MINIMUM WORD CLASSIFICATION ERROR TRAINING OF HMMS FOR AUTOMATIC SPEECH RECOGNITION

Zhi-Jie Yan, Bo Zhu, Yu Hu, and Ren-Hua Wang

iFlytek Speech Lab, University of Science and Technology of China, Hefei, P. R. China, 230027
yanzhijie@ustc.edu bozhu@mail.ustc.edu.cn yuhu@iflytek.com rhw@ustc.edu.cn

ABSTRACT

This paper presents a novel discriminative training criterion, *Minimum Word Classification Error* (MWCE). By localizing conventional string-level MCE loss function to word-level, a more direct measure of empirical word classification error is approximated and minimized. Because the word-level criterion better matches performance evaluation criteria such as WER, an improved word recognition performance can be achieved. We evaluated and compared MWCE criterion in a unified DT framework, with other commonly-used criteria including MCE, MMI, MWE, and MPE. Experiments on TIMIT and WSJ0 evaluation tasks suggest that word-level MWCE criterion can achieve consistently better results than string-level MCE. MWCE even outperforms other substring-level criteria on the above two tasks, including MWE and MPE.

Index Terms— Discriminative training, Minimum classification error, Minimum word classification error, Speech recognition

1. INTRODUCTION

It has been shown in recent years that *Discriminative Training* (DT) methods are able to produce solid and consistent performance improvements not only for small parameter sets, but also for large-vocabulary speech recognition tasks. The success of DT methods for large-scale tasks relies on the development of the following key techniques: a) proposing of new DT criteria, b) use of lattices which compactly represent competing space, and c) advance in parameter optimization methods for HMMs.

This study focuses on the first category of techniques, to develop a new discriminative training criterion. The most widely used DT criteria nowadays include conventional *Maximum Mutual Information* (MMI) [1], *Minimum Classification Error* (MCE) [2], and newly proposed *Minimum Word / Phone Error* (MWE / MPE) [3]. Being regarded as derived from the same pipeline of MMI, MWE and MPE are usually shown to outperform string-level MMI on many tasks [3], mainly due to the exploiting of local information of substring-level accuracy. By maximizing the expected word or phone accuracy on training data, MWE and MPE are often seen to be more directly related to performance evaluation measures, such as *Word Error Rate* (WER), than MMI.

Similar to MMI, conventional MCE criterion for HMM based speech recognition is also implemented on string-level [4]. String-level MCE aims at minimizing a smoothed measure of string error, which indirectly minimizes our ultimate goal of word error. It should be noted, however, that there is a considerable mismatch between string error and word error. These two measures are related to some extent, but not equivalent. In order to develop new criteria which minimize word-level error rather than string error, several attempts have been made along the MCE pipeline, including *general MCE loss function* [5], *label-based phoneme-level MCE* [6], and *phone-discriminating MCE* [7]. However, these methods are derived more from intuition, and lack of proof of their relationship with true word errors. Moreover, the performance gain reported in those work when comparing with conventional string-level MCE is only marginal. Therefore, a new MCE based criterion which has closer relation to the word-level performance evaluation measures is more desired.

In this paper, we propose a novel word-level DT criterion, namely *Minimum Word Classification Error* (MWCE). We aim at choosing appropriate discriminant functions, misclassification measure, and loss function so that a more direct measure of word-level error on the training set can be approximated and minimized. In contrast to string-level MCE, minimizing our proposed word-level error measure can directly attack the ultimate problem of minimizing WER. Therefore, an improved word recognition performance can be expected by using MWCE.

The main difference compared with previous research [5, 6, 7] is that in this study, not only an intuitive explanation, but also a theoretical analysis is carried out, to connect MWCE criterion to the word-level error measure. We show that under an ideal condition, MWCE criterion will become an estimate of the number of words being incorrectly recognized on the training data. When the ideal condition cannot be satisfied in practice, a smoothed word-level error measure can still be approximated and form an applicable criterion.

