MINIMUM CLASSIFICATION ERROR LINEAR REGRESSION (MCELR) FOR SPEAKER ADAPTATION USING HMM WITH TREND FUNCTIONS

Rathinavelu Chengalvarayan

Currently at: Speech Processing Group, Bell Labs
Lucent Technologies, Naperville, IL 60566, USA
Tel: (630) 224 6398, Fax: (630) 979 5915
Email: rathi@lucent.com

ABSTRACT

In this paper, we report our recent work on applications of the combined MLLR and MCE approach to estimating the time-varying polynomial Gaussian mean functions in the trended HMM. We call this integrated approach as the minimum classification error linear regression (MCELR), which has been described in this study. The transformation matrices associated with each polynomial coefficients are calculated to minimize the recognition error of the adaptation data and is developed using the gradient descent algorithm. A speech recognizer based on these results is implemented in speaker adaptation experiments using TI46 corpora. Results show that the trended HMM always outperforms the standard HMM and that adaptation of linear regression coefficients is always better when fewer than three adaptation tokens are used.

1. INTRODUCTION

In the last couple of years, there has been much interest in the area of *feature-space* transformation and *model-space* transformation based adaptation to reduce the recognition errors caused by acoustic mismatches between the training and testing conditions [1], [3], [5], [9], [15]. Previous experiments showed that the model-space approach results in a significant improvement over the feature-space method [13]. In the current study, we will follow this model-space transformation scheme [15], which adapts a set of speaker independent models to a specific speaker by applying a set of linear transformations to the Gaussian mean vectors. Each transformation is used for a number of Gaussian distributions, and the number of transformations is determined the amount of adaptation available. The parameters of transformation matrices are estimated to maximize the likelihood of the speaker specific data.

The formulation of the trended HMM (or trajectory-based HMM or nonstationary-state HMM) has been successfully used in automatic speech recognition applications for the past few years [2], [4]. The model parameters of the trended HMM (state-dependent time-varying means and variances) used in the past were trained using Viterbi-like algorithms based on the joint-state maximum likelihood principle (ML). The method of ML, however, need not be optimal in terms of minimizing classification error rate in recognition tasks in which the observation is assumed to be produced by one of the many source classes. Discrimination can be improved if out-of-class information is also used in training the models. Another alternative reestimation criterion, called minimum classification error training (MCE) has been developed for trended HMM to improve the discriminating ability of ML criterion [11]. This training approach aims at directly minimizing the recognition error rate of the training data by taking into account other competing models and has recently been used in speaker adaptation applications [6], [7], [8], [14].

In this paper, we extend the ML-based Viterbi algorithm to the MCE training algorithm for optimally estimating the linear transformations to the set of polynomial mean vectors in the trended HMM. We shall call this integrated approach as the MCELR, which has been described in this study. The transformations are set to be different for different trend parameters. Hence if we use quadratic trend functions, then we must have three transformation matrices, one for the intercept, one for the slope and the other for the quadratic polynomial coefficients. The MCELR takes some adaptation data from a new speaker and updates the regression matrices to minimize the classification errors on the adaptation data and is implemented using the gradient descent algorithm. The regression matrices linearly transform the trend mean parameters in order to map them to the test speakers. The other HMM parameters are not adapted since the main differences among speakers are assumed to be captured by the means. Although gains can be made by using state-dependent linear transforms [12], we consider transformations on a global basis for the case of small amount of adaptation data.

2. THE TRENDED HMM INCORPORATING LINEAR REGRESSION MATRICES

The trended HMM is of a data-generative type and can be described as

$$\mathcal{O}_t \quad = \quad \sum_{p=0}^{P} \mathcal{B}_i(p)(t-\tau_i)^p + \mathcal{R}_t(\Sigma_i), \qquad (1)$$

where $\mathcal{O}_t$, $t = 1, 2, \cdots, T$ is a modeled observation data sequence of length T, within the HMM state indexed by i; $\mathcal{B}_i(p)$ are state-dependent polynomial regression coefficients of order P indexed by state i; and the term $\mathcal{R}_t$ is the stationary residual assumed to be independent and identically distributed (IID) and zero-mean Gaussian source characterized by state-dependent, but time-invariant covariance matrix Σ_i. The term $t - \tau_i$ represents the sojourn time in state i at time t, where τ_i registers the time when state i in the HMM is just entered before regression on time takes place.

