DISCRIMINATIVE TRAINING BASED ON THE CRITERION OF LEAST PHONE COMPETING TOKENS FOR LARGE VOCABULARY SPEECH RECOGNITION

Bo Liu[12], Hui Jiang[3], Jian-Lai Zhou[1],Ren-Hua Wang[2]

[1]Microsoft Research Asia, Beijing, P. R. China
[2]Department of EEIS, University of Science and Technology of China, Hefei, Anhui, P. R. China
[3]Department of Computer Science and Engineering, York University, Toronto, CANADA
Email: liubo@ustc.edu, hj@cs.yorku.ca, jlzhou@microsoft.com, rhw@ustc.edu.cn

ABSTRACT

In this paper, we propose a new discriminative training approach based on the criterion of least phone competing tokens. In our approach, we first collect a competing token set for each physical HMM from training data. Different from the previous in-search token selection, an off-line token collection procedure is used in this work to collect the competing-tokens from word lattices. Then we re-estimate HMM parameters discriminatively to minimize the total number of competing tokens counted in the phone level. The phone token counts are approximated by a sigmoid-based objective function. The GPD algorithm is used to adjust HMM parameters to minimize the objective function. In this work, a merging mechanism and a gradient normalization in the HMM tied-state level are proposed to improve the generalization power of our discriminative training method. The proposed method is evaluated on the Resource Management (RM) and the Switchboard (a 24-hr mini-train set) tasks. Experimental results clearly show that our new discriminative training method achieves significant improvements over our best MLE models in both tasks, namely about 8% and 4.5% relative error rate reduction in RM and Switchboard respectively, over the best MLE models.

1. INTRODUCTION

It's well known that discriminative training (DT) is an important and effective approach to improve the performance of speech recognition system. The discriminative training has been extensively studied for HMM-based automatic speech recognition (ASR), such as maximum mutual information (MMI) estimation [6,7,8] and minimum classification error (MCE) method [4,5,3]. Discriminative training has been found quite effective to improve ASR performance over the ML method in small or medium vocabulary ASR tasks (see [5,6]). But no performance gain has been demonstrated in any large-scale ASR tasks until very recently. In [8], MMI method was applied to the switchboard task and some moderate improvements was consistently observed while in [2,3] the MCE/GPD method was extended to the DARPA communicator task and a slight gain was also achieved over the best MLE (maximum likelihood estimation) HMMs. In [2,3], a dynamic data selection algorithm is used to collect competing tokens from the ASR decoding process. The collected tokens contain competing information about the original HMM set and thus they can be used to improve acoustic models. In [2,3], a word-level criterion, i.e., the smoothed count of imposter words, was proposed as the objective function for the GPD-based optimization. In this paper, we adopt an off-line token collection approach for convenience in implementation. For each HMM, two different token sets, namely the competing token (CT) set and true token (TT) set, are collected from the word lattices, which are generated by a regular decoder prior to the data selection stage. In this work, we propose a new discriminative training criterion in the phone level, namely the least phone competing token (LPCT) criterion. We formulate the smoothed count of all collected competing tokens as our objective function. Then the GPD algorithm is used to minimize the function by adjusting HMM parameters. Moreover, a merging mechanism and a gradient normalization scheme are proposed to improve the performance of our discriminative training method. Experimental results on two large vocabulary tasks, namely the DARPA RM and Switchboard, demonstrate the effectiveness of the proposed DT method in large-scale ASR tasks.

The remainder of this paper is organized as follows. In section 2 we present the off-line token collection method after giving the definition of TT and CT. The criterion for the discriminative training and the corresponding GPD optimization procedure are given in section 3. In section 4, experimental results on the RM and Switchboard database are reported and discussed. Finally, we conclude the paper with our findings.

