Generalization of the Minimum Classification Error (MCE) Training Based on Maximizing Generalized Posterior Probability (GPP)

Qiang Fu, Antonio Moreno-Daniel, Biing-Hwang Juang
Department of Electrical and Computer Engineering
Georgia Institute of Technology, GA 30332
{qfu,antonio,juang}@ece.gatech.edu

Jian-Lai Zhou, Frank Soong
Microsoft Research Asia
Beijing, China
{jlzhou,frankkps}@microsoft.com

Abstract

In this paper, we generalize the training error definitions for minimum classification error (MCE) training and investigate their impact on recognition performance. Starting the conventional MCE method, we discuss with three issues in regard to training error definition, which may affect the recognizer performance and need to be extensively studied. We focus our discussions on the first two aspects in this paper. We re-visit the fact that the objective function in MCE training can be formulated into an equivalent form for maximizing the "posterior probability" of the corresponding training units. Based on the framework of the generalized posterior probability (GPP) [1], we design experiments to demonstrate effects about different training units and different constraints on segmentation boundaries for the MCE training. We also provide a performance analysis to illustrate our generalization for both phone recognition and word recognition tasks based on the wall street journal (WSJ0) [2, 3] database.

Index Terms: MCE, GPP, WSJ0.

1. Introduction

Discriminative training (DT) methods [4][5] have led to successful results in various of automatic speech recognition (ASR) tasks. Fundamentally, three popular discriminative training methods have been proposed; they are the *maximum mutual information* (MMI) method [6][7], the *minimum phone/word error* (MPE/MWE) method [8] , and the *minimum classification error* (MCE) method [4][5]. We have witnessed great effort in the whole ASR community to compare, unify, and generalize these criteria. Recent research indicates that it is possible to formulate the discriminative training methods under a unified function form with substantial generalization [5]. We choose the MCE criterion as the main investigation objective, because it is the one which elaborates the most direct connection between the Bayes decision rule and the speech recognition performance (i.e., empirical error rate). The MMI method roots in maximizing the mutual information between two probabilistic models. The MPE/MWE method shares a similar objective function with the MCE method, and can be viewed as a customization of the MCE criterion.

Following the convention of [4], the MCE method can be formulated:

1. Define the performance objective and the corresponding task evaluation measure;
2. Specify the target event (i.e., the correct label), competing events (i.e., the incorrectly hypothesis results from the recognizer) , and the corresponding models (a good organization of training events is also critical);
3. Construct the objective function and set hyper-parameters;
4. Choose a suitable optimization method to obtain model parameters.

Here an "event" could be any user-specified speech units (e.g. phones, syllables, words, etc.). The unique feature of MCE is that the objective function in step 3 is chosen to be the same as the performance objective in step 1. With this principle, we contemplate a more flexible generalization starting from the MCE method to all discriminative criteria, in which not only the functional forms, but also all relevant components are organized together based on a thorough and systematic consideration. For instance, MCE should no longer be confined as a criteria only for isolated training or a string-based method, while the training level (string-based/word-based/phone-based,etc.) is viewed only as a part of the system configuration instead of a critical differentiation factor between criteria. In this paper, only the first item which is also the most fundamental one is discussed due to limited space. This work is the first of an extensive generalization of the MCE training criterion. The purpose of this paper is to provide a discussion in regard to generalization of the training error definitions for the MCE method, Furthermore, the experiments are conducted in order to demonstrate the effect of distinct error definitions rather than aiming at an optimal system development.

As we have mentioned, the MCE training ties the classic Bayes decision rule and the task evaluation performance. Bayes decision is the basis of pattern recognitions. It teaches us the best decision is the one with the maximum posterior probability. In particular, for speech recognition, the recognized speech unit should maximize the following equation:

$$W_i = \arg\max_i P(W_i|X_i) = \frac{P(X_i|W_i)P(W_i)}{P(X_i)} \qquad (1)$$

where $P(X_i|W_i)$ is the acoustic model and $P(W_i)$ is the language model. By its definition, the MCE training is directly minimizing the empirical training errors, which is the performance measure. Later in the next section, we can see that this operation is consistent with maximizing the posterior probability. Estimating the total probability $P(X_i)$ is normally very hard in ASR problems because there is impossible to exhaust the probability space. Wessel *et al.* [9] has proposed a method to estimate the posterior probability for a speech unit using a reduced search space (i.e., word graph). In [1], a more relaxed estimation was introduced as "generalized posterior probability" (GPP). Though the original motivation for these work is to propose a new confidence measure for recognized hypotheses, a method was shown a path for applying the MCE to large vocabulary continuous speech recognition (LVCSR) [5]. Note that the term "posterior probability" is interpreted here for the pattern recognition problem which may not be identical to the conventional distribution estimation problem. That is, training based on "maximizing posterior probability" in speech recognition only means to raise the ranking of the labeled speech units.

