AGGREGATE *A POSTERIORI* LINEAR REGRESSION FOR SPEAKER ADAPTATION

Chih-Hsien Huang and Jen-Tzung Chien

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan 70101, ROC
acheron@chien.csie.ncku.edu.tw & jtchien@mail.ncku.edu.tw

ABSTRACT

In this paper, we present a rapid and discriminative speaker adaptation algorithm for speech recognition. The adaptation paradigm is constructed under the popular linear regression transformation framework. Attractively, we estimate the regression matrices from the speaker-specific adaptation data according to the aggregate *a posteriori* criterion, which can be expressed in a form of classification error function. The goal of proposed aggregate *a posteriori* linear regression (AAPLR) turns out to estimate the discriminative linear regression matrices for transformation-based adaptation so that the classification errors can be minimized. Different from minimum classification error linear regression (MCELR), AAPLR algorithm ha closed-form solution to achieve rapid speaker adaptation. The experimental results reveal that AAPLR speaker adaptation does improve speech recognition performance with moderate computational cost compared to the maximum likelihood linear regression (MLLR), maximum *a posteriori* linear regression (MAPLR) and MCELR.

1. INTRODUCTION

In general, the hidden Markov model (HMM) parameters are trained using two categories of approaches: the distribution estimation and the discriminative training. The popular algorithms for distribution estimation are based on maximum likelihood (ML) and maximum *a posteriori* (MAP) criteria [6]. Also, the criteria for discriminative training are using the minimum classification error (MCE) [8] and the maximum mutual information (MMI) [1]. Using MCE discriminative training criterion, the generalized probabilistic descent (GPD) algorithm is applied to iteratively learn the model parameters. However, it is time-consuming to fulfill iterative GPD algorithm. In this study, we focus on developing a rapid and discriminative algorithm for *speaker adaptation*. We would like to adapt the existing HMM parameters to a new speaker and his/her operating environments so as to improve the speech recognition performance. In the literature, the linear regression adaptation using MLLR [9] and MAPLR [3][5] is popular and shown to be effective. Here, we concern the issue of discriminative adaptation where the likelihoods from target HMM's as well as that from competing HMM's are considered to determine the most likely regression matrices. The adaptation performance can be significantly improved. To avoid extensive computation, our goal is to derive closed-form regression matrices for rapid adaptation.

The discriminative training algorithm based on MCE criterion has been applied to speaker adaptation [2][12]. The adapted speech models are discriminative so as to reduce the classification error rates. In [2], the MCE criterion was merged to perform the linear transformation and estimate the time-varying polynomial Gaussian mean functions for the trended HMM. This approach was called minimum classification error linear regression (MCELR). Although the speech recognition performance was improved, the major weakness came from the heavy computational cost. To compensate this weakness and reinforce the adaptation robustness, we properly incorporate the prior density of linear regression model and conduct the Bayesian model estimation. Interestingly, we present the aggregate *a posteriori* linear regression (AAPLR) where the aggregate *a posteriori* probability [10] is maximized to find the optimal regression matrices. A closed-form solution to AAPLR is derived to achieve fast and discriminative speaker adaptation. For comparison, we carry out different linear regression adaptation algorithms in the experiments.

2. RELATED ALGORITHMS

Before describing the new discriminative transformation-based adaptation algorithm, we are introducing the MCE criterion and the linear regression adaptations using MLLR, MAPLR and MCELR.

2.1. MCE Criterion

Juang et al. [8] presented the MCE criterion with a three-step procedure. For the case of *M*-category classification, the first step is to determine the discriminant functions $\{g_m(X;\lambda_m), m=1,\cdots,M\}$. Second, a misclassification measure is introduced as follows

$$d_m(X) = -g_m(X;\lambda_m) + \log\left[\frac{1}{M-1}\sum_{j\neq m}\exp\left[\eta g_j(X;\lambda_j)\right]\right]^{1/\eta}, \quad (1)$$

where η is a positive number and $\Lambda=\{\lambda_m\}$ is the model parameter. This function is continuous and flexible with varying η. Notably, all competing classes $j\neq m$ are considered during parameter learning. At the third step, the loss function measuring the classification errors is formulated by

$$l_m(X;\lambda_m) = l(d_m(X)) = \frac{1}{1+\exp(-\gamma d_m(X)+\theta)}. \quad (2)$$

