MINIMUM CLASSIFICATION ERROR LINEAR REGRESSION FOR ACOUSTIC MODEL ADAPTATION OF CONTINUOUS DENSITY HMMS

Xiaodong He † *and Wu Chou* ‡

† CECS Department, University of Missouri, Columbia, MO 65211
‡ Avaya Labs Research, 233 Mt. Airy Rd., Basking Ridge, NJ 07920
Email: xhb1a@mizzou.edu , wuchou@avaya.com

ABSTRACT

In this paper, a concatenated "super" string model based minimum classification error (MCE) model adaptation approach is described. We show that the error rate minimization in the proposed approach can be formulated into maximizing a special ratio of two positive functions. The proposed string model is used to derive the growth transform based error rate minimization for MCE linear regression (MCELR). It provides an effective solution to apply MCE approach to acoustic model adaptation with sparse data. The proposed MCELR approach is studied and compared with the maximum likelihood linear regression (MLLR) based model adaptation. Experiments on large vocabulary speech recognition tasks are performed. Experimental results indicate that the proposed MCELR model adaptation can lead to significant speech recognition performance improvement and its performance advantage over the MLLR based approach is observed even when the amount of adaptation data is sparse.

1. INTRODUCTION

Minimum classification error (MCE) based discriminative approach has various applications in speech recognition [1,2]. Instead of assuming that the parametric model used in speech recognition characterizes the true distribution of the data, MCE approach is a discriminant function based pattern classification method, and for a given family of discriminant function, optimal classifier/recognizer design involves finding a set of parameters which minimize the empirical recognition error rate. The reason of taking a discriminant function based approach to classifier design is due mainly to the fact that we lack complete knowledge of the form of the data distribution and that training data are always inadequate, particularly in dealing with speech and language problems.

However, minimizing the functional form of the empirical error rate function in MCE based classifier design often presents a great challenge. The most common optimization method used in MCE is based on the generalized probability descent (GPD) algorithm that iteratively adapts the model parameters at an utterance-by-utterance basis [2]. However, there are three major issues in GPD based approach despite its popularity. Firstly, the selection of the step size vector ϵ is empirical and has a critical impact on the model performance. In order to improve the model performance, ϵ needs to be carefully determined. Moreover, different model parameter requires different step size in MCE training. Secondly, the sample-by-sample based parameter adjustments in GPD approach are noisy and the performance of the model fluctuates. Although the performance advantage of MCE is observed in many applications, there is no theoretical guarantee that the selected stopping point in MCE training gives a better model. This is because the benefit of the GPD based optimization is from an asymptotic process. Thirdly, the computational efficiency of GPD algorithm is plagued by the requirement of doing repeated sample-by-sample adjustments.

These issues become critical and acute for MCE based model adaptation when only a small amount of adaptation data is available. If structured adaptation, such as linear regression based model adaptation [3], is considered, there is a huge number of parameters to adapt. These parameters are not real model parameters, but parameters from the "hyperstructure" (e.g. regression matrices). This makes the selection of the step size vector ϵ even more difficult. In [4], the GPD algorithm was directly applied to adapt the linear regression matrices. In this paper, we present a new string model based MCE linear regression (MCELR) approach and derive its growth transform based solution for acoustic model adaptation in large vocabulary speech recognition. The contributions of this paper are:

- A "super" string model based MCELR adaptation approach is described. It utilizes the error correlation between adaptation utterances.
- A growth transform based solution is derived for super string model based MCELR model adaptation.
- The algorithmic effects of the proposed MCELR algorithm are studied and compared with the conventional MLLR based model adaptation. Performance advantage of MCELR is observed on the standard large vocabulary recognition task with a small amount of adaptation data.

2. A FRAMEWORK OF MCE MODEL TRAINING

In string model based MCE approach [2], the classification error count function is represented at the string level model matching and embedded in a smooth loss function

$$L_c(X,\Lambda)=\frac{1}{1+e^{-d_c(X,\Lambda)}}. \quad (1)$$

where $d_c(X, \Lambda)$ is the string level misclassification measure. When N-best competing string models are used,

$$d_c(X,\Lambda) = -\log f(X,W_c;\Lambda) + \log[\frac{1}{N}\sum_{\substack{i=1 \\ W_i \neq W_c}}^{N} \exp[\eta \log f(X,W_i;\Lambda)]]^{1/\eta}. \quad (2)$$

where W_c is the correct transcript lexical word string, and $\{W_i \mid W_i \neq W_c, i = 1,\ldots,N\}$ is the set of N most confusing word strings that are different from W_c. These confusion word strings are typically identified by the recognizer through a N-best search. In conventional MCE training, the GPD algorithm is applied to minimize the expected loss over all training utterances. Each utterance is considered as an independent observation, assuming that there is no correlation between errors in different utterances.