The rest of paper is organized in the following way. In section 2, we will present some issues in error definition generalization. We will construct a relationship between the MCE and maximization of the posterior probability for a specific speech unit in section 3. A brief review of the definition of GPP is also provided in this section. The experimental results that illustrate the effect of different training units are exhibited in section 4. The conclusion of this paper is drawn in section 5 and the future work is proposed, too.

2. Error Definitions Generalization in MCE Training

There are three important issues in defining an error: the level of training, the speech unit boundaries, and error types in regard to the unit alignment (substitution/insertion/deletion).

2.1. Error Definition in Different Level of Training

There are a number of training units in different levels which can be used to measure recognizer performance. Thus, to count training errors. Practically, a sequence of observations in ASR are indexed by frame, which is the smallest unit in calculating features. Therefore, we may list a set of speech units bottom-up: **state, phoneme, syllable, word, phrase, string/sentence/utterance**. This list may not be optimal or exhaustive, but it is a good demonstration for the flexibility of a hierarchical MCE training structure. For each level, an error is defined as the number of discrepancies between the manually transcribed units and the recognized units. Any level of models can be formed from a basic level. For example, the speech unit to construct acoustic models for English recognition usually is phoneme, hence the higher level models such as word level training can be carried out by simply concatenating phoneme models. On the other hand, a lower level training like state level is conducted by aligning phoneme sequences into state sequences. Later in this paper, three levels of training units (word/phone/state) will be discussed.

2.2. Error Definition for Speech Unit Boundaries

The principle of MCE training is built upon the identification of the numerically competitive situation between the transcribed target event and the competing events. In continuous speech recognition tasks, it is in general difficult to obtain a set of competing events with identical boundaries with the transcribed event. A relaxed boundary constraint was successfully applied in discriminative training [5]. For example, at each time frame *t* in a word graph, where new word hypotheses are to be started, not only the word hypotheses starting at exactly this frame are allowed to be one of the competing events, but also those starting in time interval $[t - \delta t, t + \delta t]$ need to be counted as well. We analyze MCE formulations with both strict and relaxed boundaries in the following.

2.3. Error Definition in and with Unit Alignment

In continuous speech recognition, recognition errors can be classified into three types in terms of their position in the alignment between the recognized string and the transcription. They are deletion errors, insertion errors, and substitution errors. These three kinds of errors have been considered in conventional speech **recognition** problems. The original MCE method aims at minimizing the substitution errors. However, if we re-interpret the ASR problem as a **detection** problem, the deletion/insertion/substitution errors can be respectively viewed as miss errors, false alarm errors, and miss/false-alarm errors happening together. Now we can directly minimize all errors under the framework of detection theory. This is one of the essence of the detection-based ASR [10], which would be discussed separately.

3. MCE Training and Generalized Posterior Probability

3.1. MCE Objective Re-formulation

The heart of the MCE method is to embed the empirical training error into a function that is easy to optimize. Let us recall the error counting function:

$$l_i(X_i, W_i) = \frac{1}{1 + \exp[-\gamma(-g_i(X_i, W_i) + G_i(X_i, W_i)) + \theta]} \tag{2}$$

where X_i is the ith training observation, and W_i is the corresponding label. g_i and G_i are log likelihood of the labeled model (target model) and the wrongly recognized model (competing model). when we set $\gamma = 1$ and $\theta = 0$, The above function can be rewritten as

$$\begin{aligned} l_i(X;\lambda) &= 1 - \frac{P_C(X_i, W_i)}{P_C(X_i, W_i) + P_W(X_i, W_i)} && (3) \\ &= 1 - \frac{P(X_i|W_i)P(W_i)}{\sum_{\forall W_j} P(X_i|W_j)P(W_j)} && (4) \\ &= 1 - P(W_i|X_i) && (5) \end{aligned}$$

in which P_C denotes the probability computed using the labeled target model, and P_W are probability from competing models. This derivation indicates how the MCE method associates the empirical training errors with the Bayes decision rule.

3.2. Generalized Posterior Probability (GPP)

In [1], a confidence measure using GPP for recognized word is proposed. When the notion of word level decision is broadened to other levels, the MCE training criterion allows integration of the corresponding GPP. As well known, a word graph bears a richer search space than an N-best list. In this paper, all training experiments are conducted in the context of word graphes. Assume there is a labeled word sequence $W_1^M = w_1, w_2, \ldots, w_M$ with observation $X_1^T = x_1, x_2, \ldots, x_T$. We represents a word w_i starting from time s and ending at time t as $[w_i; s, t]$. Hence, the posterior probability can be written as [9]

$$P([w_i; s, t]|X_1^T) = \sum_{\forall n, [w_n; s_n, t_n] = [w_i; s, t]} \frac{P^\alpha(X_s^t|w_i)P^\beta(w_i|w_{i-1})}{P(X_1^T)} \tag{6}$$