The MCELR approach to speaker adaptation accepts a small of data from a new speaker and modifies the speaker independent polynomial mean parameters to minimize the classification error rate of the adaptation data. The remaining model parameters are not updated since the previous studies observed that the mean parameters are the most effective in representing the essential characteristics of a particular speaker [6], [10]. The adaptation of the mean parameter is performed by applying a global transformation matrix to each of the state-dependent polynomial coefficients according to

$$\hat{\mathcal{B}}_i(p) \quad = \quad \mathcal{W}(p)\mathcal{B}_i(p)$$

where $\mathcal{W}(p)$ is an $d \times d$ matrix, with d being the dimension associated with each polynomial coefficients which minimizes the recognition errors of the adaptation data. Each state of the adapted model is characterized by a multivariate Gaussian density function with diagonal covariance matrices in the form of

$$b_i(\mathcal{O}_t) = \frac{(2\pi)^{\frac{-d}{2}}}{|\Sigma_i|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}\left[\mathcal{O}_t - \sum_{p=0}^{P}\mathcal{W}(p)\mathcal{B}_i(p)(t-\tau_i)^p\right]^{Tr}\right.$$
$$\left.\Sigma_i^{-1}\left[\mathcal{O}_t - \sum_{p=0}^{P}\mathcal{W}(p)\mathcal{B}_i(p)(t-\tau_i)^p\right]\right)$$

where $\mathcal{B}_i(p)$, Σ_i denotes the polynomial coefficients for the time-varying mean functions and the variances for the i-th state, respectively; $(t - \tau_i)$ is the sojourn time in state i at time t and d is the dimensionality of vector $\mathcal{O}$. Superscripts Tr,-1 and the symbol $||$ denote the matrix transposition, inversion, and determinant, respectively.

3. ESTIMATION OF LINEAR REGRESSION MATRICES

In this section, the MCELR training process is briefly described. One major purpose of this study is to develop and implement the MCE-based discriminative training paradigm in the context of the trended HMM for achieving optimal estimation of the global regression matrices associated with each polynomial coefficients. Let Φ_j, $j = 1, 2, \cdots, \mathcal{K}$, denote the parameter set characterizing the trended HMM for the j-th class, where $\mathcal{K}$ is the total number of classes. The classifier based on these $\mathcal{K}$ class models can be characterized by $\Phi = \{\Phi_1, \Phi_2, \cdots, \Phi_{\mathcal{K}}\}$. The purpose of the MCE-based discriminative training is to find the parameter set Φ such that the number of misclassifying all the adaptation tokens is minimized.

3.1. Definition of Loss Function

Let $g_j(\mathcal{O}, \Phi)$ denote the log-likelihood associated with the optimal state sequence Θ for the input token $\mathcal{O}$, obtained by applying the Viterbi algorithm using model Φ_j for the j-th class. Then, for the utterance $\mathcal{O}$ (from class c), the misclassification measure $d_c(\mathcal{O}, \Phi)$ is determined by

$$d_c(\mathcal{O}, \Phi) \quad = \quad -g_c(\mathcal{O}, \Phi) + g_\chi(\mathcal{O}, \Phi), \qquad (2)$$

where χ denote the incorrect model with the highest log-likelihood (i.e., the most confusible class). In this definition, a negative value of $d_c(\mathcal{O}, \Phi)$ corresponds to a correct classification. The definition in Eqn. (2) focuses on the comparison between the true model and the best wrong model, an approximation which we adopt in this study for computation efficiency. A more general form of the misclassification measure using the log-likelihoods from all models can be found in [12]. A loss function with respect to the input token is finally defined in terms of the misclassification measure given by

$$\Upsilon(\mathcal{O}, \Phi) \quad = \quad \frac{1}{1 + e^{-d_c(\mathcal{O}, \Phi)}}, \qquad (3)$$

which projects $d_c(\mathcal{O}, \Phi)$ into the interval [0,1]. Note that the loss function $\Upsilon(\mathcal{O}, \Phi)$ is directly related to the classification error rate and is first-order differentiable with respect to each global regression matrix parameters.