We embedded our MWCE criterion into the unified DT criterion proposed in [8] and [9]. Because all criteria share most of the implementation details in a common framework, the unified DT criterion can provide a fair evaluation of different criteria. We compared MWCE criterion with string-level MMI, MCE, as well as substring-level MWE and MPE. Experimental results on TIMIT and WSJ0 tasks suggest that MWCE not only achieves an improved performance over conventional MCE, but also outperforms other criteria including MMI, MWE, and MPE.

The rest of this paper is organized as follows: In Section 2, the implementation of conventional string-level MCE in the unified DT criterion is briefly reviewed. In Section 3, the proposed MWCE criterion is introduced, and the relationship between MWCE and word-level error measure will be explained. In Section 4, experimental results of MWCE and the comparison with other DT criteria are presented. Finally in Section 5, we will draw our conclusions.

2. STRING-LEVEL MCE AND ITS IMPLEMENTATION IN UNIFIED DT CRITERION

In order to compare conventional string-level MCE with our proposed word-level MWCE, the implementation of string-level MCE

in unified DT criterion is briefly reviewed here.

For the r_{th} training utterance, MCE criterion chooses discriminant functions to form a string-level misclassification measure [4], which can be formulated as:

$$\begin{aligned} d_r \;=\; & -\log p_\theta(X_r|W_r)p(W_r) \\ & +\log\left[\frac{1}{|\mathcal{M}_r|}\sum_{W\in\mathcal{M}_r} p_\theta^\alpha(X_r|W)p^\alpha(W)\right]^{1/\alpha}, \end{aligned} \tag{1}$$

where θ represents the set of all parameters of the emission probabilities, X_r is the observation sequence, W_r the reference word sequence, and α the weighting exponent, respectively. For MCE criterion, $\mathcal{M}_r$ is chosen to be all possible word sequences excluding the reference, i.e., $\mathcal{M}_r = \mathcal{M}\setminus\{W_r\}$, and $|\mathcal{M}_r|$ is the total number of word sequences in $\mathcal{M}_r$.

In order to approximate string error, a loss function of

$$\mathcal{L}(d_r) = \frac{1}{1+e^{-2\gamma(d_r+\xi)}} \tag{2}$$

needs to be introduced. This smoothed loss will be close to 0 if the entire string is recognized correctly, and close to 1 otherwise. In an extreme case when $\alpha\to\infty$ and $\gamma\to\infty$, Eq. (2) will become an indicator function of string error. Therefore, minimizing the sum of this loss over all training utterances will minimize the number of empirical string errors, which indirectly minimizes WER.

By choosing appropriate γ and ξ, an equivalent variant of the string-level loss in Eq. (2) can be embedded into the unified DT criterion [8]:

$$\mathcal{F}_{\text{MCE}} = \sum_{r=1}^{R} f\left(\log\frac{p_\theta^\alpha(X_r|W_r)\cdot p^\alpha(W_r)}{\sum_{W\in\mathcal{M}_r} p_\theta^\alpha(X_r|W)\cdot p^\alpha(W)}\right), \tag{3}$$

where $f(z) = -1/(1+e^{2\rho z})$, and ρ is a smoothing factor (note the negative sign in f because the unified criterion is designed to be maximized for parameter optimization). Following this way, MCE criterion can be evaluated and compared with other DT criteria in a same framework.

3. MINIMUM WORD CLASSIFICATION ERROR CRITERION

Although conventional MCE has successfully embedded the string based *Dynamic Programming* (DP) procedure during decoding into the training process, the drawback of this method is almost obvious. The main disadvantage is that there exists a mismatch between the string-level criterion and word-level performance evaluation measures (e.g., WER). Minimizing string error does lead to a minimization of word error, but this kind of optimization is quite indirect. A new criterion which focuses on word-level error may provide a more effective way to optimize our ultimate goal of WER.

In the following subsections, a word-level, MWCE criterion is proposed. We aim at choosing appropriate discriminant functions, misclassification measure, and loss function so that a more direct measure of empirical word-level error can be approximated and minimized. Meanwhile, we will also give an investigation on the relationship between MWCE criterion and word-level error.