The sigmoid function $l(\cdot)$ has parameters γ and θ. In (1), the positive value of $d_m(X)$ reflects the misclassification while $d_m(X)<0$ implies the correct classification. Then, the generalized probabilistic decent algorithm is developed to iteratively fulfill the MCE criterion. The learning algorithm of Λ is given by

$$\Lambda^{i+1} = \Lambda^i - \varepsilon U \nabla l(X;\Lambda^i) \ . \tag{3}$$

Here, i is the iteration index, X are the training samples, U is the positive definite matrix and ε is the learning rate. This paper concerns the discriminative linear regression adaptation rather than discriminative HMM training. In what follows, we are describing several variants of linear regression adaptation and the conceptual evolution from MCE discriminative training to proposed discriminative linear regression adaptation.

2.2. MLLR, MAPLR and MCELR

The linear regression speaker adaptation aims to estimate the cluster-dependent regression matrices, which are used to transform/adapt the speaker-independent HMM parameters to a new speaker. By properly controlling the sharing of regression matrices, MLLR can effectively find the maximum likelihood estimate of regression matrices for adaptation of HMM mean vectors. Assume that a HMM distribution of λ_m having $D\times 1$ mean vector μ_m, the adapted mean vector $\hat{\mu}_m$ using $D\times(D+1)$ regression matrix $\mathbf{W}_{r(m)}$ is expressed by

$$\hat{\mu}_m = \mathbf{W}_{r(m)}\xi_m \ . \tag{4}$$

Here, D is feature dimension, $r(m)$ is the regression/cluster class of HMM label m and ξ_m is the extended mean vector $\xi_m = [1, \mu_m^T]^T$. The maximum likelihood estimate of regression matrices $\mathbf{W}_{\text{ML}} = \{\mathbf{W}_{r(m)}\}$ using adaptation data X is determined by

$$\mathbf{W}_{\text{ML}} = \arg\max_{\mathbf{W}} p(X|\mathbf{W},\Lambda) \ . \tag{5}$$

The expectation-maximization (EM) algorithm can be applied to find the optimal $\mathbf{W}_{\text{ML}}$. The regression matrices are calculated through solving a system of linear equations [9].

Further, when the amount of adaptation data is sparse, the estimated regression matrices are biased. It is helpful to achieve desirable adaptation performance by constraining the distribution shape of regression matrices using prior densities. In [3][5], the matrix-variate normal density served as the prior distribution for $\mathbf{W}$ so as to perform the maximum *a posteriori* estimation

$$\mathbf{W}_{\text{MAP}} = \arg\max_{\mathbf{W}} p(\mathbf{W}|X,\Lambda) = \arg\max_{\mathbf{W}} p(X|\mathbf{W},\Lambda)g(\mathbf{W}) \ . \tag{6}$$

The prior distribution of a regression matrix is defined by

$$g(\mathbf{W}) \propto |\Delta|^{-1/2} \cdot q\left(\sum_{d=1}^{D}(\mathbf{w}_d - \mathbf{m}_d)\Sigma_d^{-1}(\mathbf{w}_d - \mathbf{m}_d)^T\right), \tag{7}$$

where $\mathbf{m}_d$ and Σ_d are mean vector and covariance matrix for *d*th row of regression matrix $\mathbf{w}_d$, respectively and Δ is a block diagonal matrix $\text{diag}(\Sigma_1,\ldots,\Sigma_\text{D})$. Usually, q is an exponential function. The resulting maximum *a posteriori* linear regression (MAPLR) has better adaptation performance than MLLR.

