A Novel Learning Method for Hidden Markov Models in Speech and Audio Processing

Xiaodong He and Li Deng
Microsoft Research
1 Microsoft Way
Redmond, WA 98052, USA
{xiaohe|deng}@microsoft.com

Wu Chou
Avaya Labs Research
233 Mt. Airy Rd
Basking Ridge, NJ 07920, USA
wuchou@avaya.com

***Abstract* — in recent years, various discriminative learning techniques for HMMs have consistently yielded significant benefits in speech recognition. In this paper, we present a novel optimization technique using the Minimum Classification Error (MCE) criterion to optimize the HMM parameters. Unlike Maximum Mutual Information training where an Extended Baum-Welch (EBW) algorithm exists to optimize its objective function, for MCE training the original EBW algorithm cannot be directly applied. In this work, we extend the original EBW algorithm and derive a novel method for MCE-based model parameter estimation. Compared with conventional gradient descent methods for MCE learning, the proposed method gives a solid theoretical basis, stable convergence, and it is well suited for the large-scale batch-mode training process essential in large-scale speech recognition and other pattern recognition applications. Evaluation experiments, including model training and speech recognition, are reported on both a small vocabulary task (TI-Digits) and a large vocabulary task (WSJ), where the effectiveness of the proposed method is demonstrated. We expect new future applications and success of this novel learning method in general pattern recognition and multimedia processing, in addition to speech and audio processing applications we present in this paper.**

Keywords—Speech recognition and audio processing, pattern recogntion, machine learning, discriminative learning, hidden Markov model, rational-function optimization, growth transformation, extended Baum-Welch algorithm

Topic area—Multimedia processing (speech and audio).

I. INTRODUCTION

In the history of speech recognition, discriminative training has been applied to and analyzed for many state-of-the-art Hidden Markov Model (HMM) [15] based systems using different training criteria. These include Maximum Mutual Information (MMI) [1][4][10][12], Minimum Classification Error (MCE) [2][3][6][7][8][9][14][16], and Minimum Phone Error (MPE) [11][13][3]. In these systems, a crucial component is the optimization method for model parameter estimation given the discriminative training criterion.

In this work, we employ the MCE criterion to optimize the HMMs as the acoustic models for speech recognition. The essence of MCE is to define the objective function for optimization that is closely related to classification errors. This is more desirable than other types of discriminative training that are less relevant to classification errors.

In the context of MCE, the conventional optimization method has been the sequential, sample-by-sample gradient descent technique called Generalized Probabilistic Descent (GPD) [2][7][8]. However, GPD requires precise tuning of the learning rate, and, due to its sequential processing nature, cannot be parallelized over multiple processors. Therefore, applying GPD to large vocabulary speech recognition tasks is very difficult. Batch-mode optimization methods including Batch-mode PD, QuickProp, Rprop, and partial BFGS have been explored in [16]. Although performance improvements were reported on larger vocabulary tasks, they still require careful, empirical tuning of learning parameters for stable convergence and they tend to be slow. It is highly desirable to develop new optimization methods that have a more solid theoretical basis, stable convergence, and are easy to parallelize.

In [13], an optimization method based on the weak sense auxiliary function (WSAF) was proposed. With this method, a WSAF is constructed first. Then, the extend Baum-Welch (EBW) algorithm for HMM parameter re-estimation is derived to optimize the WSAF so as to optimize the targeting objective function. The theoretical basis of the WSAF method, however, is weak. This is because the WSAF is a function which only has the same gradient as the targeting objective function at a local point in the model space as detailed in [13]. Hence, optimizing the WSAF cannot guarantee optimization of the targeting objective function (MPE as discussed in [13]).

In [4], the EBW method was originally developed for optimizing the MMI criterion associated with a discrete HMM. In this method, "growth transformation" as the re-estimation formulas is established for optimizing a special rational function that is directly related to the MMI objective function. The elegant theoretical development in [4] has proved that the growth transformation on the HMM parameters guarantee a non-decrease of the MMI objective function. The EBW method was extended from the discrete HMM to the continuous HMM in [10][5]. This EBW method is theoretically well founded for the MMI criterion, and it has been successfully applied to large vocabulary tasks [12]. However, applying the EBW method to the MCE criterion is very difficult. This is because EBW is suited for optimizing a single rational objective function such as MMI, but the MCE objective function is a sum of multiple rational functions as shown in section II.B.