It is known that recognition errors often exhibit a strong correlation with phonetic contexts and are correlated across different utterances. When the amount of adaptation data is small, such correlation should be utilized in model adaptation. To improve the effect of MCE based model adaptation, we introduce a "super string" based string model. The super string X in our approach is constructed by concatenating the limited adaptation utterances into one string. The string model based MCE training becomes to minimize the loss function $L_c(X,\Lambda)$ of the super string X, with the added constraint that the word sequence content of each utterance is aligned within its original start/end boundaries.

In the "super" string model framework, we consider

$$P(\Lambda) = 1 - L_c(X,\Lambda). \quad (3)$$

It is obvious that minimizing $L_c(X,\Lambda)$ is equivalent to maximizing

$$P(\Lambda) = \frac{N^{1/\eta} f(X,W_c;\Lambda)}{[\sum_{i=1}^{N} f(X,W_i;\Lambda)^{\eta}]^{1/\eta} + N^{1/\eta} f(X,W_c;\Lambda)}. \quad (4)$$

If we set the smooth factor $\eta = 1$, it simplifies to

$$P(\Lambda) = \frac{N \cdot f(X,W_c;\Lambda)}{\sum_{i=1}^{N} f(X,W_i;\Lambda) + N \cdot f(X,W_c;\Lambda)}. \quad (5)$$

However, $P(\Lambda)$ is a complicated ratio of two positive functions. We sketch the main steps that are used to derive the growth transform solution for optimizing $P(\Lambda)$ in MCE based model adaptation.

$P(\Lambda)$ is the ratio of $\frac{G(\Lambda)}{H(\Lambda)}$, where

$$G(\Lambda) = N \cdot f(X,W_c;\Lambda), \quad (6)$$

$$H(\Lambda) = \sum_{i=1}^{N} f(X,W_i;\Lambda) + N \cdot f(X,W_c;\Lambda) \cdot \quad (7)$$

Then a function can be constructed as follows

$$F(\Lambda;\Lambda') = G(\Lambda) - P(\Lambda')H(\Lambda) + D, \quad (8)$$

with D a suitable positive constant. The important property of $F(\Lambda;\Lambda')$ is that, if $F(\Lambda;\Lambda') \geq F(\Lambda';\Lambda')$, then $P(\Lambda) \geq P(\Lambda')$ [5]. Furthermore, if $F(\Lambda;\Lambda')$ can be represented in the form

$$F(\Lambda;\Lambda') = \sum_s \int_{\chi} h(\chi,s,\Lambda) d\chi, \quad (9)$$

increasing the value of $F(\Lambda;\Lambda')$ can be achieved by maximizing

$$\sum_s \int_{\chi} h(\chi,s,\Lambda') \log h(\chi,s,\Lambda) d\chi, \quad (10)$$

where $h(\chi, s, \Lambda)$ is a positive function [6]. For super string model based MCE adaptation,

$$h(\chi,s,\Lambda) = [\Gamma(\Lambda') + d(s)] \cdot f(\chi \mid s;\Lambda) \quad (11)$$

and

$$\Gamma(\Lambda') = 1_{\chi}(X)\left[N \cdot \frac{f(s,W_c)\sum_{i=1}^{N} f(\chi,W_i;\Lambda') - f(\chi,W_c;\Lambda')\sum_{i=1}^{N} f(s,W_i)}{\sum_{i=1}^{N} f(\chi,W_i;\Lambda') + N \cdot f(\chi,W_c;\Lambda')}\right],$$

where $1_{\chi}(X)$ is the indicator function of X, and s is the hidden Gaussian component sequence. If we only adapt mean vectors and covariance matrices of the acoustic model and denote Λ as those parameters, (W, s) is independent from Λ. Moreover, we have $f(X \mid s) f(W, s) = f(X \mid W, s) f(W, s) = f(X, W, s)$ for arbitrary word string W. The constant D in (8) is determined by $D = \sum_s d(s)$, where $d(s)$ for each s is chosen to guarantee that $h(\chi, s, \Lambda)$ is positive.