3.2. Minimization of Loss Function

Let ϕ be a parameter in the model Φ. Provided that $\Upsilon(\mathcal{O}, \Phi)$ is differentiable with respect to ϕ, that parameter is adjusted in the gradient decent method according to

$$\hat{\phi} \quad = \quad \phi - \epsilon \frac{\partial \Upsilon(\mathcal{O}, \Phi)}{\partial \phi}, \qquad or$$
$$\hat{\phi} \quad = \quad \phi - \epsilon \underbrace{\Upsilon(\mathcal{O}, \Phi)(\Upsilon(\mathcal{O}, \Phi) - 1)}_{\psi} \frac{\partial d_c(\mathcal{O}, \Phi)}{\partial \phi}. \qquad (4)$$

In Eqn. (4), $\hat{\phi}$ is the new estimate of the parameter and ϵ is a small positive constant which monotonically decreases as the iteration number increases. This gradient descent method is iteratively applied to all training tokens in a sequential manner (for each global regression matrix parameters) to minimize the loss function during the training process. Some intuitive explanations for Eqn. (4) are given here. In the case of near error-free classification, $\Upsilon(\mathcal{O}, \Phi) \approx 0$, In the case of a complete loss (very poor classification), $\Upsilon(\mathcal{O}, \Phi) \approx 1$, the magnitude of ψ in Eqn. (4) would be close to zero and therefore the change of ϕ would become very small. On the other hand, if $\Upsilon(\mathcal{O}, \Phi) \approx 0.5$ (i.e., the likelihoods for the correct and the best wrong model about the same, then the magnitude of ψ would reach a maximum. Therefore, the training procedure as described in Eqn. (4) will focus on input tokens which are likely to be misclassified but can be classified correctly after proper adjustment of the model parameters.

In order to determine $\frac{\partial d_c(\mathcal{O}, \Phi)}{\partial \phi}$ in Eqn. (4), we note that in the trended HMM, each state is characterized by a multivariate Gaussian density function as given in section 2.

Based on the trended model j, the optimal state sequence $\Theta^j = \theta_1^j, \theta_2^j, \cdots, \theta_T^j$ for an input token $\mathcal{O} = \mathcal{O}_1, \mathcal{O}_2, \cdots, \mathcal{O}_T$ (T frames in total) is obtained by means of modified Viterbi algorithm [2]. Then, the log-likelihood is given by

$$g_j(\mathcal{O}, \Phi) \quad = \quad \sum_{t=1}^{T} \log b_{\theta_t^j}(\mathcal{O}_t | \tau_{\theta_t^j}), \qquad (5)$$

which will be used to compute the gradient $\frac{\partial d_c(\mathcal{O},\Phi)}{\partial \phi}$ in Eqn.(4) for global regression matrix parameters in the trended HMM to be described in the remaining part of this section.

3.3. Gradient Formula for Global Regression Matrices

By applying the chain rule results in eqn. (4), the gradient calculation of i-th state parameter $\mathcal{W}_{i,j}(r)$, $r = 0, 1, \cdots, P$, for the j-th model becomes

$$\begin{aligned}
\frac{\partial \Upsilon(\mathcal{O}, \Phi)}{\partial \mathcal{W}_{i,j}(r)} &= \psi \frac{\partial d_c(\mathcal{O}, \Phi)}{\partial \mathcal{W}_{i,j}(r)} \\
&= \psi \frac{\partial}{\partial \mathcal{W}_{i,j}(r)} \left(-g_c(\mathcal{O}, \Phi) + g_\chi(\mathcal{O}, \Phi)\right) \\
&= \psi \frac{\partial}{\partial \mathcal{W}_{i,j}(r)} \left(- \sum_{t=1}^{T} \log b_{\theta_t^c}(\mathcal{O}_t | \tau_{\theta_t^c}) \right. \\
&\quad \left. + \sum_{t=1}^{T} \log b_{\theta_t^\chi}(\mathcal{O}_t | \tau_{\theta_t^\chi}) \right) \\
&= \psi_j \sum_{t \in T_i(j)} \Sigma_{i,j}^{-1} \left[\mathcal{O}_t - \sum_{p=0}^{P} \hat{\mathcal{B}}_{i,j}(p)(t - \tau_i)^p \right] \\
&\qquad [\mathcal{B}_{i,j}(p)]^{Tr} (t - \tau_i)^r
\end{aligned}$$

where the adaptive step size is defined as

$$\psi_j \quad = \quad \begin{cases} \psi & if\ j = c \ \ (correct - class) \\ -\psi & if\ j = \chi \ \ (wrong - class) \\ 0 & otherwise \end{cases}$$

and the set $T_i(j)$ includes all the time indices such that the state index of the state sequence at time t of belongs to state ith in the N-state Markov chain

$$T_i(j) \quad = \quad \{t | \theta_t^j = i\}, \quad 1 \leq i \leq N, \quad 1 \leq t \leq T.$$

