

MINIMUM CLASSIFICATION ERROR (MCE) 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 framework of minimum classification error (MCE) model adaptation for continuous density HMMs is proposed based on the approach of "super" string model. We show that the error rate minimization in the proposed approach can be formulated into maximizing a special ratio of two positive functions, and from that a general growth transform algorithm is derived for MCE based model adaptation. This algorithm departs from the generalized probability descent (GPD) algorithm, and it is well suited for model adaptation with a small amount of training data. The proposed approach is applied to linear regression based variance adaptation, and the close form solution for variance adaptation using MCE linear regression (MCELR) is derived. The MCELR approach is evaluated on large vocabulary speech recognition tasks. The relative performance gain is more than doubled on the standard (WSJ Spoke 3) database, comparing to maximum likelihood linear regression (MLLR) based variance adaptation for the same amount of adaptation data.

1. INTRODUCTION

Model adaptation is critical for speech recognition in adverse environment, and it is an active research area in the past ten years. Minimum classification error (MCE) based discriminative approach is effective in acoustic model training and has found various applications in speech recognition [1] [2]. The MCE based classifier/recognizer design involves finding a set of parameters that 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 is 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]. The optimality of this algorithm is its asymptotic convergence property, making the algorithm more suitable for acoustic model training with sufficient data, not for model adaptation where usually only a small amount of adaptation data is available. Another problem in GPD based MCE approach is the selection of the step size vector ϵ . In order to improve the model performance, ϵ needs to be carefully determined, and different model parameter requires different step size. This process is empirical and has a critical impact on the model performance. Moreover, in model adaptation, many parameters to be estimated are so-called hyper-parameters [3] [4]. These parameters are not real model parameters, but parameters from the added "hyper-structure", which generally do not have a strong physically meaningful interpretation. This makes the determination of ϵ in GPD approach even more difficult. Because of the above-mentioned problems, MCE approach is rarely used in model adaptation with sparse adaptation data.

In this paper, we develop an efficient MCE model adaptation framework based on the concept of "super string" models and establish a growth transform based minimization framework for minimizing the loss function in MCE adaptation. This approach has a monotonic loss minimization property, which is critical for model adaptation when only a small amount of adaptation data is available, and when there are lots of hyper-parameters to estimate (e.g. regression matrices). We derived in [5] the growth-transform solution for mean vector adaptation in MCE linear regression (MCELR). In this paper, it is shown that MCELR variance adaptation also has a growth-transform based solution under the proposed framework.

2. A "SUPER STRING" MODEL BASED MCE MODEL ADAPTATION FRAMEWORK

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,

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

where W_c is the correct transcript lexical word string, and $\{W_i | W_i \neq W_c, i = 1, \dots, N\}$ is the set of N most confusing word strings that are different from W_c . These confusion word strings are typically identified by a 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, a "super string" based string model is introduced. 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.

3. FORMULATION OF GROWTH TRANSFORM IN MCE MODEL ADAPTATION FRAMEWORK

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

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

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

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 $P(\Lambda) \geq P(\Lambda')$ [6]. Furthermore, if $F(\Lambda; \Lambda')$ can be represented in the form

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

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

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

where $h(\chi, s, \Lambda)$ is a positive function [7], and the integration domain χ is a space with $P \times T$ dimensions, given P is the feature dimension and T is the number of data frames. For super string model based MCE approach,

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

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 , s is the state mixture sequence, and Λ is the set of parameters for mean vectors and covariance matrices in Gaussian observation densities of all HMMs. From HMM structure, s is formed by Markov chain and Gaussian mixture weights, and is independent of Λ . So $f(X, s, W | \Lambda) = f(X | s, W, \Lambda) f(s, W | \Lambda) = f(X | s, \Lambda) f(s, W)$ 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

$$V(\Lambda) = \sum_{s, \chi} \int [\Gamma(\Lambda') + d(s)] f(\chi | s, \Lambda') \log f(\chi | s, \Lambda) d\chi. \quad (12)$$

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

$$U(\Lambda) = \sum_{t,r} [\Delta\gamma(t,r)] \log f(x_t | s_t = r, \Lambda) \quad (13)$$

$$+ \sum_{t,r} d'(t,r) \int_{\mathcal{X}_t} f(\mathcal{X}_t | s_t = r, \Lambda') \log f(\mathcal{X}_t | s_t = r, \Lambda) d\mathcal{X}_t,$$

where \mathcal{X}_t is a P-dimensional space, and

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

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

Eq. (13) is the EM formulations of the growth transform solution for MCE model adaptation. It applies to all model parameters or model parameter transforms in Λ . As an application of Eq. (13), we derive the growth transform for MCELR based variance transformation in the next section.