where w_n is a hypothesized word, and $P(w_i|w_{i-1})$ is the language model. α and β are acoustic and language model scale factors, respectively. $[w_n; s_n, t_n] = [w_i; s, t]$ implies the recognized word has same identity and exact starting and ending time with the labeled one. This definition has one obvious drawback. In recognition, the word might be correctly recognized but the time registration s and t of that hypothesis often does not satisfy the exact match. Since the word content is more important than the timing information (unless it negatively impacts the recognition decision on the neighboring segments), it is unnecessary to impose the strict constraint upon word boundaries at this stage. Therefore, the GPP is defined in [1]

$$\begin{aligned} P_G([w_i; s, t]|X_1^T) \quad &= \sum_{\forall n, w_n = w_i} \frac{P^{\alpha}(X_s^t|w_i)P^{\beta}(w_i|w_{i-1})}{P(X_1^T)} \\ when \quad &[s_n, t_n] \cap [s, t] \neq \emptyset \end{aligned} \tag{7}$$

This equation relaxes the timing constraint, indicating that once the hypothesis identity matches the labeled one, we will count it as a correct recognition as long as there is a reasonable segmentation overlap. Of course, the constraint of timing can be tightened somehow in order to achieve different recognition error definitions in regard to boundaries. As we have mentioned this is the generalization for speech unit boundaries in error definitions. Based on the derivation above, we can conclude the objective function to be maximized under the MCE/GPP framework as

$$\mathcal{F}_{GPP}(\Lambda) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} P_G([w_i; s, t]|X_1^T) \tag{8}$$

where K is the number of total training tokens. We can use either the GPD method [4] or the EBW method [5] to optimize the objective function (8). In this paper, the GPD method is employed for simplicity. The update equations for all parameters can be found in [4].

3.3. Two Implementation Issues

One issue for graph-based MCE training is that we need to exclude the spoken word sequence from the recognized word graphes [11]. It is not appropriate to physically remove any utterance hypothesis because some words in that sequence may be a part of the other word sequences. Hence, the total likelihood of the graph is calculated using forward-backward method [9] then the contribution of the correctly recognized word sequences are subtracted. The other issue is to select the acoustic and language model scale factor α and β. In [1] some techniques in searching the optimal values for them are proposed. In this paper, we apply some heuristic values that prepared in advance for simplicity.

4. Experiments and Results

4.1. System Description

The experiments were conducted on the WSJ0 database [2, 3]. The baseline recognizer followed the large vocabulary continuous speech recognition recipe using HTK (see http://www.inference.phy.cam.ac.uk/kv227/htk/) , which was based on representing training classes using continuous density Gaussian mixture hidden markov models (CDHMM). A word internal context-dependent tri-phone set is formed with 7,385 physical models and 19,075 logic models. All models are represented by 3-state strict left-to-right HMMs, with 8 Gaussian mixture components per state. These models were trained first by Maximum Likelihood (ML) method implemented by the HTK toolkit. The experiments were then carried out by comparing the performance of systems trained using different MCE criterion.

We generated feature vectors for all 7,077 utterances by 84 speakers in the training set of the WSJ0 corpus. Each feature vector has 12MFCC+12Δ+12Δ^2 and 3 log energy values so that total 39 features are used. The feature generation process is also applied on the Now'92 evaluation set with 330 utterances by 8 speakers. The CMU6 recognition lexicon are employed, which contains 126,834 words. The word graphes are generated using HTK toolkit, too. At most 3 tokens are allowed to survive at the same time during word graph generation. Other baseline system details can be found in [12].

The MCE training criterion are tested on three aspects based on the GPP framework. First, we conducted experiments on different training levels. Second, we investigated the effect of selecting different constraint for the word boundaries. Third, we presented an investigation of the effects for different training constraints, i.e., different word graphes. In each set of experiments, the other two factors are keeping identical for consistency.

4.2. Experiment Results

4.2.1. MCE on different training levels

The first experiments are investigating the effects of different training levels for the MCE method. There are three popular training levels for word recognition experiments: the state level, the phone level and the word level. At each specific level, the GPP is computed at that level and the GPD method is used for optimization. For example, the word level training means we first calculated the GPP for each word in the graph, then all parameters of that word model are updated based on the GPP value. Assuming the spoken word occupies the time interval $[s, t]$, We allow any words falling into the time interval $[s - \delta t, t + \delta t]$ are counted into the calculation of GPP. In this paper, we are using a relative constraint $\delta t = 1/3(t - s)$. The word boundaries are read from the word graph files, and the phone boundaries and state boundaries are set using Viterbi alignment. To generate the word graph, a bigram language model is applied. The word insertion penalty and the language model scale factor are set to be -14.0 and 5.0, respectively.