Since $[\Gamma(\Lambda') + d(s)]$ is not a function of Λ, the growth transform is the one that maximizes

$$V(\Lambda) = \sum_s \int_{\chi} [\Gamma(\Lambda') + d(s)] f(\chi \mid s;\Lambda') \log f(\chi \mid s;\Lambda) d\chi. \quad (12)$$

Divide through (12) by $f(X, W_c; \Lambda')$, for continuous density HMMs, the maximizing objective function is as follows

$$U(\Lambda) = \sum_{t,m} [\Delta\gamma(t,m)] \log f(x_t \mid s_t = m;\Lambda) + \sum_{t,m} d'(t,m) \int_{\chi_t} f(\chi_t \mid s_t = m;\Lambda') \log f(x_t \mid s_t = m;\Lambda) d\chi_t, \quad (13)$$

where

$$\Delta\gamma(t,m) = \frac{N \cdot \sum_{i=1}^{N} f(X,W_i;\Lambda')[\gamma(t,m,W_c) - \gamma(t,m,W_i)]}{\sum_{i=1}^{N} f(X,W_i;\Lambda') + N \cdot f(X,W_c;\Lambda')},$$

with $\gamma(t, m, W) = p(s_t = m \mid X, W; \Lambda')$ is the *a posteriori* probability of occupying the Gaussian component m at time t, given data X and a referenced word string W, and $d'(t, m)$ is computed by $d'(t,m) = \sum_{s,s_t=m} d(s) / f(X,W_c;\Lambda') \cdot$

3. MCELR MODEL ADAPTATION

In the linear regression based model adaptation framework, usually all Gaussian components of the acoustic model are clustered into several regression classes through a regression tree [3]. For class m with R Gaussian components $\{\lambda_{mr} \mid r = 1,\ldots,R\}$, a transform matrix W_m is estimated. Then for the m_r-th Gaussian component $N(\mu_{mr}, \Sigma_{mr})$, the adapted mean vector is given by:

$$\hat{\mu}_{mr} = W_m \cdot \xi_{mr} \quad (14)$$

where $\xi_{mr} = [1, \mu_{mr}(1),\ldots,\mu_{mr}(D)]^T$ is the extended vector of the D-dimensional mean vector μ_{mr}. In MLLR based model adaptation, W_m is estimated based on the maximum likelihood (ML) criterion. In MCELR based approach, the MCE criterion is used for W_m estimation.

In this paper, we adopt the same notation used to derive MLLR in [3] for the purpose of easy comparison. For simplification, the subscript of class m is omitted in following equations.

Denoting $h(x,r) = (x - W\xi_r)^T \Sigma_r^{-1}(x - W\xi_r)$, and ignoring terms that are irrelevant to maximization, the maximizing objective function has the following form

$$Q(W) = \sum_t \sum_r [\Delta\gamma(t,r) h(x_t,r)] + \sum_t \sum_r d'(t,r) \int_{\chi_t} f(\chi_t \mid s_t = r, \Lambda') h(\chi_t, r) d\chi_t. \quad (15)$$

Set $\partial Q(W)/\partial W = 0$, and notice that

$\int_{\chi_t} f(\chi_t \mid s_t = r;\Lambda')d\chi_t = 1$, $\int_{\chi_t} \chi_t f(\chi_t \mid s_t = r;\Lambda')d\chi_t = \mu_r$,

W can be solved through the following equation,

$$\sum_r \Sigma_r^{-1}[\sum_t \Delta\gamma(t,r)x_t + D_r\mu_r]\xi_r^T = \sum_r [\sum_t \Delta\gamma(t,r) + D_r]\Sigma_r^{-1} W \xi_r \xi_r^T, \quad (16)$$

where $D_r = \sum_t d'(t,r)$. Follows the notation in [3], we denote

$$\tilde{V}^{(r)} = [\sum_t \Delta\gamma(t,r) + D_r] \cdot \Sigma_r^{-1},$$

$$\tilde{Z} = \sum_r \Sigma_r^{-1}[\sum_t \Delta\gamma(t,r)x_t + D_r\mu_r] \cdot \xi_r^T.$$

When diagonal covariance matrices are used, $\tilde{V}^{(r)}$ is a diagonal matrix, and W can be computed on a row-by-row basis,

$$w_i^T = \tilde{G}^{(i)-1} \cdot \tilde{z}_i^T, \quad (17)$$

where w_i and $\tilde{z}_i$ are the i-th rows of W and $\tilde{Z}$, respectively. $G^{(i)}$ is a $(D+1)\times(D+1)$ matrix that is computed by $g_{j,q}^{(i)} = \sum_r \tilde{v}_{i,i}^{(r)} d_{j,q}^{(r)}$, with $D^{(r)} = \xi_r \xi_r^T$, where the individual matrix elements at the j^{th} row and the k^{th} column of $G^{(i)}$, $\tilde{V}^{(r)}$ and $D^{(r)}$ are denoted by $g_{j,k}^{(i)}$, $\tilde{v}_{j,k}^{(r)}$ and $d_{j,k}^{(r)}$, respectively.