3.1. Word-Level MWCE Loss Function

Suppose the reference word sequence of the r_{th} training utterance is consisted of N_r words, i.e., $W_r = \{w_r^1, w_r^2, \ldots, w_r^{N_r}\}$. For each reference word w_r^n, we first define the *correct string set* $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ and *incorrect string set* $\mathcal{M}_{w_r^n}^{\mathcal{J}}$ such that:

$$\begin{aligned} &\forall W\in\mathcal{M}_{w_r^n}^{\mathcal{K}}, \exists w\in W, w\equiv w_r^n; \\ &\forall W'\in\mathcal{M}_{w_r^n}^{\mathcal{J}}, \forall w'\in W', w'\neq w_r^n. \end{aligned} \tag{4}$$

In Eq. (4), $w\equiv w_r^n$ means that we restrict the word w to have a same label and same time alignment as the reference word w_r^n. Therefore, the correct string set $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ will include all DP strings that pass through a "matched" word w for the corresponding interval of speech. Conversely, the incorrect string set $\mathcal{M}_{w_r^n}^{\mathcal{J}}$ is consisted of all strings that do not pass any "matched" word for w_r^n. Obviously, $\mathcal{M}_{w_r^n}^{\mathcal{K}}\cap\mathcal{M}_{w_r^n}^{\mathcal{J}}=\emptyset$, and $\mathcal{M}_{w_r^n}^{\mathcal{K}}\cup\mathcal{M}_{w_r^n}^{\mathcal{J}}=\mathcal{M}$. So the discriminant functions for each string set can be formulated as:

$$g_{\mathcal{K}}(\theta) = \log\left[\frac{1}{|\mathcal{M}_{w_r^n}^{\mathcal{K}}|}\sum_{W\in\mathcal{M}_{w_r^n}^{\mathcal{K}}} p_\theta^\alpha(X_r|W)\cdot p^\alpha(W)\right]^{1/\alpha}, \tag{5}$$

and

$$g_{\mathcal{J}}(\theta) = \log\left[\frac{1}{|\mathcal{M}_{w_r^n}^{\mathcal{J}}|}\sum_{W'\in\mathcal{M}_{w_r^n}^{\mathcal{J}}} p_\theta^\alpha(X_r|W')\cdot p^\alpha(W')\right]^{1/\alpha}. \tag{6}$$

And the misclassification measure related to the reference word w_r^n can be written as:

$$d_{w_r^n} = -g_{\mathcal{K}}(\theta) + g_{\mathcal{J}}(\theta). \tag{7}$$

Consequently, a loss function can be naturally chosen as:

$$\mathcal{L}(d_{w_r^n}) = \frac{1}{1+e^{-2\gamma(d_{w_r^n}+\xi)}}, \tag{8}$$

like in the case of string-level MCE.

3.2. Relation Between MWCE Loss Function and Word Error

Firstly, let us give an intuitive explanation of the MWCE loss function in Eq. (8). The definition of this loss is based on the nature of the DP strategy in decoding: If a reference word w_r^n is to be recognized correctly, the best DP string must pass through this word for the corresponding time frames. Conversely, if the best DP string fails to contain w_r^n in its correct position, we can say that a word-level recognition error has occurred. Comparing with string-level MCE, word-level MWCE focuses on, and only on a particular local interval of the DP strings. That is to say, a string is considered to belong to the correct or incorrect string set depending on if it contains the reference word w_r^n in a given local segment. Outside that segment, any word sequence is allowed. It can be argued that the definition of correct strings in Eq. (4) is somewhat too strict, i.e., correct strings must contain a word that exactly matches the reference word. Although we follow this definition in all of our experiments in this study, one may loose this constraint to allow certain degree of differences in time alignment.