In this work, we show that the problem of minimizing a sum of rational functions for MCE can be mapped to a problem of maximizing a specially constructed single rational function. Then the original EBW algorithm is extended to construct a new growth transformation for optimizing the MCE objective function for the HMM parameters. This new learning technique gives a solid theoretical basis for stable and monotonic convergence for the algorithm, and is well suited for large-scale batch-mode HMM training.

The organization of this paper is as follows. In section 2, we present our proposed method of optimizing the MCE loss function using growth transformation or EBW, moving away from the traditional GPD method. Detailed theoretical development is presented, including re-formulation of the MCE objection function into a novel form of a rational function suitable for the application of the EBW algorithm. This section also includes major steps in the EBW derivation of the HMM parameter estimation formulas. In section 4, experimental results on TI-Digits and Wall Street Journal (WSJ) tasks are presented to demonstrate the effectiveness of the proposed method.

II. Proposed Method

A. Initial Construction of the MCE Objective Function

MCE learning was originally introduced for multiple-category classification problems where the smoothed error rate is minimized for isolated "tokens"[7]. It was later generalized to minimize the smoothed "sentence token" or string-level error rate [2][8], which is known as "embedded MCE". The MCE objective function is defined first based on a set of discriminant functions and a special type of loss function. Then the model is estimated to minimize the expected loss that is closely related to the recognition error rate of the classifier.

In the following sections, we denote by Λ the HMM parameter set that needs to be estimated. Let r=1,…, R be the index for multiple training utterances. Each utterance consists of an acoustic observation vector sequence X_r=$x_{r,1}$, …, $x_{r,Tr}$, which is computed from the audio signal of the speech waveform, and the corresponding correctly labeled word string S_r = $w_{r,1}$,…, $w_{r,Nr}$. Further, we use s_r to represent all possible label sequences (strings) for the r-th utterance, including the correct label sequence S_r and all other incorrect label sequences.

At embedded MCE training, first a set of discriminant functions is defined based on the correct string S_r and other competing strings,

$$g_{s_r}(X_r;\Lambda) = \log\ p(X_r, s_r; \Lambda)\ . \tag{1}$$

Then the decision rule for the speech recognizer is the one that for the observation data sequence X_r,

$$C(X_r) = s_r^* \qquad iff\ \ s_r^* = \arg\max_{s_r} g_{s_r}(X_r;\Lambda)\ . \tag{2}$$

For continuous speech recognition, usually only the N most confusable competing strings, $s_{r,1}$,…, $s_{r,N}$, are considered in (2). They are defined inductively as follows.

$$\begin{aligned} s_{r,1} &= \arg\max_{s_r : s_r \neq S_r}\ \log p(X_r, s_r; \Lambda) \\ s_{r,i} &= \arg\max_{s_r : s_r \neq S_r,\, s_r \neq s_{r,1},\ldots,s_{r,i-1}} \log p(X_r, s_r; \Lambda)\ \ i=2,\ldots,N. \end{aligned} \tag{3}$$

Unlike the multiple-category classification problems, these N confusable competing strings change dynamically after every estimation iteration. In practice, they are re-generated at every iteration through an N-best decoding, based on the model obtained at the immediate previous iteration.