4. EXPERIMENTS

4.1. Experimental condition

The speech recognition experiments were performed on the Wall Street Journal (WSJ) speaker adaptation task using the official 1993 Spoke 3 speaker adaptation and evaluation data (ET_S3). The data set includes 10 speakers, each of which provides 40 utterances for adaptation and other 40~43 utterances for testing. The standard 5K trigram language model specified for the evaluation was used. The speech feature vector is MFCC based with 39 dimensions (c, Δc, $\Delta\Delta c$, e, Δe, $\Delta\Delta e$). The speaker independent (SI) model was trained on the standard speaker independent WSJ SI-84 portion of the training corpus. Crossword triphones were used as the recognition units and the baseline SI model was obtained by using phonetic decision tree based state tying. For the baseline system, an average word error rate (WER) of 27.5% was achieved over these 10 speakers.

In our experiments, 1-best competing string model based MCE approach was implemented. Correspondingly, $\Delta\gamma(t, r)$ is

$$\Delta\gamma(t,r) = \frac{f(X,W_e;\Lambda')[\gamma(t,r,W_c) - \gamma(t,r,W_e)]}{f(X,W_e;\Lambda') + f(X,W_c;\Lambda')}, \quad (18)$$

where W_e is the most confusing string that is different from W_c.

In the 1-best competing string model MCE approach, a large portion of W_c and W_e are the same, except those words that correspond to recognition errors. Furthermore, referring to (18), many data are "neutralized" except those "effective data" which correspond to the confusing error words between W_c and W_e. Correspondingly, in MCELR, the criterion to estimate a transform matrix for a regression class should be based on the adequate "effective data" that are accumulated in that class.

The constant D_r in (16) is a factor to control the "learning rate". As suggested in MMI training [7], for the r-th Gaussian mixture, D_r is given as

$$D_r = \tau + E \cdot \sum_t \gamma(t,r,W_e), \quad (19)$$

where E is a global smoothing factor to scale the value of D_r, and τ is a small constant to make sure D_r is always positive. In our experiments, τ was always set to 2, and the WSJ 20K trigram language model was used to generate the competitor W_e.

4.2. Experimental result

Two sets of experiments were conducted to evaluate the proposed MCELR based model adaptation method. Firstly, the MLLR based mean adaptation was performed. Secondly, by using the MLLR adapted model as the seed model, a series of MCELR adaptation experiments were performed. In our experiments, the sample count threshold of generating a transform matrix in MLLR was set to 1000, and the "effective data" amount threshold of generating a transform matrix in MCELR was set to 100. In adaptation, the silence model was not adapted. Furthermore, we found that a better performance could be obtained if the value of $f(X, W_c; \Lambda')$ is decreased slightly by a factor F, where $f'(X, W_c; \Lambda') = f(X, W_c; \Lambda')^F$, and F was set to 1.003 in the following experiments.

4.2.1. Objective function optimization

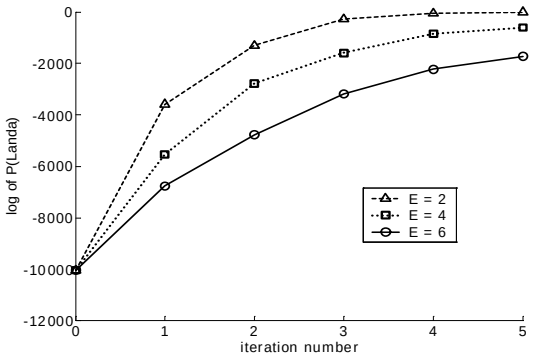

Fig. 1. The value of log[$P(\Lambda)$] vs. iteration number, given different value of the global smoothing factor E.

The change in MCE objective function $P(\Lambda)$ as adaptation iteration proceeds is shown in Figure 1. The effects of the global smoothing factor with $E = 2$, $E = 4$, and $E = 6$ are evaluated. The monotone increasing of the value of $P(\Lambda)$ indicates that the decreasing of empirical loss function $L_c(X,\Lambda)$ of MCE is achieved as adaptation iteration continues. Furthermore, compared with a larger value of E, a smaller value of E leads to a smaller constant D_r, and results to a faster "learning rate".

4.2.2. MCELR performances on training and testing set

The proposed MCELR approach is evaluated on both of adaptation and testing data. The 20K language model is used in decoding the adaptation utterances. It is desirable to update the competitor when the model is updated. In experiments, a new competitor is generated after every three iterations.