2. COMPETING TOKEN COLLECTION

2.1 Definition of TT and CT

In speech recognition, given a speech segment X as input, the recognizer usually gives a unit a as output (The unit a may range from a phone to a sentence, but in this paper it always denotes a phone). If X's true transcription matches phone a, then X is called a true token of phone a. Otherwise, it is a competing token of a. So we give the definition of the CT set of a as

$$S_c(a) = \{Y \mid P_r(Y \mid a) > P_r(Y \mid a'), \forall a' \neq a, \; Y \notin a, P_r(Y \mid a) \geq \xi\} \quad (1)$$

and the definition of the TT set of a is given as

$$S_T(a) = \{X \mid P_r(X \mid a) > P_r(X \mid a'), \forall a' \neq a, \; X \in a\} \quad (2)$$

Where $Y \notin a$ denotes Y doesn't correspond to phone a, $Y \notin a$ denotes Y's label matches with a, and $P_r(a \mid Y) \geq \xi$ means any

competing token with too small observation probability are excluded.

2.2 Off-line Token Collection

Given the definition of TT and CT sets, a token collection procedure should be designed to automatically collect the two token sets from speech data. In [2], an on-line token collection procedure was used. However, in this paper, we propose an off-line token selection method to avoid the difficulty of modifying an existing decoder.

In our off-line token method, the collection process of TTs is relatively simple: Firstly, the forced-alignment is performed on every utterance in the training set to generate its reference segmentation by using the best MLE trained models. Then every phone segment, say a, in the reference segmentation is simply treated as a TT of a. As for the CT set, the collection process is a little complex. Firstly, a word lattice is generated for every utterance in the training set. Secondly, for every word arc in the word lattice, we perform a phone-level forced-alignment to obtain the phone boundaries within each word arc based on the original MLE tri-phone HMMs used for lattice generation. In this way, the word-lattice is converted into a tri-phone lattice. Finally, we use the token selection method in [2,3] to decide whether each phone arc in the tri-phone lattice is a CT or not by comparing its phone id and boundary information with the reference segmentation: if the maximum overlap of the arc with all phone segments (with the same phone id) in the reference segmentation exceeds a preset threshold AND the difference of the log likelihood score of this arc and its corresponding reference model is larger than another pre-set threshold, then this arc, i.e., the phone segment together with its id, to say a, and boundary registration, is classified as a CT of phone a. The above selection procedure is repeated for all utterances in the training data set to obtain the whole CT sets for all distinct phones in the system.

3. DISCRIMINATIVE TRAINING

Once the TT and CT sets are collected, we can adjust original acoustic models to improve their discrimination capability to improve speech recognition performance. In this work, a GPD (generalized probabilistic descent) algorithm based discriminative training is employed to minimize the objective function formulated according to the criterion of Least Phone Competing Tokens (LPCT).

At first, we assume that each phone is modeled by an N-state CDHMM, and the state (e.g. state i) observation p.d.f. is assumed to be a mixture of multivariate Gaussian distribution with diagonal precision matrix:

$$\begin{aligned} p(x \mid \theta_i) &= \sum_{k=1}^{K} \omega_{ik} N(x \mid m_{ik}, r_{ik}) \\ &= \sum_{k=1}^{K} \omega_{ik} \prod_{d=1}^{D} \sqrt{\frac{r_{ikd}}{2\pi}} e^{-\frac{1}{2} r_{ikd} (x_d - m_{ikd})^2} \end{aligned} \tag{3}$$

Where K denotes the number of Gaussian mixtures in each state, D is the dimension of feature vector and $\theta_i = \{\omega_{ik}, m_{ik}, r_{ik}\}_{k=1,2,\cdots,K}$ is mixture parameters for state i.

For simplicity, only mean vectors of CDHMM are updated in this paper and all other parameters remain constant in our GPD-based discriminative training process.