To reduce the model complexity as well as to get robust estimates from a small amount of adaptation data, we tied all the state and model dependent transformation matrices $\mathcal{W}_{i,j}(r)$ to a global parameter $\mathcal{W}(r)$ in our experiments. For this special case, the gradient is given by

$$\begin{aligned}
\frac{\partial \Upsilon(\mathcal{O}, \Phi)}{\partial \mathcal{W}(r)} &= \sum_{j=1}^{\mathcal{K}} \psi_j \sum_{i=1}^{N} \sum_{t \in T_i(j)} (t - \tau_i)^r \Sigma_{i,j}^{-1} \\
&\quad \left[\mathcal{O}_t - \sum_{p=0}^{P} \hat{\mathcal{B}}_{i,j}(p)(t - \tau_i)^p \right] [\mathcal{B}_{i,j}(p)]^{Tr}
\end{aligned}$$

The other model parameters are not adopted since the main differences between speakers are assumed to be purely represented by the mean parameters.

4. SPEAKER ADAPTATION EXPERIMENTS

The experiments conducted to evaluate the MCELR approach are aimed at recognizing the 26 letters in the English alphabet, contained in the TI46 speaker dependent isolated word corpus. It is produced by 16 speakers, eight males and eight females. The speaker-independent (SI) training set consists of 26 tokens per word from each of six male and six female speakers. For the remaining four speakers (f1, f2, m1 and m2), up to ten tokens of each word are used as adaptation training data, and the remaining 16 tokens used as speaker dependent test data.

The preprocessor produces a vector of 26 elements consisting of 13 Mel-frequency cepstral coefficients (MFCCs) and 13 delta MFCCs for every 10 msec of speech. The delta MFCCs are constructed by taking the difference between two frame forward and two frame backward of the MFCCs. This window length of 50ms is found to be optimal in capturing the slope of the spectral envelope, i.e. the transitional information [12]. The augmented MFCCs and delta MFCCs are provided as the data input for every frame of speech into the modeling stage. Each word is represented by a single left-to-right, three-state HMM (no skips) with mixture Gaussian state observation densities. The covariance matrices in all the states of all the models are diagonal and are not tied. All transition probabilities are uniformly set to 0.5 (all transitions from a state are considered equally likely) and are not learned during the training process.

The speaker-dependent (ML) models are trained from adaptation data using five-iterations of the modified Viterbi algorithm with single mixture for each state in the HMMs [2]. To set up a baseline speaker-independent (SI) performance on the test data set, we created the ML models, which had been well trained using the SI training set, with a single mixture distribution for each state in the HMMs [10]. For the MCELR approach, the global transformation matrix is initialized by the $d \times d$ identity matrix. Since good initialization of transformation matrices is important to avoid local optimum that would necessarily occur due to the use of gradient descent. Note that the above initialization gives rise to the trended HMM model parameters without adapting the time-varying means. We perform a total of five ML and MCELR iterations and only the best-incorrect-class is used in the misclassification measure. Alphabet classification is performed directly from the modified Viterbi score calculation [2].

The average recognition rates (averaged over two males and two females) are summarized in Table 1 for three experimental setups: 1) benchmark speaker-independent (SI) experiments; 2) speaker-dependent (ML) experiments; 3) speaker-adaptation experiments adapting only polynomial coefficients for the time-varying means (MCELR). These results demonstrate effectiveness of the MCELR training on the trended HMM. Compared with speaker-independent models, the MCELR adaptive training procedure achieves

Number of Adaptation Tokens	Polynomial Order			
	P=0, SI=69.95%		P=1, SI=75.48%	
	ML	MCELR	ML	MCELR
1	58.35%	76.44%	46.82%	79.44%
2	71.15%	78.13%	74.58%	82.69%
3	77.7%	80.29%	82.52%	84.74%

Table 1. Summary of speaker adaptation results.

consistently better performance even with a single token in the adaptation data. The results clearly show that the regular training procedure (ML) is not as good as SI rate when the amount of available training data is limited to one adatation token per word. In the MCELR experiments, the best error rate reduction of 22.58% is obtained when moving from $P = 0$ (80.29%) model to $P = 1$ (84.74%) model with three adaptation takens. The rate drops gradually with fewer adaptation tokens for MCELR experiments. In contrast, for ML experiments, the rate drops rapidly when the training tokens reduce from three to one. The best recognition rate of 84.74% is achieved when polynomial coefficients are adapted using all three tokens of adaptation data.