Secondly, let us try to give a theoretical analysis of the MWCE loss function. Again like in the case of string-level MCE, consider an extreme case when $\alpha\to\infty$ and $\gamma\to\infty$, the misclassification measure in Eq. (7) will become:

$$\begin{aligned} d_{w_r^n} \;=\; & -\log\max_{W\in\mathcal{M}_{w_r^n}^{\mathcal{K}}} p_\theta(X_r|W)\cdot p(W) \\ & +\log\max_{W'\in\mathcal{M}_{w_r^n}^{\mathcal{J}}} p_\theta(X_r|W')\cdot p(W'), \end{aligned} \tag{9}$$

and the loss function in Eq. (8) becomes a step function of:

$$\mathcal{L}(d_{w_r^n}) = \begin{cases} 0 & \text{if } d_{w_r^n} < 0 \\ 1 & \text{if } d_{w_r^n} > 0 \end{cases} . \quad (10)$$

Since the DP procedure in decoding will automatically choose the string with highest probability as recognition output, for any reference word w_r^n, the best DP output, say W^*, will either belong to $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ or $\mathcal{M}_{w_r^n}^{\mathcal{J}}$. Therefore, depending on the sign of $d_{w_r^n}$, the word-level MWCE loss $\mathcal{L}$ can be discussed in the following two cases:

Case 1. $d_{w_r^n} < 0 \Rightarrow \mathcal{L} = 0$: In this case, the best string in $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ has a higher probability than the best string in $\mathcal{M}_{w_r^n}^{\mathcal{J}}$. Consequently, $W^* \in \mathcal{M}_{w_r^n}^{\mathcal{K}}$. Recalling how we choose $\mathcal{M}_{w_r^n}^{\mathcal{K}}$ in Eq. (4), W^* will contain a "matched" word for w_r^n by definition. This case means that the reference word w_r^n is to be recognized correctly, and the MWCE loss in Eq. (10) is 0 accordingly.

Case 2. $d_{w_r^n} > 0 \Rightarrow \mathcal{L} = 1$: In this case, $W^* \in \mathcal{M}_{w_r^n}^{\mathcal{J}}$. As a result, W^* will not contain any "matched" word for w_r^n. This case means w_r^n could not be recognized correctly, so the MWCE loss in Eq. (10) will be 1.

Based on the two cases analyzed above, it is relatively easy to see that we are trying to design the MWCE loss function so as to approximate the number of word errors in training set. Optimizing model parameters with respect to this word-level MWCE loss better matches our ultimate goal of WER than string-level MCE. Please note that the discussion above is only based on an ideal case when α and $\gamma \to \infty$. In practice, however, α and γ are usually set to relatively smaller values (the same like in string-level MCE). This will take more competing strings into account, which is believed to be able to improve generalization.

3.3. MWCE in Unified DT Criterion

It is quite straightforward to embed our MWCE loss function into the unified DT criterion. If we set $\gamma = \alpha\rho$ in Eq. (8), and ξ to cancel the number of alternative strings in Eqs. (5) and (6), a new criterion in the unified form can be rearranged as the sum of smoothed word errors on the training data:

$$\mathcal{F}_{\text{MWCE}} = \sum_{r=1}^{R}\sum_{n=1}^{N_r} f\left(\log \frac{\sum_{W\in\mathcal{M}_{w_r^n}^{\mathcal{K}}} p_\theta^\alpha(X_r|W)\cdot p^\alpha(W)}{\sum_{W'\in\mathcal{M}_{w_r^n}^{\mathcal{J}}} p_\theta^\alpha(X_r|W')\cdot p^\alpha(W')}\right), \quad (11)$$

in which $f(z) = -1/(1+e^{2\rho z})$ keeps unchanged as string-level MCE. Please note again that f is a negative loss function because we always maximize the unified criterion for parameter optimization.

4. EXPERIMENTS

4.1. Implementation Details

The *Hidden Markov Model Toolkit* (HTK) implemented a unified DT criterion including MMI, MWE and MPE in its latest release [10]. The toolkit uses an initial model trained using MLE, to generate two sets of lattices (the so called "numerator" and "denominator" lattices) for discriminative training. Model parameters are updated using Extended Baum-Welch (EB) algorithm [11], so the traing process can be done in a parallel mode when a cluster of processors is available.