3.1 The Criterion of Least Phone Competing Tokens (LPCT)

The tokens in the collected CT sets can be viewed as the strong competitors of the correct hypothesis in the phone level, which could cause potential recognition errors since they have relatively large likelihood score. Intuitively, we could choose the objective function in discriminative training as the total number of the phone CTs as counted by the current HMM set. If we can minimize the objective function with respect to the HMM parameters, it is very likely that the final word error rate (WER) of recognition will be reduced. This criterion is named as *least phone competing tokens (LPCT)*. Here, we follow the idea in the MCE to formulate a smoothed count of phone competing tokens (PCT) as follows:

Suppose we have a CT , say Y, of phone a, which is a speech segment of total T frames. Then the misclassification measure for Y can be defined as

$$d_a = \frac{1}{T}[l(\vec{Y} \mid \Lambda_a) - l(\vec{Y} \mid \Lambda_{ref}(\vec{Y}))] \tag{4}$$

Where $l(\cdot)$ denotes log likelihood function. $\Lambda_{ref}(\vec{Y})$ stands for the reference model for the segment according to the optimal Viterbi path obtained in force-alignment against reference transcription, and Λ_a is the HMM for phone a. If the misclassification measure $d_a > 0$ then this token is counted as one phone competing token (PCT). Next, the above d_a is plugged into a sigmoid function to approximate the zero-one decision in the actual counting of the phone competing tokens:

$$\ell(d_a) = \frac{1}{1+\exp(-\gamma \cdot d_a + \vartheta)} \tag{5}$$

Where γ and θ are the parameters to control the shape of sigmoid function. In this way, $\ell(d_a)$ can be viewed as the "smoothed" count of phone competing token for Y. Finally, the total number of competing tokens can be calculated by summarizing the above smoothed count over all collected tokens in the whole CT set, $S_c(a)$, as follows:

$$L(\vec{\Lambda}) \propto \sum_{a \in S_c} \ell(d_a) \tag{6}$$

In this work, $L(\vec{\Lambda})$ is the objective function in our discriminative training method, which represents a smoothed count of total number of phone competing tokens (PCT). Thus, the HMM parameters $\vec{\Lambda}$ are optimized to minimize it.

3.2 GPD Optimization

The GPD algorithm is adopted to minimize the objective function $L(\vec{\Lambda})$. Only those HMM parameters related to competing tokens are adjusted. Suppose $Y(a)$ is a competing token of phone a, the GPD will be used to adjust the HMM model Λ_a and its reference model, denoted as $\Lambda_{ref}(\vec{Y})$.

Given a competing token $Y(a) = \{y_1, y_2, \cdots, y_T\}$ in the CT set $S_c(a)$ of phone a, and $\{s_1, s_2, \cdots, s_T\}$ is assumed to be its

corresponding optimal Viterbi state sequences in the model Λ_a, Then the gradient for each mean vector $\{m_{ik} \mid 1 \le i \le N, 1 \le k \le K\}$ is accumulated over all collected tokens in the CT set as follows:

$$G_{ik} = \sum_{Y \in S_c(a)} \frac{\partial \ell(d_a)}{\partial m_{ik}} \Big|_{m_{ik}=m'_{ik}} = \sum_{Y \in S_c(a)} \gamma \cdot \ell(d_a)(1-\ell(d_a)) \cdot \frac{1}{T} \sum_{t=1}^{T} \xi_{ikt} \cdot (y_t - m'_{ik}) \cdot \delta(s_t - i) \quad (7)$$

Where i and k are indices of state and mixture component respectively, $\delta(\cdot)$ denotes the Kronecher delta function, and ξ_{ikt} is the mixture component occupation rate:

$$\xi_{ikt} = (b_i(\vec{y}_t))^{-1} \cdot \omega_{ik} \prod_{d=1}^{D} \sqrt{\frac{r_{ikd}}{2\pi}} \exp\{-\frac{1}{2} r_{ikd} (y_{td} - m_{ikd})^2\} \quad (8)$$

$$b_i(\vec{y}_t) = \sum_{k=1}^{K} \omega_{ik} \prod_{d=1}^{D} \sqrt{\frac{r_{ikd}}{2\pi}} \exp\{-\frac{1}{2} r_{ikd} (y_{td} - m_{ikd})^2\} \quad (9)$$

where $b_i(\vec{y}_t)$ is the emission probability of the frame t.