5. CONCLUSIONS

In this study, the global linear regression based speaker adaptation technique using MCE-based discriminative training paradigm (MCELR) is developed, implemented and evaluated for optimally estimating the time-varying polynomial Gaussian mean functions in the trended HMM. Compared with speaker-independent models, the MCELR adaptive training procedure achieves consistently better performance even with a single token in the adaptation data. An error rate reduction of 61% is achieved when moving from ML to MCELR adaptation scheme in case of linear trended models using a single token in the adaptation data. When three training tokens are used to obtain adaptive estimates for the polynomial coefficients, the recognizer achieves the best recognition rate of 84.74% (averaged over four speakers). We conclude that the time-varying mean parameters in the trended HMM represent the essential characteristics of a particular speaker and can be better estimated with MCELR training approach even with limited amount of training data by using the discriminatively derived global regression matrices. A much more details on experiments with higher order trend polynomial function using MCELR approach is under way and will be reported soon.

REFERENCES

[1] C. M. Ayer, M. J. Hunt and D. M. Brookes, "A Discriminatively Derived Linear Transform for Improved Speech Recognition", *Proc. EUROSPEECH*, Vol. 1, pp. 583-586, Berlin, September, 1993.

[2] L. Deng, M. Aksmanovic, X. Sun and J. Wu, "Speech Recognition Using Hidden Markov Models with Polynomial Regression Functions as Nonstationary States", *IEEE Trans. on Speech and Audio Processing*, Vol.2, No. 4, pp. 507-520, October 1994.

[3] T. Eisele, R. H. Umbach and D. Langmann, "A Comparative Study of Linear Feature Transformation Techniques for Automatic Speech Recognition", *Proc. ICSLP*, Vol. 1, pp. 252-255, Philadelphia, October 1996.

[4] H. Gish and K. Ng, "Parametric Trajectory Models for Speech Recognition," *Proceedings of ICSLP,* Vol. 1, 1996, pp. 466-469.

[5] F. T. Johansen and M. H. Johnsen, "Non-Linear Input Transformations for Discriminative HMMS", *Proc. ICASSP*, Vol. 1, pp. 225-228, Adelaide, 1994.

[6] C. H. Lin, P. C. Chang and C. H. Wu, "An Initial Study on SPeaker Adaptation for Mandarin Syllable Recognition with Minimum Error Discriminative Training", *Proc. ICSLP*, Vol. 1, pp. 307-310, Yokohama, 1994.

[7] C. S. Liu, C. H. Lee, W. Chou, B. H. Juang and A. Rosenberg, "A Study on Minimum Error Discriminative Training for Speaker Recognition", *Journal of Acoustical Society of America*, Vol. 97, No. 1, pp. 637-648, January, 1995.

[8] T. Matsui and S. Furui, "A Study of Speaker Adaptation Based on Minimum Classification Error Training", *Proc. EUROSPEECH*, Vol. 1, pp. 81-84, Madrid, September, 1995.

[9] M. Padmanabhan, L. Bahl, D. Nahamoo and M. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Large-Vocabulary Speech Recognition Systems", *Proc. ICASSP*, Vol. 2, pp. 701-704, Atlanta, May 1996.

[10] C. Rathinavelu and L. Deng, "Speaker Adaptation Experiments Using Nonstationary-State Hidden Markov Models: A MAP Approach", *Proc. ICASSP*, Vol. 2, pp. 1415-1418, Munich, April 1996.

[11] C. Rathinavelu and L. Deng, "The Trended HMM with Discriminative Training for Phonetic Classification", *Proc. ICSLP*, Vol. 2, pp. 1049-1052, Philadelphia, October 1996.

[12] C. Rathinavelu and L. Deng, "HMM-Based Speech Recognition Using State-Dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features", *IEEE Transactions on Speech and Audio Processing*, Vol. 5, N0. 3, pp-243-256, May, 1997.

[13] A. Sankar, L. Neumeyer and M. Weintraub, "An Experimental Study of Acoustic Adaptation Algorithms", *Proc. ICASSP*, Vol. 2, pp. 713-716, Atlanta, May 1996.

[14] J. Takahashi and S. Sagayama, "Minimum Classification Error Training for a Small Amount of Data Enhanced by Vector-Field-Smoothed Bayesian Learning", *Proc. ICASSP*, Vol. 2, pp. 597-600, Atlanta, May 1996.

[15] 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.