In [2], Chengalvarayan proposed the minimum classification error linear regression (MCELR) adaptation algorithm where the parameters of global regression matrix were estimated according to the GPD algorithm. Wu and Huo [12] further performed MCELR adaptation using multiple regression classes. In general, the learning algorithm of $\mathbf{W}$ is similar to (3). Having the overall loss function $l(X;\mathbf{W}^i)$ due to current regression matrices $\mathbf{W}^i$, the updating of $\mathbf{W}$ is done iteratively by

$$\mathbf{W}^{i+1} = \mathbf{W}^i - \varepsilon\nabla l(X;\mathbf{W}^i) \ . \tag{8}$$

3. AGGREGATE *A POSTERIORI* CRITERIA

Subsequently, we are introducing the generalized minimum error rate (GMER) [10] criterion, which was proposed for discriminative training and done by Li and Juang. An aggregate *a posteriori* (AAP) probability was defined and rearranged to fit the formula of classification error criterion for discriminative training. Under certain assumptions, a closed-form solution to HMM training was obtained. Such solution allows us to perform efficient model training. In this study, we use AAP probability as an objective function to estimate the regression matrices for discriminative speaker adaptation rather than HMM parameters for model training.

3.1. GMER Criterion

In GMER training algorithm, the AAP probability is defined by

$$J_{\text{AAP}}(\Lambda) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\frac{p(X_{m,n}|\lambda_m)P_m}{p(X_{m,n})}, \tag{9}$$

where $X_{m,n}$ is the *n*th training sentence from the *m*th model λ_m and its length is T_n, $X_{m,n} = \{\mathbf{x}_{m,n,t}\}_{t=1}^{T_n}$. P_m represents the prior probability of class m. The aggregate *a posteriori* probability can be viewed as the averaged posterior probability over all aligned speech segments. Assume that the training data are i.i.d., the joint likelihood of training data $X_{m,n}$ and model λ_m is expressed by $p(X_{m,n}|\lambda_m) = \prod_{t=1}^{T_n} p(\mathbf{x}_{m,n,t}|\lambda_m)$. The AAP probability can be written as

$$J_{\text{AAP}}(\Lambda) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m} l(d_{\text{AAP}}(X_{m,n})) \ . \tag{10}$$

Here, $l(\cdot)$ is a loss function under the case of using $\gamma=1$ and $\theta=0$ in sigmoid function. The misclassification measure becomes

$$d_{\text{AAP}}(X_{m,n}) = \log p(X_{m,n}|\lambda_m)P_m - \log\sum_{j\neq m} p(X_{m,n}|\lambda_j)P_j \ . \tag{11}$$

In [10], a closed-form AAP solution was derived to estimate the HMM parameters

$$\Lambda_{\text{AAP}} = \arg\max_{\Lambda} J_{\text{AAP}}(\Lambda) \ . \tag{12}$$

3.2. AAPLR

However, for speaker adaptation, we are adjusting the HMM mean vectors to a new speaker using cluster-dependent linear regression matrices $\mathbf{W}=\{\mathbf{W}_{r(m)}\}$. The joint AAP probability of class and regression matrix is established by

$$J_{\mathrm{AAP}}(\mathbf{W})=\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\frac{p(X_{m,n}\left|\mathbf{W}_{r(m)},\lambda_m\right.)P_m g(\mathbf{W}_{r(m)})}{p(X_{m,n})}. \quad (13)$$

Again, the objective function of AAPLR can be rearranged as

$$J_{\mathrm{AAP}}(\mathbf{W})=\sum_{m=1}^{M}\sum_{n=1}^{N_m} l(d_{\mathrm{AAP}}(X_{m,n},\mathbf{W}_{r(m)})), \quad (14)$$

where

$$\begin{aligned} d_{\mathrm{AAP}}(X_{m,n},\mathbf{W}_{r(m)}) &= g_r(X_{m,n};\lambda_m,\mathbf{W}_{r(m)}) \\ &-\log\left\{\frac{1}{M-1}\sum_{j\neq m}\exp[g_{\bar{r}}(X_{m,n};\lambda_j,\mathbf{W}_{r(j)})]\right\}. \end{aligned} \quad (15)$$