4. MCELR VARIANCE ADAPTATION

In continuous density HMMs with mixture Gaussian densities, the Gaussian component is characterized by its mean and covariance matrix and denoted generically as $N(\mu_r, \Sigma_r)$. The covariance matrix Σ_r is a positive definite matrix that can be represented in the following form:

$$\Sigma_r = B_r^T B_r = \Sigma_r^{\frac{1}{2}} \Sigma_r^{\frac{1}{2}}, \quad (14)$$

where $B_r = C_r^{-1}$ and $\Sigma_r^{-1} = C_r C_r^T$. In the linear regression based model adaptation framework, all Gaussian components of the acoustic model are clustered into several regression classes through a regression tree [4]. For class m with R Gaussian components $\{\lambda_{m,r} | r = 1, \dots, R\}$, a transform matrix H_m is estimated. Then for the m_r -th Gaussian component $N(\mu_{m,r}, \Sigma_{m,r})$, the adapted covariance is given by:

$$\hat{\Sigma}_{m,r} = B_{m,r}^T H_m B_{m,r}. \quad (15)$$

In MLLR based model adaptation, H_m is estimated based on the maximum likelihood (ML) criterion. In MCELR based approach, the MCE criterion is used for H_m estimation. In the following derivation, the subscript m is omitted for simplification.

To estimate the variance transformation matrix H based on the MCE criterion, the optimization object function (13) becomes:

$$U(\Lambda) = \sum_{t,r} \Delta\gamma(t,r) Q(x_t, r, \Lambda) \quad (16)$$

$$+ \sum_{t,r} d'(t,r) \int_{\mathcal{X}_t} f(\mathcal{X}_t | s_t = r, \Lambda') Q(\mathcal{X}_t, r, \Lambda) d\mathcal{X}_t,$$

where $Q(x_t, r, \Lambda) = [\ln |H| + (x_t - \mu_r)^T C_r H^{-1} C_r^T (x_t - \mu_r)]$.

Set $\partial U(\Lambda) / \partial H = 0$, notice that $\int_{\mathcal{X}_t} f(\mathcal{X}_t | s_t = r, \Lambda') d\mathcal{X}_t = 1$ and $\int_{\mathcal{X}_t} f(\mathcal{X}_t | s_t = r, \Lambda') [(\mathcal{X}_t - \mu_r)(\mathcal{X}_t - \mu_r)^T] d\mathcal{X}_t = \Sigma_r$, H can be solved through the following equation,

$$H^T \left[\sum_t \sum_r \Delta\gamma(t,r) + \sum_r D_r \right] - \sum_t \sum_r [d'(t,r) \cdot C_r^T \cdot \Sigma_r \cdot C_r] \quad (17)$$

$$= \sum_t \sum_r [\Delta\gamma(t,r) (C_r^T x_t - C_r^T \mu_r) (C_r^T x_t - C_r^T \mu_r)^T]$$

where $D_r = \sum_t d'(t,r)$.

Finally, H is adapted as follows:

$$H = \frac{\sum_t \sum_r [\Delta\gamma(t,r) (C_r^T x_t - C_r^T \mu_r) (C_r^T x_t - C_r^T \mu_r)^T] + \sum_r D_r \cdot I}{\left[\sum_t \sum_r \Delta\gamma(t,r) + \sum_r D_r \right]} \quad (18)$$

5. EXPERIMENTS

5.1. Experimental Conditions

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 ($c, \Delta c, \Delta \Delta c, e, \Delta e, \Delta \Delta e$). 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 super-string-model based MCE approach was implemented. Correspondingly, $\Delta\gamma(t, r)$ is

$$\Delta\gamma(t,r) = \frac{f(X, W_e | \Lambda') [\gamma(t,r, W_c) - \gamma(t,r, W_e)]}{f(X, W_e | \Lambda') + f(X, W_c | \Lambda')} \quad (19)$$

where W_e is the most confusing string different from W_c .