We extend the HTK implementation to support conventional string-level MCE and our proposed word-level MWCE. Because all criteria share most of the implementation details, we believe this unified DT framework can provide a reasonably fair evaluation and comparison of all criteria.

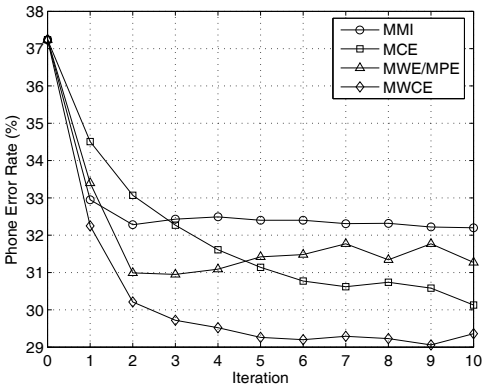

Fig. 1. Phone error rate on TIMIT database

Criterion	MLE	MMI	MCE	MW(P)E	MWCE
PER(%)	37.24	32.20	30.13	30.95	29.06
Relative(%)	-	13.5	19.1	16.9	22.0

Table 1. Phone Error Rate (PER) on TIMIT database, and relative improvement over MLE baseline.

String-level MCE only needs to accumulate statistics once for each training utterance. However, MWCE has to accumulate statistics once for each reference word (ref. to Eqs. (3) and (11)). This will make MWCE much more time-consuming in training than MCE. In our implementation of MWCE, for each reference word w_r^n, only the statistics within its local interval are calculated and accumulated. This simplification therefore reduces the training time to a comparable level of string-level MCE.

"I-smoothing" is used in our experiments to improve generalization. For MMI, MWE and MPE criteria, the i-smoothing factor τ is set to the recommended values in [10]. For MWCE criteria, τ is set according to the denominator counts, as suggested in [12]. Finally, the weighting exponent α is set to $1/15$ in all of our experiments, with a smoothing factor $\rho = 0.04$ as in [9].

4.2. Experimental Results

4.2.1. Experiments on TIMIT phone recognition task

Although not being an LVCSR task, the TIMIT phone recognition task focuses on pure acoustic modeling, and provides us an efficient way to evaluate new DT criteria. Our experimental conditions are close to that of [13]. The standard 3696 training and 192 core-test sets are used. 48 phones are chosen to train tri-phone HMMs, and they are then folded to 39 phones when calculating results. We obtain a total number of 990 tied-states in our system, and each state is modeled using an 8-component Gaussian mixture. A phone-loop network is used in decoding (no language model used). The phone recognition accuracy of the initial MLE model is 62.76%, which is comparable with [13].

For discriminative training, the "numerator" lattices are obtained according to manual labels. The "denominator" lattices are generated using the same network in decoding. Because of the phone recognition task, MWE criterion becomes equivalent to MPE, and MWCE is actually conducted on phone-level.

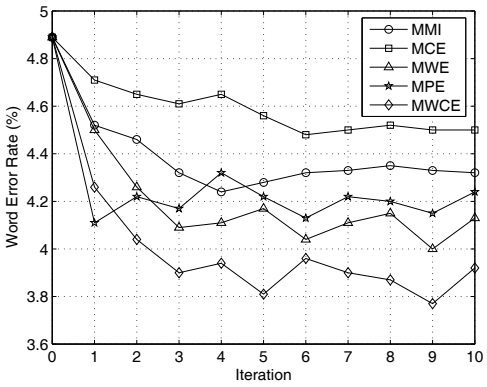

Fig. 2. Word error rate on WSJ0 Nov'92 5k evaluation

Criterion	MLE	MMI	MCE	MWE	MPE	MWCE
WER(%)	4.89	4.24	4.48	4.00	4.11	3.77
Relative(%)	-	13.3	8.4	18.2	16.0	22.9

Table 2. WER on WSJ0 Nov'92 5k evaluation, and relative improvement over MLE baseline.