In table 1, we can see that the phone-level training achieved slightly better performance than the the word level and state level training. The reason for this observation is that the time interval for state level training when calculating GPP may be too short, and the time interval for word level may be too long. Too short interval could lead over-optimization. Too long interval contains too many parameters so that the effect for each parameter is weakened when maximizing the corresponding GPP.

Table 1: *Word Error Rate (WER) and Sentence Error Rate (SER) for WSJ0-eval using different training levels*

Training level	WER	SER
Baseline	8.41	57.88
Word-level	8.05	56.97
Phone-level	7.96	56.67
State-level	8.02	56.97

4.2.2. MCE on different speech unit boundaries

We only apply the word level experiment in this section. The word graphes are also identically generated. Four values of δt are tested. The $\delta t = 0$ means that we use the strict boundary for calculating the GPP. From Table 2, we can see that the best performance happens when we applied the relaxed boundary $\delta t = 1/3(t - s)$. As expected, the strict boundary didn't work as well as the relaxed boundary. The comparison between two relaxed boundaries show that the time interval $\delta t = 1/2$ is too wide.

Table 2: *Word Error Rate (WER) and Sentence Error Rate (SER) for WSJ0-eval using different boundary constraints*

δt	WER	SER
Baseline	8.41	57.88
$\delta t = 0$	8.10	57.27
$\delta t = 1/3(t - s)$	8.05	56.97
$\delta t = 1/2(t - s)$	8.15	57.58

4.2.3. MCE on different word graphes

We compare the training effect of the MCE method under two word graphes in this section. The first graph is built using the configuration listed before, and the second one is generated with the word insertion penalty to be -8.0 and the language scale factor to be 4.0. Still, the experiments in this section are only for word-level training. The relative constraint for word boundary is set to be $\delta t = 1/3(t - s)$. The second type of word graphes have higher graph density (i.e., it contains more recognized word candidates) than the first one because it loosens the constraint of the language model and the word insertion penalty. We can see that the performance of using the second type of word graph is slightly better than the first type. We believe the reason for this observation is that the second type of graphes contains more competing words. As well known, one performance bottleneck for the MCE method is that the competing events may probably be rare so that the GPP is dominated by the correct likelihoods.

Table 3: *Word Error Rate (WER) and Sentence Error Rate (SER) for WSJ0-eval using different word graphes*

	WER	SER
Baseline	8.41	57.88
$p = -14, s = 5$	8.05	56.97
$p = -8, s = 4$	8.01	56.67

5. Conclusions

In this paper, we generalize the training error definitions for minimum classification error (MCE) training and investigate their impact on recognition performance.This paper is the first part of an extensive generation of the MCE training. The experiments are conducted based on the framework of "maximizing posterior probability". Three factors are investigated. They are: the impact of different training levels, the impact of different word boundaries, and the impact of different word graphes. For the first factor, we observed the best performance in phone level training. For the second one, the relaxed word boundary shows better performance than the fixed word boundary, and two different relative constraints are compared. For the third factor, training using word graphes with more recognized competitors decreases the word error rate in this task.

6. Acknowledgements

Part of this work was finished when the first author was working for Microsoft Research Asia in 2005 as a summer intern.

7. References

[1] F. K. Soong, W.-K. Lo, and S. Nakamura, "Generalized word posterior probability (gwpp) for measuring reliability of recognized words," in *SWIM-2004*, Maui, Hawaii, Jan. 2004.

[2] D. S. Pallet, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki, "1994 benchmark test for the arpa spoken language program," in *ARPA Human language Technology Workshop*, Austin, Texas, Jan. 1995, pp. 5–36.

[3] F. Kubala, "Design of the 1994 csr benchmark tests," in *ARPA Human Language Technology Workshop*, Austin, Texas, Jan. 1995, pp. 41–46.

[4] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.

[5] R. Schluter, W. Macherey, B. Muller, and H. Ney, "Comparison of discriminative training criteria and optimization methids for speech recognition," *Speech Communication*, vol. 34, no. 1, pp. 287–310, May. 2001.

[6] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, "Mmie training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, 1997.

[7] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden markov models for speech recognition," *Computer Speech and Language*, vol. 16, pp. 25–47, 2001.

[8] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved dsicriminative training," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, May 2002, pp. 105–108.

[9] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, Mar. 2001.

[10] C.-H. Lee and B.-H. Juang, "A new detection paradigm for collaborative automatic speech recgnition and understanding," in *SWIM-2004*, Maui, Hawaii, Jan. 2004.

[11] W. Macherey, L. Haferkamp, R. Schluter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Interspeech-2005*, Lisbon, Portugal, Sep. 2005, pp. 2133–2136.

[12] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using htk," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Adelaide, Australia, Apr. 1994, pp. 125–128.