Next, a misclassification measure $d_r(X_r,\Lambda)$ is defined to emulate the decision rule for utterance r, i.e., $d_r(X_r,\Lambda)>0$ implies misclassification and $d_r(X_r,\Lambda)<0$ implies a correct classification,

$$d_r(X_r,\Lambda) = -g_{S_r}(X_r;\Lambda) + G_{S_r}(X_r;\Lambda)\ . \tag{4}$$

where $G_{S_r}(X_r;\Lambda)$ counts for the score of competitors competing with the correct string S_r. It is usually a soft-max of the discriminant functions of the N most confusable competing strings,

$$G_{S_r}(X_r;\Lambda) = \log\ \left\{ \frac{1}{N}\sum_{i=1}^{N} \exp\ \left[\eta\ g_{s_{r,i}}(X_r;\Lambda) \right] \right\}^{\frac{1}{\eta}} \tag{5}$$

For the widely used 1-best MCE training [17], only the best-incorrect-string is considered, and $G_{S_r}(X_r;\Lambda)$ is:

$$G_{S_r}(X_r;\Lambda) = g_{s_{r,1}}(X_r;\Lambda)\ . \tag{6}$$

Finally, the loss function for a single utterance r is typically a sigmoid function as originally proposed in [2][8]:

$$l_r\left(d_r(X_r,\Lambda)\right) = \frac{1}{1+e^{-\alpha d_r(X_r,\Lambda)}} \tag{7}$$

where we assume $\alpha = 1$ for simplicity in exposition without loss of generality. This loss function emulates the zero-one recognition error count function. i.e., when $d_r(X_r,\Lambda)>0$, which implies misclassification, the loss function approaches to one, which essentially becomes a recognition error count.

Given the loss function for each sentence r, the loss function over the whole training set with all R training utterances is

$$L_{MCE}(\Lambda)=\frac{1}{R}\sum_{r=1}^{R} l_r\left(d_r(X_r,\Lambda)\right). \tag{8}$$

(8) is the empirical loss defined on the R independent training utterances. It converges to the expected loss as R increase. (8) is closely related to the sentence error rate and is the objective function to minimize for MCE training. The traditional MCE methods minimize (8) via the technique of probabilistic gradient descent or GPD, which we refer the readers to an excellent review [2].

B. Re-formulation of the MCE Objective Function

In this paper, we will discuss the 1-best MCE training, i.e., $s_r \in \{S_r, s_{r,1}\}$ for utterance r. Based on (1) (4) and (6), Equation (7) can be re-written to

$$l_r\left(d_r(X_r,\Lambda)\right)=\frac{p(X_r,s_{r,1}\mid\Lambda)}{p(X_r,s_{r,1}\mid\Lambda)+p(X_r,S_r\mid\Lambda)} \tag{9}$$

Minimizing the MCE objective function $L_{MCE}(\Lambda)$ of (8) is equivalent to maximizing the following function $Q(\Lambda)$ (where R is the fixed number of training utterances):

$$Q(\Lambda)=R\left(1-L_{MCE}(\Lambda)\right)=\sum_{r=1}^{R}\frac{p(X_r,S_r\mid\Lambda)}{p(X_r,s_{r,1}\mid\Lambda)+p(X_r,S_r\mid\Lambda)} \tag{10}$$

$Q(\Lambda)$ is a sum of multiple rational functions, not amenable to the EBW-style optimization which requires a single rational function. In the following, we will show how $Q(\Lambda)$ can be re-formulated to a special rational function in the form of

$$P(\Lambda)=\frac{\sum_{s_1}\cdots\sum_{s_R}\left[p(X_1,\ldots,X_R,s_1,\ldots,s_R\mid\Lambda)\sum_{r=1}^{R}\delta(s_r,S_r)\right]}{\sum_{s_1}\cdots\sum_{s_R}p(X_1,\ldots,X_R,s_1,\ldots,s_R\mid\Lambda)} \tag{11}$$

where s_r is a variable of the r-th utterance, which can be either the correct training string S_r, or the best incorrect string $s_{r,1}$, and $\delta(s_r,S_r)=0$ if $s_r=s_{r,1}$; $\delta(s_r,S_r)=1$ if $s_r=S_r$.