As for the reference model, we assume $\{\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_T\}$ denotes the optimal state path in it. The gradient for each mean vector $\{\bar{m}_{ik} \mid 1 \le i \le M, 1 \le k \le K\}$ is calculated as:

$$\bar{G}_{ik} = \sum_{Y \in S_c(a)} \frac{\partial \ell(d_a)}{\partial \bar{m}_{ik}} \Big|_{\bar{m}_{ik}=\bar{m}'_{ik}} = \sum_{Y \in S_c(a)} \gamma \cdot \ell(d_a)(1-\ell(d_a)) \cdot \frac{1}{T} \sum_{t=1}^{T} \bar{\xi}_{ikt} \cdot (y_t - \bar{m}'_{ik}) \cdot \delta(\bar{s}_t - i) \quad (10)$$

$\bar{\xi}_{ikt}$ is the mixture component occupation rate:

$$\bar{\xi}_{ikt} = (\bar{b}_i(\vec{y}_t))^{-1} \cdot \bar{\omega}_{ik} \prod_{d=1}^{D} \sqrt{\frac{\bar{r}_{ikd}}{2\pi}} \exp\{-\frac{1}{2} \bar{r}_{ikd} (y_{td} - \bar{m}_{ikd})^2\} \quad (11)$$

$$\bar{b}_i(\vec{y}_t) = \sum_{k=1}^{K} \bar{\omega}_{ik} \prod_{d=1}^{D} \sqrt{\frac{\bar{r}_{ikd}}{2\pi}} \exp\{-\frac{1}{2} \bar{r}_{ikd} (y_{td} - \bar{m}_{ikd})^2\} \quad (12)$$

where $\bar{b}_i(\vec{y}_t)$ is the emission probability of the frame t.

Finally, the mean vectors of the phone model of a and the reference model are updated as follows:

$$m_{ik}^{(t+1)} = m_{ik}^{(t)} - \varepsilon_1 \cdot \frac{1}{N_i} \cdot G_{ik} \quad (13)$$

$$\bar{m}_{ik}^{(t+1)} = \bar{m}_{ik}^{(t)} + \varepsilon_2 \cdot \frac{1}{\bar{N}_i} \cdot \bar{G}_{ik} \quad (14)$$

where N_i and $\bar{N}_i$ denote the total number of distinct physical frames used to compute the gradient for state i in the above gradient accumulation stage as in eq.(7) and eq.(10). It is found that the dynamic range of N_i and $\bar{N}_i$ is so large that the adjustment for the mean vectors of some states may be already too large while others virtually remain constant if we don't normalize the gradient as in the conventional MCE/PGD learning. Therefore, we believe it is very important to perform such a gradient normalization based on the accumulated frame numbers N_i and $\bar{N}_i$, especially when we update a large HMM set as in large-scale discriminative training. We believe this is a critical technique to make the GPD work for a large HMM set.

Moreover, it is likely that the same physical frame is assigned to different phone CTs during our token selection process. But this physical frame actually corresponds to the same physical HMM tied-state in the optimal Viterbi paths of these different CTs. We also find that it is necessary to merge these physically identical frames corresponding to the same tied-state as appearing from different tokens. Otherwise, the gradient value of a particular HMM mean from the same data frame could be accumulated more than once in eq.(7) or eq.(10). After merging, it is guaranteed that each physical frame will be used to update a physical tied-state for at most once.

The above optimization process can be repeated to improve the model discrimination capability iteratively. Every newly updated model is used as the starting model for next iteration. In theory, the token collection procedure should also be repeated by using the updated model for every iteration. However, the token collection process is so time-consuming that we come to a compromise to keep the token sets unchanged during the whole GPD training procedure.

