63	18.77	22.20

4.3. Comparison of Different Criteria for FWE

In this section, we show the performances obtained by different objective functions for FWE with MAP-based ATE. Table 2 lists the results of MAP-based ATE, with ML-based, MAP-based and discriminative FWE as "ML+MAP", "MAP+MAP", and "D+ MAP" in the first, second, and third rows, respectively. Note that "D+MAP" stands for the proposed DFCMAPLR results.

From Table 1, we observed that "D+MAP" outperforms both "ML+MAP" and "MAP+MAP". This result confirms that DFCMAPLR provides the best performance among different combinations. Next from the results in Tables 1 and 2, we observe that "ML+MAP", "MAP+MAP" and "D+MAP" achieve better performance than both "Baseline" and "MLLR". By comparing the results of "ML+ML" in Table 1 and "D+MAP" in Table 2, we verified that DFMAPLR outperforms the conventional FCMLLR. When compared to the baseline result, DFMAPLR gives a clear improvement of 9.86% (24.04% to 21.67%) relative PER reduction on the set All.

Table 2: PERs(%) with different criteria for FWE (with fixed MAP-based ATE) on sets F, M, and All.

	set F	set M	set All
ML+MAP	25.51	18.64	22.08
MAP+MAP	25.34	18.41	21.88
D+MAP	**25.03**	**18.31**	**21.67**

5. Conclusions

We proposed a new speaker adaptation approach—DFCMAPLR, to extend the FCMLLR approach by incorporating the discriminative information and prior knowledge. We evaluated the proposed DFCMAPLR approach in a supervised adaptation mode. The evaluation results first confirmed the effectiveness of the use of fuzzy clustering weights to improve the model adaptation capability. Moreover, the results verified that DFCMAPLR outperforms both the conventional MLLR and FCMLLR approaches with a same amount of adaptation utterances. When compared to the baseline result, DCFMAPLR gives a significant 9.86% relative average PER reduction over ten testing speakers.

6. 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, pp 171-185, 1995.

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

[3] O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech and Language,* vol. 16, pp. 5-24, 2002.

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

[5] J. Wu and Q. Huo, "A study of minimum classification error (MCE) linear regression for supervised adaptation of MCE-trained continuous-density hidden Markov models," *IEEE Trans. Speech Audio Processing*, vol. 15, pp. 478-488, 2007.

[6] X. He and W. Chou, "Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs," in *Proc. ICASSP*, pp 556-559, 2003.

[7] L. Wang and P. C. Woodland, "MPE-based discriminative linear transform for speaker adaptation," *Computer Speech and Language*, vol. 22, pp 256-272, 2008.

[8] S. Matsuda, Y. Tsao, J. Li, S. Nakamura, and C. Lee, "A study on soft margin estimation of linear regression parameters for speaker adaptation", in *Proc. Interspeech,* 2009, pp.1603-1606.

[9] Gales M.J.F., "The generation and use of regression class trees for MLLR adaptation," Technical Report CUED/F-INFENG/TR263, Cambridge University, 1996.

[10] Ing-Jr Ding, "Incremental MLLR speaker adaptation by fuzzy logic control," *Pattern Recognition*, vol. 40 no. 11, pp. 3110-3119, 2007.

[11] S. Young et al., The HTK Book, Cambridge University Engineering Department, 2005.