To show that $P(\Lambda) = Q(\Lambda)$, we make the usual assumption that training utterances are independent with each other:

Then, starting from (11), we have,

$$P(\Lambda)=$$

$$\frac{\sum_{s_1}\cdots\sum_{s_R}\left[p(X_1,s_1\mid\Lambda)p(X_2,\ldots,X_R,s_2,\ldots,s_R\mid\Lambda)\left[\delta(s_1,S_1)+\sum_{r=2}^{R}\delta(s_r,S_r)\right]\right]}{\sum_{s_1}\cdots\sum_{s_R}p(X_1,s_1\mid\Lambda)p(X_2,\ldots,X_R,s_2,\ldots,s_R\mid\Lambda)}$$

$$=\frac{\sum_{s_1}p(X_1,s_1\mid\Lambda)\delta(s_1,S_1)\sum_{s_2}\cdots\sum_{s_R}p(X_2,\ldots,X_R,s_2,\ldots,s_R\mid\Lambda)}{\sum_{s_1}p(X_1,s_1\mid\Lambda)\sum_{s_2}\cdots\sum_{s_R}p(X_2,\ldots,X_R,s_2,\ldots,s_R\mid\Lambda)}$$

$$+\frac{\sum_{s_1}p(X_1,s_1\mid\Lambda)\sum_{s_2}\cdots\sum_{s_R}\left[p(X_2,\ldots,X_R,s_2,\ldots,s_R\mid\Lambda)\sum_{r=2}^{R}\delta(s_r,S_r)\right]}{\sum_{s_1}p(X_1,s_1\mid\Lambda)\sum_{s_2}\cdots\sum_{s_R}p(X_2,\ldots,X_R,s_2,\ldots,s_R\mid\Lambda)}$$

$$=\frac{\sum_{s_1}p(X_1,s_1\mid\Lambda)\delta(s_1,S_1)}{\sum_{s_1}p(X_1,s_1\mid\Lambda)}+\frac{\sum_{s_2}\cdots\sum_{s_R}\left[p(X_2,\ldots,X_R,s_2,\ldots,s_R\mid\Lambda)\sum_{r=2}^{R}\delta(s_r,S_r)\right]}{\sum_{s_2}\cdots\sum_{s_R}p(X_2,\ldots,X_R,s_2,\ldots,s_R\mid\Lambda)}$$

$$=\frac{\sum_{s_1}p(X_1,s_1\mid\Lambda)\delta(s_1,S_1)}{\sum_{s_1}p(X_1,s_1\mid\Lambda)}+\ldots+\frac{\sum_{s_R}p(X_R,s_R\mid\Lambda)\delta(s_R,S_R)}{\sum_{s_R}p(X_R,s_R\mid\Lambda)}$$

$$=\sum_{r=1}^{R}\frac{p(X_r,S_r\mid\Lambda)}{p(X_r,s_{r,1}\mid\Lambda)+p(X_r,S_r\mid\Lambda)}=Q(\Lambda).$$

If we denote by $X=X_1,\ldots,X_R$, $s=s_1,\ldots,s_R$, and $C(s)=\sum_{r=1}^{R}\delta(s_r,S_r)$, $P(\Lambda)$ can be re-written to

$$P(\Lambda)=\frac{\sum_{s}p(X,s\mid\Lambda)C(s)}{\sum_{s}p(X,s\mid\Lambda)} \tag{12}$$

After the re-formulation of (12), $P(\Lambda)$ as the new objective function for MCE (i.e., maximizing $P(\Lambda)$ gives a minimum MCE loss) is strictly in the form of a (single) rational function. This enables us to invoke the EBW optimization technique [4][5] that was previously applied to MMI only. In the next section, we will give the EBW or growth transformation formulas for re-estimating HMM parameters that optimize $P(\Lambda)$ and thus the MCE criterion. Due to the space limitation, we only give a brief derivation for estimation of Gaussian mean and covariance. A detailed version will be presented in a full paper in the future.

C. Growth Transformation for MCE – Auxiliary Function

Growth transformation is an iterative optimization scheme where if the parameter set Λ is subject to a transformation $\Lambda=T(\Lambda')$, then the objective function "grows" in its value $O(\Lambda)>O(\Lambda')$ unless $\Lambda=\Lambda'$. Hence the name "growth transformation". For the interest of this paper, let Λ consist of all mean vector and covariance matrix parameters of Gaussian HMMs, and let Λ' be the parameter set obtained from the immediately previous iteration.