The recognition performances on adaptation and testing data are shown in Figure 2 and 3, respectively. As we expect, the recognition error rate on the training adaptation data set drops sharply, with a relative error reduction around 60%. However, this dramatic improvement is not maintained on the testing set, on which a 6.2% error reduction is achieved after 12 iterations, when $E = 6$. The effects of the global smoothing factor E on recognition performance are also evaluated. As illustrated in Figure 3, a small value of $E = 2$, which corresponds to the fastest learning rate, leads to an unstable performance on testing set. On the other hand, although the best result is obtained by using a larger value of $E = 6$, much more iterations are needed for it.

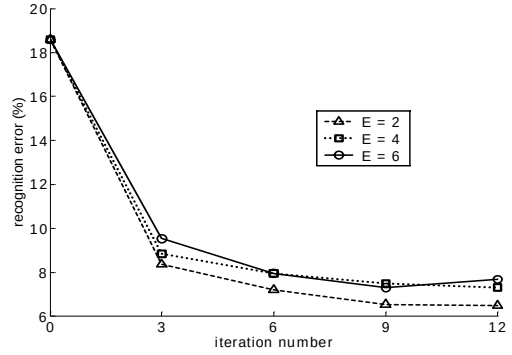

Fig. 2, recognition error rate vs. iteration number on adaptation set, given different value of the global smoothing factor E.

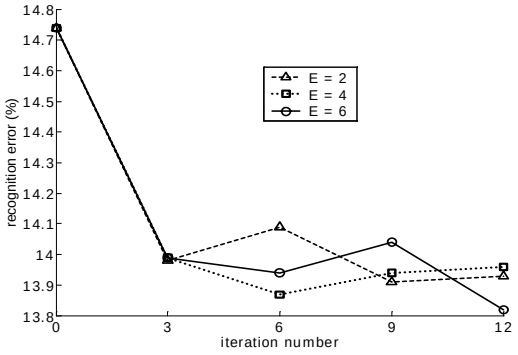

Fig. 3, recognition error rate vs. iteration number on testing set, given different value of the global smoothing factor E.

4.2.3.Comparison of MCELR and MLLR approaches

Table 1 summarizes the performance comparison of the MLLR based model adaptation and the MCELR approach, with respect to the amount of adaptation data (in adaptation utterance). In all MCELR adaptation experiments, the seed models are the corresponding MLLR adapted models. The global smoothing factor E is set to 6 for tests with 10 and 20 adaptation utterances, and set to 4 for tests with 30 and 40 utterances. The iteration number of MCELR based model adaptation is fixed to 6. Compared with MLLR approach, MCELR based adaptation can further reduce the recognition WER by 5.0% ~ 7.7%, relatively.

TABLE I, Recognition performance (WER %) comparison of the MLLR based adaptation and the MCELR based adaptation.

# Adpt. utter.	10	20	30	40
MLLR	19.31	16.88	15.56	14.74
MCELR	17.82	15.75	14.78	13.87
Err. reduction	-7.7%	-6.7%	-5.0%	-5.9%

5. SUMMARY

In this paper, a "super" string model based minimum classification error linear regression (MCELR) adaptation approach was described. It was shown that the error rate minimization in the proposed approach could be formulated into maximizing a special ratio of two positive functions. Furthermore, a growth transform based estimation of the MCE linear regression transform matrix has been derived. It provides an effective solution to apply MCE approach to acoustic model adaptation with sparse data. The implementation details were studied and experimental results on the 1993 Spoke 3 test set of the WSJ task show that significant performance advantage over the MLLR based approach was achieved even when the amount of adaptation data is sparse.

6. REFERENCES

[1] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. Speech Audio Proc.*, vol. 5, May 1997.

[2] W. Chou, "Discriminant - Function - Based Minimum Recognition Error Rate Pattern - Recognition Approach to Speech Recognition," *Proceedings of the IEEE*. Vol. 88, No.8, pp.1201 – 1223. August 2000.

[3] 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, April 1995.

[4] J. Wu and Q. Huo, "Supervised Adaptation of MCE-Trained CDHMMs Using Minimum Classification Error Linear Regression," *ICASSP'02*, pp. 605 – 608, May 2002.

[5] A. Gunawardana and W. Byrne, "Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression," *Proc. EuroSpeech'01*, September 2001.

[6] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems," *IEEE Trans. Inf. Thry.*, Vol 37, pp.107 – 113, January. 1991.

[7] P. C. Woodland and D. Povey, "Large Scale Discriminative Training for Speech Recognition," *Proc. ITRW ASR*, ISCA, 2000.