Here, we specify $P_m=1$ and $P_j=1/(M-1)$ for $j\neq m$ and

$$g_r(X_{m,n};\lambda_m,\mathbf{W}_{r(m)})=\log\{p(X_{m,n}\left|\mathbf{W}_{r(m)},\lambda_m\right.)g(\mathbf{W}_{r(m)})\}. \quad (16)$$

where $g(\mathbf{W}_{r(m)})$ represents the prior density of regression matrix $\mathbf{W}_{r(m)}$. To simplify the estimation, we assume that the covariance matrix of HMM Gaussian distribution is diagonal, i.e.

$$\begin{aligned} N(\mathbf{x}_{m,n,t}\left|\xi_m,\Sigma_m,\mathbf{W}_{r(m)}\right.) &= (2\pi)^{-D/2}\left|\Sigma_m\right|^{-1/2} \\ &\times\exp\left[-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_{m,n,t,d}-\mathbf{w}_{r(m),d}\xi_m)^2}{\sigma_{md}^2}\right]. \end{aligned} \quad (17)$$

We estimate the D rows of regression matrix $\{\mathbf{w}_{r(m),d}, d=1,\cdots,D\}$ by solving the optimization problem

$$\mathbf{W}_{\mathrm{AAP}}=\arg\max_{\mathbf{W}} J_{\mathrm{AAP}}(\mathbf{W}). \quad (18)$$

The gradient of $J_{\mathrm{AAP}}(\mathbf{W})$ with respect to $\mathbf{w}_{r(m),d}$ is derived to be

$$\begin{aligned} \nabla_{\mathbf{w}_{r(m),d}} J_{\mathrm{AAP}}(\mathbf{W}) &= \sum_{m=1}^{M}\sum_{n=1}^{N_m} L_{r(m)}(X_{m,n})\cdot\Bigg\{\sum_t\Big(\frac{x_{m,n,t,d}-\mathbf{w}_{r(m),d}\xi_m}{\sigma_{md}^2}\Big)\xi_m^T \\ &+2(\mathbf{w}_{r(m),d}-\mathbf{m}_{r(m),d})\Sigma_{r(m),d}^{-1}-\Psi_{\bar{r}}(X_{m,n})\sum_{j\neq m}\Phi_j(X_{m,n}) \\ &\times\left\{\sum_t\Big(\frac{x_{m,n,t,d}-\mathbf{w}_{r(m),d}\xi_j}{\sigma_{jd}^2}\Big)\xi_j^T+2(\mathbf{w}_{r(m),d}-\mathbf{m}_{r(m),d})\Sigma_{r(m),d}^{-1}\right\}\Bigg\} \end{aligned} \quad (19)$$

where $\Psi_{\bar{r}}(X_{m,n})=\dfrac{1}{\sum_{j\neq m}\exp[g_{\bar{r}}(X_{m,n};\lambda_j,\mathbf{W}_{r(j)})]}$, $\Phi_j(X_{m,n})=\exp[g_j(X;\lambda_j,\mathbf{W}_{r(j)})]$ and $L_{r(m)}(X_{m,n})=l(d_{\mathrm{AAP}}(X_{m,n},\mathbf{W}_{r(m)}))\times(1-l(d_{\mathrm{AAP}}(X_{m,n},\mathbf{W}_{r(m)})))$. By setting $\nabla_{\mathbf{w}_{r(m),d}} J_{\mathrm{AAP}}(\mathbf{W})=0$, a closed-form solution to $\mathbf{w}_{r(m),d}$ is yielded by solving the linear equation