As in [4], we construct the auxiliary function of

$$F(\Lambda;\Lambda')=G(\Lambda)-P(\Lambda')H(\Lambda)+D \tag{13}$$

where $G(\Lambda)=\sum_s p(X,s\,|\,\Lambda)C(s)$, $H(\Lambda)=\sum_s p(X,s\,|\,\Lambda)$.

Then, as shown in [4], as long as D is a quantity not relevant to Λ, increasing $F(\Lambda;\Lambda')$ guarantees to increase $P(\Lambda)$; i.e., $P(\Lambda)-P(\Lambda')=\frac{1}{H(\Lambda)}\left[F(\Lambda;\Lambda')-F(\Lambda';\Lambda')\right]$.

After substitution, we first expand (13) into

$$F(\Lambda;\Lambda')=\sum_q\sum_s p(X,q,s\,|\,\Lambda)[C(s)-P(\Lambda')]+D \tag{14}$$

where q is the Gaussian state sequence, and $s = s_1,\ldots,s_R$, is a word string for each of the training utterances. And since Λ consists of only mean and variance parameters, (q,s) is independent of Λ, and hence $p(X,q,s|\Lambda) = p(X|q,\Lambda)p(q,s)$. Therefore,

$$\begin{aligned}F(\Lambda;\Lambda')&=\sum_q\left[\sum_s p(q,s)[C(s)-P(\Lambda')]\right]p(X\,|\,q,\Lambda)+D\\&=\sum_q\int_\chi\left[\Gamma(\Lambda')+d(q)\right]p(\chi\,|\,q,\Lambda)d\chi\end{aligned} \tag{15}$$

where $\Gamma(\Lambda')=\delta(\chi,X)\sum_s p(q,s)[C(s)-P(\Lambda')]$, and $D=\sum_q d(q)$ is a quantity independent of Λ.

If $d(q)$ is selected to make the term $[\Gamma(\Lambda')+d(q)]$ positive, according to Jensen's inequality, increasing the value of $F(\Lambda;\Lambda')$ can be achieved by maximizing the following function $V(\Lambda;\Lambda')$:

$$\begin{aligned}V(\Lambda;\Lambda')&=\sum_q\int_\chi\left[\Gamma(\Lambda')+d(q)\right]p(\chi\,|\,q,\Lambda')\log p(\chi\,|\,q,\Lambda)d\chi\\&=\sum_q\left[\sum_s p(X,q,s\,|\,\Lambda')\big(C(s)-P(\Lambda')\big)\right]\log p(X\,|\,q,\Lambda)\\&\quad+\sum_q d(q)\int_\chi p(\chi\,|\,q,\Lambda')\log p(\chi\,|\,q,\Lambda)d\chi\end{aligned} \tag{16}$$

Dividing $V(\Lambda;\Lambda')$ by a factor $p(X\,|\,\Lambda')$, which is positive and independent of Λ, we have the auxiliary function

$$\begin{aligned}&U(\Lambda;\Lambda')\\&=\sum_q\left[\sum_s p(s\,|\,X,\Lambda')p(q\,|\,X,s,\Lambda')\big(C(s)-P(\Lambda')\big)\right]\log p(X\,|\,q,\Lambda)\\&\quad+\sum_q d'(q)\int_\chi p(\chi\,|\,q,\Lambda')\log p(\chi\,|\,q,\Lambda)d\chi\end{aligned} \tag{17}$$

where $d'(q)=d(q)/p(X\,|\,\Lambda')$.