The phone recognition error rates of each iteration for the five criteria are shown in Fig. 1. We also compared the best recognition performance for each criterion, as given in Table. 1. The experimental results suggest that MCE and MWCE criteria outperform other three criteria on this task. And the proposed word-level MWCE criterion can achieve an improved performance over string-level MCE, with a relative error rate reduction of 22.0%.

4.2.2. Experiments on WSJ0 Nov'92 5k evaluation

To evaluate our MWCE criterion on an LVCSR task, experiments are carried out on WSJ0 database. The training corpus is SI-84 set, with 7133 utterances from 84 speakers. Evaluation is performed on standard Nov'92 non-verbalized 5k closed vocabulary test set, with 330 utterances from 8 speakers. The training setups are similar to the WSJ HTK recipe proposed in [14] (a system that performs similar to [15]). Cross-word tri-phone HMMs with a total number of 2774 tied-states are trained, and each state has 8 Gaussian components. The WERs of the MLE baseline using standard bi-gram and tri-gram language models are 7.34% and 4.89%, respectively. These results are comparable with the numbers reported in [14].

For discriminative training, a weakened, uni-gram language model is used to generate lattices. The recognition performances of the five criteria when decoded using the standard tri-gram language model are given in Fig. 2 and Table. 2. We observe that MWE outperforms MPE on this task, which is consistent with [3]. String-level MCE only achieves a relative error rate reduction of 8.4% on this task, which is much lower than that of MMI, MWE, and MPE. Comparing with other criteria, our proposed MWCE criterion again achieves the best recognition performance, with a relative error rate reduction of 22.9%.

5. CONCLUSIONS AND FUTURE WORK

We proposed a new discriminative training criterion MWCE in this paper. By localizing conventional string-level MCE loss function to word-level, a more direct objective which better matches the performance evaluation measures (such as WER) can be derived. Both intuitive explanation and theoretical analysis were carried out in this study, to investigate the relationship between MWCE criterion and word classification error. Finally, MWCE was embedded into the unified DT criterion and evaluated with other commonly-used criteria. Experimental results on TIMIT and WSJ0 tasks suggested that consistent performance improvement can be obtained by MWCE over conventional string-level MCE. MWCE also outperformed MMI and other substring-level criteria MWE / MPE, on the above two tasks. To evaluate MWCE criterion on larger and more difficult tasks will be our future work.

6. REFERENCES

[1] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Proc. ICASSP1986*, 1986, vol. 1, pp. 49–52.

[2] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, 1992.

[3] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2004.

[4] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum error rate training based on n-best string models," in *Proc. ICASSP1993*, 1993, vol. 2, pp. 652–655.

[5] E. McDermott and S. Katagiri, "Minimum error training for speech recognition," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, 1994, pp. 259–268.

[6] E. McDermott and S. Katagiri, "String-level MCE for continuous phoneme recognition," in *Proc. EuroSpeech1997*, 1997, vol. 1, pp. 123–126.

[7] Q. Fu, X. He, and L. Deng, "Phone-discriminating minimum classification error (P-MCE) training for phonetic recognition," in *Proc. InterSpeech2007*, 2007, pp. 2073–2076.

[8] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, vol. 34, pp. 287–310, 2001.

[9] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Proc. EuroSpeech2005*, 2005, pp. 2133–2136.

[10] S. Young, et al., *The HTK Book*, 2006, Revised for HTK version 3.4.

[11] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem*, Ph.D. thesis, McGill University, 1991.

[12] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to MPE for large scale discriminative training," in *Proc. ICASSP2007*, 2007, vol. 4, pp. 321–324.

[13] S. Young, "The general use of tying in phoneme-based hmm speech recognisers," in *Proc. ICASSP1992*, 1992, vol. 1, pp. 569–572.

[14] K. Vertanen, "HTK Wall Street Journal (WSJ) training recipe," http://www.inference.phy.cam.ac.uk/kv227/htk/.

[15] P. Woodland, J. Odell, V. Valtchev, and S. Young, "Large vocabulary continuous speech recognition using HTK," in *Proc. ICASSP1994*, 1994, vol. 2, pp. 125–128.