$$\begin{aligned} &\mathbf{w}_{r(m),d}^{\mathrm{AAP}}\cdot\left\{\begin{aligned} &\sum_{m=1}^{M}\sum_{n=1}^{N_m} L_{r(m)}(X_{m,n})\Bigg[T_n\Big(\frac{\xi_m\xi_m^T}{\sigma_{md}^2}\Big)- \\ &T_n\Psi_{\bar{r}}(X_{m,n})\sum_{j\neq m}\Phi_j(X_{m,n})\Big(\frac{\xi_j\xi_j^T}{\sigma_{jd}^2}\Big) \\ &-2(1-\Psi_{\bar{r}}(X_{m,n})\sum_{j\neq r}\Phi_j(X_{m,n}))\Sigma_{r(m)d}^{-1}\Bigg] \end{aligned}\right\} \\ &=\left\{\begin{aligned} &\sum_{m=1}^{M}\sum_{n=1}^{N_m} L_{r(m)}(X_{m,n})\Bigg[\sum_{t=1}^{T_n}\Big(\frac{x_{m,n,t,d}}{\sigma_{md}^2}\Big)\xi_m^T- \\ &\Psi_{\bar{r}}(X_{m,n})\sum_{j\neq m}\Phi_j(X_{m,n})\sum_{t=1}^{T_n}\Big(\frac{x_{m,n,t,d}}{\sigma_{jd}^2}\Big)\xi_j^T \\ &-2(1-\Psi_{\bar{r}}(X_{m,n})\sum_{j\neq r}\Phi_j(X_{m,n}))\mathbf{m}_{r(m)d}\Sigma_{r(m)d}^{-1}\Bigg] \end{aligned}\right\} \end{aligned}. \quad (20)$$

Without loss of generality, we can express (20) in matrix form $\mathbf{w}_{r(m),d}\cdot\mathbf{L}=\mathbf{r}$ where $\mathbf{L}$ is a $(D+1)\times(D+1)$ matrix and $\mathbf{r}$ is a $1\times(D+1)$ vector. Finally, we solve a linear equation to find $\mathbf{W}_{r(m)}=\{\mathbf{w}_{r(m),d}, d=1,\cdots D\}$ via

$$\mathbf{w}_{r(m),d}=\mathbf{r}\cdot\mathbf{L}^{-1}. \quad (21)$$

4. EXPERIMENTS

In the experiments, we carried out the linear regression speaker adaptation algorithms of MLLR, MAPLR, MCELR and AAPLR for continuous Mandarin speech recognition. Mandarin is a tonal and syllabic language. We conducted the base syllable recognition for comparative study. A broadcast news speech recognition task was performed to examine the performance of speaker adaptation. In this study, we prepared two speech corpora for HMM training and adaptation. The speaker-independent HMM's were trained using the benchmark Mandarin speech corpus TCC300 [4] which was recorded in office environments using close-talking microphones. We sampled 14266 sentences (about 16 hours) recorded by 100 males and 100 females for training. The construction of context-dependent sub-syllable HMM's for Mandarin speech was referred to [4]. The adaptation and testing data sets were sampled from the MATBN database [11]. MATBN database contained Mandarin Chinese broadcast news (PTSN) utterances, which were shared by the Public Television Service Foundation of Taiwan and collected by the Institute of Information Science at Academia Sinica, Taiwan. MATBN was a 120-hour broadcast news corpus. We performed two-pass adaptation prior to speech recognition; task adaptation and speaker adaptation. In task adaptation, we used 200 utterances (about 30 minutes) randomly sampled from MATBN database to adapt the existing HMM's to fit the broadcast news transcription task. In speaker adaptation, we collected sixty utterances (about 14 minutes) from one male reporter and one female reporter and performed linear regression adaptation. The other 40 utterances (about 9 minutes) from the same speaker were adopted for speech recognition. For all experiments, we used 26 dimensional feature vectors consisted of twelve Mel-frequency cepstral coefficients, one log energy and their first derivatives. Several sets of experiments on supervised adaptation were reported. To evaluate the performance versus adaptation data length, we performed speaker adaptation using ten, thirty and seventy adaptation utterances. The number of regression classes was fixed to be four for all cases. Each utterance was about three seconds. In Table 1, the syllable error rates (%) are reported through five-fold cross-validation over the adaptation data set.

		MLLR	MAPLR	MCELR	AAPLR
Number of Adaptation Data	10	32.5	31.5	31.1	30.6
	20	30.5	29.7	29.6	29.2
	60	29.0	28.4	28.1	27.9

Table 1 Syllable error rates (%) of supervised adaptation using various adaptation algorithms.

		MLLR	MAPLR	MCELR	AAPLR
Number of Adaptation Data	10	42.7	43.4	52.5	46.6
	20	52.7	53.4	65.6	57.1
	60	86.8	85.1	106.0	93.1

Table 2 Adaptation time (second) of supervised adaptation with different adaptation algorithms.