Define $\gamma_{m,r,s_r}(t)\triangleq p(q_{r,t}=m\,|\,X_r,s_r,\Lambda')$ as the posterior probability of being in state m in the corresponding HMM at time t given utterance r for word string s_r, where $\gamma_{m,r,s_r}(t)$ is ready to compute through the forward-backward algorithm as in [15]. Then after several steps of algebraic expansion, we obtain

$$\begin{aligned}U(\Lambda;\Lambda')&=\sum_r\sum_t\sum_m\Delta\gamma_{m,r}(t)\log p(x_{r,t}\,|\,m,\Lambda)\\&\quad+\sum_r\sum_t\sum_m d'(r,t,m)\int_{\chi_t}p(\chi_t\,|\,m,\Lambda')\log p(\chi_t\,|\,m,\Lambda)d\chi\end{aligned} \tag{18}$$

where $d'(r,t,m)=\sum_{q:q_{r,t}=m}d'(q)$, and χ_t is a vector-valued variable in the feature space, and

$$\Delta\gamma_{m,r}(t)=p(S_r\,|\,X_r,\Lambda')p(s_{r,1}\,|\,X_r,\Lambda')\left(\gamma_{m,r,S_r}(t)-\gamma_{m,r,s_{r,1}}(t)\right) \tag{19}$$

D. Growth Transformation for MCE – Optimization

We now optimize the auxiliary function given by (18) in order to establish the growth transformation formulas for estimating the HMM parameters. To proceed, we set

$$\partial U(\Lambda;\Lambda')/\partial\Lambda=0 \tag{20}$$

Using $\int_{\chi_t}p(\chi_t\,|\,m;\Lambda')d\chi_t=1$, $\int_{\chi_t}\chi_t\cdot p(\chi_t\,|\,m;\Lambda')d\chi_t=\mu'_m$, and $\int_{\chi_t}(\chi_t-\mu'_m)(\chi_t-\mu'_m)^T p(\chi_t\,|\,m;\Lambda')d\chi_t=\Sigma'_m$, we obtain the results for the mean vector and covariance matrix associated with the Gaussian at state m of the HMM as:

$$\mu_m=\frac{\sum_r\sum_t\Delta\gamma_{m,r}(t)x_{r,t}+D_m\mu'_m}{\sum_j\sum_t\Delta\gamma_{m,r}(t)+D_m}, \tag{21}$$

$$\begin{aligned}\Sigma_m=\frac{1}{\sum_r\sum_t\Delta\gamma_{m,r}(t)+D_m}\Big\{&\sum_r\sum_t\left[\Delta\gamma_{m,r}(t)(x_{r,t}-\mu_m)(x_{r,t}-\mu_m)^T\right]\\&+D_m\Sigma'_m+D_m(\mu_m-\mu'_m)(\mu_m-\mu'_m)^T\Big\}\end{aligned} \tag{22}$$

where $D_m=\sum_{r=1}^{R}\sum_t d'(r,t,m)$.

E. Setting quantity D_m

According to the above derivation, quantity D_m in the above growth transformation or EBW estimates (21) and (22) is set based on $d(q)$, and $d(q)$ is chosen so that the term $[\Gamma(\Lambda')+ d(q)]$ in (15) is positive. However, this may lead to a very large D_m and slow down the training process. In practice, we found that the following form of D_m works well,

$$\begin{aligned}D_m=E\cdot\sum_{r=1}^{R}p(S_r\,|\,X_r,\Lambda')\Big[&\;p(S_r\,|\,X_r,\Lambda')\sum_t\gamma_{m,r,S_r}(t)\\&+p(s_{r,1}\,|\,X_r,\Lambda')\sum_t\gamma_{m,r,s_{r,1}}(t)\;\Big]\end{aligned} \tag{23}$$

Readers may notice that the final formulas (21) and (22) for MCE training have a similar form as to EBW equations for MMI training (e.g., [10]). However, the computation of several key terms including $\Delta\gamma_{m,r}(t)$ and D_m are very different.

III. Experimental Results and Discussion

The experiments reported in this section are designed to evaluate the new HMM learning technique based on the EBW formulas (20) (21) for the MCE training, which is very different from the traditional gradient descent (GPD) optimization method in [8]. We tested on two standard speech recognition tasks, one small and one large, with details of the experimental setups and results described below.