In the 1-best competing super-string-model MCE approach, most parts of W_c and W_e are the same, except those words that correspond to recognition errors. Furthermore, referring to (19), 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 amount of “effective data” accumulated in the class.

The constant D_r in (18) is a factor to control the “learning rate”. As suggested in MMI training [8], for the r -th Gaussian mixture, D_r is given as

$$D_r = \tau + E \cdot \sum_t \gamma(t, r, W_e), \quad (20)$$

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, E was set to 4, τ was set to 2, and a 60K trigram language model was used to generate the competitor W_e .

Two sets of experiments were conducted to evaluate the proposed MCELR based variance adaptation method. Firstly, the conventional MLLR based mean and variance adaptations were evaluated. Secondly, a series of experiments of MCELR variance adaptation plus MLLR mean adaptation 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. The seed model for MCELR variance adaptation is the SI model. In adaptation, diagonal transformation matrices are used for variance adaptation, and the silence model is not adapted.

5.2. Experimental Results

Table 1 evaluated the additional gain of doing MLLR adaptation on both mean and variance parameters, compared with doing MLLR mean adaptation only. As illustrated in the table, the MLLR based mean+variance adaptation only provides a slight performance improvement over the MLLR based mean-only adaptation, which is consistent with results reported in other studies [4]. The relative error rate reduction of MLLR variance adaptation is around 2.8%, which is averaged over all adaptation conditions.

TABLE I: WORD ERROR RATES OF MLLR BASED MODEL ADAPTATION METHODS (%)

# Adpt. utter.	10	20	30	40
MLLR mean only	19.31	16.88	15.56	14.74
MLLR vari+mean	18.48	16.55	15.11	14.45

The performance of the proposed MCELR variance adaptation was evaluated and the results were tabulated in Table 2. In our experiments, the iteration number of MCELR based variance adaptation was set to six, and a new competitor was generated after every other iteration. Compared with MLLR mean-only adaptation, a further error rate reduction of about 6.2% is achieved by an additional MCELR variance adaptation. Moreover, compared with the conventional MLLR variance adaptation, the performance gain resulted by variance

adaptation is more than doubled (from 2.8% to 6.2%) by the proposed MCELR variance adaptation approach.

TABLE II: WORD ERROR RATES OF MLLR BASED MEAN ADAPTATION AND MCELR BASED VARIANCE ADAPTATION PLUS MLLR BASED MEAN ADAPTATION (%)

# Adpt. utter.	10	20	30	40
MLLR mean only	19.31	16.88	15.56	14.74
MCELR vari + MLLR mean	18.53	15.85	14.40	13.71

6. SUMMARY

In this paper, a general framework of “super” string model based minimum classification error (MCE) model adaptation for continuous density HMMs was presented. It was shown that the error rate minimization in the proposed approach could be formulated into maximizing a special ratio of two positive functions. The proposed MCE model adaptation was applied to variance adaptation based on the principle of MCE linear regression (MCELR). A growth transform algorithm was derived for MCELR based variance adaptation. Experimental results on 1993 WSJ Spoke 3 speaker adaptation task indicated that comparing to MLLR, the average relative performance gain of variance adaptation by MCELR was more than doubled, under the same test condition and using the same amount of adaptation data.

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] M. J. F. Gales and P. C. Woodland, “Mean and Variance Adaptation Within the MLLR Framework,” *Computer Speech and Language*, Vol. 10, pp. 249-264. 1996.
- [5] X. He and W. Chou, “Minimum Classification Error Linear Regression for Acoustic Model Adaptation of Continuous Density HMMs,” *ICASSP’03*, April 2003.
- [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] A. Gunawardana and W. Byrne, “Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression,” *Proc. EuroSpeech’01*, September 2001.
- [8] P. C. Woodland and D. Povey, “Large Scale Discriminative Training for Speech Recognition,” *Proc. ITRW ASR, ISCA*, 2000.