Without performing adaptation, the baseline syllable error rate (SER) is 53.3%. After performing task adaptation, SER is greatly reduced to 41.6%. This implies that the environmental mismatch between databases of TCC300 and MATBN is significant and can be compensated by task adaptation. Namely, it is important to perform environmental adaptation for a new task of broadcast news transcription. Further, when performing linear regression speaker adaptation, we find that the SER's are reduced by using MLLR, MAPLR, MCELR and AAPLR under different numbers of adaptation data. At the case of ten adaptation utterances, AAPLR obtains SER of 30.6%, which is better than those of MLLR (32.5%), MAPLR (31.5%) and MCELR (31.1%). As the number of adaptation data increases to sixty, all recognition results are improved accordingly. The performance of AAPLR (27.9%) is still superior to those of MLLR (29.0%), MAPLR (28.4%) and MCELR (28.1%). The reasons are due to the incorporation of prior regression information and the discriminant capability when using AAPLR.

Also, we investigate the computational costs of MLLR, MAPLR, MCELR and AAPLR adaptation algorithms. The processing time for different adaptation algorithms was measured on a personal computer with CPU Pentium 4 2.0 GHz and RAM 256 MB. Table 2 reveals that MCELR spends additional 20% computation time relative to MLLR. However, the computation costs of AAPLR and MLLR are comparable. It is because that both algorithms realize the closed-form linear equations. These results show the superiority of AAPLR for speaker adaptation.

5. CONCLUSION

We have presented a new AAPLR algorithm for rapid and discriminative speaker adaptation. The adapted speech HMM's using discriminative regression matrices were able to enhance the speech recognition performance for broadcast news transcription. The AAP criterion was introduced to achieve model discriminability and simultaneously derive a closed-form solution for rapid parameter estimation. More importantly, we established AAPLR algorithm in which a closed-form solution to regression matrices was derived to achieve desirable adaptation performance. HMM parameters of all acoustic units could be effectively adapted. To improve system robustness, the prior information of transformation matrix was incorporated for constrained estimation. Such Bayesian approach enabled the proper estimation of regression matrices. From the experiments, we find that the recognition rate of using AAPLR for supervised speaker adaptation is higher than those of MLLR, MAPLR and MCELR for different numbers of adaptation data. Also, the computation cost of AAPLR is smaller than MCELR and moderate compared to MLLR and MAPLR.

6. REFERENCES

[1] L. Bahl, P. Brown, P. de Souza and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", in *Proc. ICASSP*, vol. 11, pp. 49-52, 1986.

[2] R. Chengalvarayan, "Speaker adaptation using discriminative linear regression on time-varying mean parameters in trended HMM", *IEEE Signal Processing Letters*, vol. 5, pp. 63-65, 1998.

[3] C. Chesta, O. Siohan, and C.-H. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation," in *Proc. EUROSPEECH*, vol. 1, pp. 211-214, 1999.

[4] J.-T. Chien and C.-H. Huang, "Bayesian learning of speech duration models", *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 558-567, 2003.

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

[6] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains", *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 291-298, 1994.

[7] P. S. Gopalakrishnan, D. Kanevsk, A. Nádas and D. Nahamoo, "An inequality for rational function with applications to some statistical estimation problems," *IEEE Trans. Inform. Theory*, vol. 37, pp. 107-113, 1991.

[8] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum classification error rate methods for speech recognition", *IEEE Trans. Speech Audio Processing*, vol. 5, no. 3, pp. 257-265, 1997.

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

[10] Q. Li and B.-H. Juang, "Fast discriminative training for sequential observations with application to speaker identification", in *Proc. ICASSP*, vol. 2, pp. 397-400, 2003.

[11] H. M. Wang, "MATBN 2002: A Mandarin Chinese broadcast news corpus," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, 2003.

[12] J. Wu and Q. Huo, "Supervised adaptation of MCE-trained CDHMMs using minimum classification error linear regression," in *Proc. ICASSP*, vol. 1, pp. 605-608, 2002.