A. *Experiments on the TI-Digits Task*

TI-Digits is a speaker independent connected-digit task. Each utterance in this corpus has an unknown length with a maximum of 7 digits. The training set includes 8623 utterances and testing set includes 8700 utterances. In our experiment, a word-based HMM is built for each of the ten digits from *ZERO* to *NINE*, plus word *OH*. The number of states of each HMM ranges from 9 to 15, depending on the average duration of each word, and each state has an average of six Gaussian mixture components. The speech feature vector is computed from audio signal analysis, which gives 12 MFCCs (Mel-Frequency Cepstral Coefficients) and the audio energy, plus their first-order and second-order temporal differences.

The experimental results are shown in Table 1. The baseline HMMs are trained with the maximum likelihood (ML) criterion. The ML model gives a word error rate (WER) of 0.30% on testing data. With growth transformation or EBW-based MCE training, after four iterations, the algorithm convergence is reached and the WER is reduced to 0.23%. As a comparison, a conventional GPD-based MCE is also implemented for this task. As shown in Table 1, the best GPD MCE result is with WER of 0.24%, which is obtained after 12 iterations over the full training data set (i.e., 12 epochs). The results of this small-task experiment show that the new EBW-based MCE learning method is slightly better than the conventional GPD-based MCE, and it gives significantly improved efficiency in the training by providing much faster algorithm convergence.

	ML	GPD MCE	EBW MCE
WER	0.30%	0.24%	0.23%
WER Reduction	–	20.0%	23.3%

Table 1. Comparative recognition-accuracy performance (measured by WER – the lower the better) of the new and traditional MCE training methods, as well as the ML method

B. *Experiments on the WSJ Task*

Experimental speech recognition studies have also been conducted on the large-vocabulary WSJ0 task. The standard WSJ0 SI-84 training set is used to train the baseline ML HMMs and the standard 5K-vocabulary trigram test set is used for model testing. The training data set has 15 hours of speech data from 84 speakers, and the test data set has 330 sentences of speech from 8 speakers. Context-dependent cross-word triphone units are used. State clustering is performed based on a phonetic decision tree and 3400 tied HMM states are generated to form the acoustic model, with an average of 12 Gaussian mixture components for each HMM state. The speech feature vector is 12 MFCCs and the audio energy, plus their first and second temporal differences.

For large-vocabulary speech recognition tasks such as the WSJ one, the conventional sample-by-sample sequential GPD algorithm is not feasible to implement due to the difficulty of parallelizing the training process. Hence, in our experiments, only the proposed growth transformation or EBW-based MCE training (which always runs in a batch mode) is tested. A 60K trigram language model is used to decode the training data for generating the competing candidates in the MCE training. After each iteration of the EBW algorithm, the training data are re-decoded using the model generated from the previous iteration. Then the new best-incorrect-strings are used in the next iteration of the MCE training.

The recognition results are summarized in Table 2. The ML-trained baseline system (denoted by Iteration 0) gives a WER of 4.6% on the test set. This baseline HMMs give the overall loss function $R{\cdot}L_{MCE}(\Lambda)$ of 2973.2 (from Eq. (8)) for the entire training data. After eight iterations of the proposed growth transformation or EBW-based MCE training, the WER is reduced progressively to 4.2%, which corresponds to an 8.7% WER reduction. Correspondingly, the $L_{MCE}(\Lambda)$ value of the overall MCE loss function is also progressively reduced, as shown in Table 2. As expected, the reduction of the MCE loss function is stable with each iteration; i.e., no oscillation over the iteration (which often happens with the traditional gradient descent technique of GPD). This property of stable convergence associated with the proposed optimization approach, as well its greater efficiency over the GPD method has been observed throughout our experiments.

Iterations	$R{\cdot}L_{MCE}(\Lambda)$ (training set)	WER (test set)
0 (ML)	2973.2	4.6%
2	2302.3	4.4%
4	1924.3	4.3%
6	1662.0	4.3%
8	1468.4	4.2%

Table 2. Increasing performance (measured by decreasing WER for the test set) of the new method for MCE training as a function of the training iteration. The MCE loss function $L_{MCE}(\Lambda)$ in the training over the iterations is also shown.

IV. Conclusion and Future Work

In this paper, a novel growth-transformation or EBW-based method is developed for MCE-based discriminative learning of HMM parameters. We present the solid theoretical foundation for this new method with mathematical rigor, which has not been published in the earlier literature. In contrast to the conventional sample-by-sample sequential gradient descent methods for MCE optimization such as GPD, the proposed method has stable convergence and is easy to parallelize in implementation over multiple processors, in addition to being more theoretically appealing. We have implemented and evaluated this new learning/training method on two speech recognition tasks where audio analysis of the speech waveform provides the fixed MFCC acoustic features. In the small vocabulary task, we find that the new training method provides significantly faster convergence and is more stable than the traditional GPD method. It also gives better recognition performance. For the large-vocabulary speech recognition task, the traditional MCE optimization method such as GPD is not feasible, because its sequential nature makes it difficult to parallelize the training which is needed for the very large amount of training data. Our new batch-mode optimization method is directly parallelizable, as we have implemented in the WSJ task. We have observed highly stable and fast convergence of the EBW algorithm and achieved significant recognition performance advantages over the baseline ML training.

Our future work involves further refinement of the learning technique discussed in this paper in two ways. First, we plan to use recognition lattices as the source of competing candidates in MCE-base discriminative training. The lattices give a much richer representation of the competing candidates than the 1-best string as we presented in this paper. Second, we plan to modify the MCE criterion so that it can be applied to substring-level (e.g., word) performance optimization. Further, we plan to apply the novel discriminative learning and optimization method present in this paper to other pattern recognition problems, including those involving joint audio-visual patterns.

References

[1] P. Brown. The Acoustic Modeling Problem in Automatic Speech Recognition, Ph.D. thesis, Carnegie Mellon University, 1987.

[2] W. Chou. "Minimum classification error approach in pattern recognition," in *Pattern Recognition in Speech and Language Processing*, (W. Chou and B.-H. Juang eds.) 20043, CRC Press, Boca Raton, pp. 1-49.

[3] L. Deng, J. Wu, J. Droppo, and A. Acero. "Analysis and comparison of two feature extraction/compensation algorithms," IEEE Signal Processing Letters, Vol. 12, No. 6, June, 2005, pp. 477-480.

[4] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems," IEEE Trans. Information Theory., Vol 37, pp. 107-113, January. 1991.

[5] A. Gunawardana and W. Byrne, "Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression," Proc. EUROSPEECH, 2001.

[6] X. He and W. Chou, "Minimum Classification Error Linear Regression For Acoustic Model Adaptation of Continuous Density HMMs," Proc. ICASSP, April 2003.

[7] B.-H. Juang, and S. Katagiri. "Discriminative Learning for Minimum Error Classification, IEEE, Trans. on SP., Vol. 40, No. 12, 1992, pp. 3043-3054.

[8] 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.

[9] W. Macherey, L. Haferkamp, R. Schluter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech Recognition," Proc. Interspeech, Sept., 2005, Lisbon, Portugal, pp. 2133-2136.

[10] Y. Normandin, "Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem," PhD. dissertation, McGill University, Montreal, 1991.

[11] D. Povey and P.C. Woodland, "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," Proc. ICASSP, 2002.

[12] D. Povey, M.J.F. Gales, D.Y. Kim, and P.C. Woodland, "MMI-MAP and MPE-MAP for Acoustic Model Adaptation," Proc. Eurospeech, September 2003.

[13] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, Cambridge, UK, 2003.

[14] C. Rathinavalu and L. Deng. "Speech trajectory discrimination using the minimum classification error learning," IEEE Trans. Speech and Audio Processing, Vol.6, No.6, Nov. 1998, pp. 505-515.

[15] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[16] J. L. Roux and E. McDermott, "Optimization for Discriminative Training," Proc. INTERSPEECH, 2005.

[17] E. McDermott and T. Hazen, "Minimum Classification Error Training of Landmark Models for Real-Time Continuous Speech Recognition," *ICASSP04', Proc.* May, 2004.