4. RECOGNITION EXPERIMENTS

We evaluate our method on the RM and Switchboard (a 24-hr minitrain) tasks. In both tasks, the above discriminative training starts from our best MLE and performs several iterations to minimize the objective function in eq.(6) w.r.t. all HMM mean vectors.

4.1 Resource Management Experiments

For the RM experiments, we use the standard training set, including 3979 utterances, and the test set with 1199 utterances. The speech feature used here is Mel-frequency cepstral coefficients (MFCCs) and the log energy, together with their first and second differentials, in total 39 dimensions. The best MLE model is a set of tied-state cross-word tri-phones HMMs, which 1605 distinct tied-states with 6 Gaussian per state. We use the standard word-pair grammar in decoding.

iteration	Training set		Test set	
	WER(%)	Err Red	WER(%)	Err Red
0 (ML)	1.26	N/A	4.30	N/A
1	1.19	8%	4.16	3%
5	1.06	16%	**3.96**	**8%**
6	1.03	18%	4.06	6%

Table 1: Results on RM database

Firstly, the off-line token collection procedure in section 2 is conducted on the whole training set. Then GPD optimization is repeated for 6 iterations using the collected token sets. Parameters of sigmoid function $\gamma = 0.5, \theta = 0.0$ and step size $\varepsilon_1 = 0.006, \varepsilon_2 = 0.006$ are used for RM database. The curve of objective function on RM is drawn in left part of Figure 1. The curve descends in the iterative GPD training just as we expect. On right part of the Figure 1, Word Error Rate (WER) curves on both training and test sets of RM database are drawn respectively. Both curves decrease as the iteration goes on. The WER's are also partially shown in Table 1. The best results achieved are 1.03% WER on training set and 3.96% WER on test set while their baseline results with the best MLE are 1.28% and 4.30% respectively. We have observed about 18% and 8% WER reduction in training and test sets respectively after our discriminative training method.

4.2 Switchboard Experiments

For the Switchboard task, the speech feature used is perceptual linear prediction (PLP) coefficients and the log energy, together with their first and second order derivatives. The baseline MLE model is also tied-state cross-word tri-phones HMMs, which include 1280 distinct tied-states with 8 Gaussian per state. The MLE mode is trained on a mini-train set which includes 18266 utterances (totally 24-hr). We use 60K lexicon and a tri-gram language in decoding. The test set is the eval2000 set, including 1831 utterances. The recognition performance on training set is evaluated on a subset of training set composed of 1803 utterances randomly selected from the whole training set. The sigmoid function parameters $\gamma = 1.0, \theta = 0.0$ and step size $\varepsilon_1 = 0.01, \varepsilon_2 = 0.01$ are used in the Switchboard task. Figure 2 plots the objective function curve and the WER for both the training and the test sets as a function of the number of iterative discriminative training procedure. The WER's are also partially shown in Table 2. After six iterations, the new recognizer achieved the best results of 29.0% WER on training set and 46.0% on test set, while their baseline results (with the MLE models) are 33.2% and 48.1%. We have observed roughly 13% and 4% WER reduction in training and test sets respectively.

iteration	Training set		Test set	
	WER(%)	Err Red	WER(%)	Err Red
0 (ML)	33.2	N/A	48.1	N/A
1	31.5	5%	47.1	2%
4	29.4	11%	**46.0**	**4%**
6	29.0	13%	46.4	4%

Table 2: Results on Switchboard task

All the above results show that our new discriminative training method can improve the recognition performance significantly even in large vocabulary ASR systems.

5. CONCLUSION

In order to improve the performance of large vocabulary speech recognition system, we propose a discriminative training method based on the criterion of least phone competing tokens (LPCT). An off-line token collection approach is used to collect competing tokens from speech data. A new merging scheme and a gradient normalization in the GPD algorithm are also presented in this paper. The experimental results on the RM and Switchboard tasks clearly demonstrate the effectiveness of our approach. As our future work, there are still many promising aspects that need more investigation, such as variance normalization, quick line search in GPD, more new criteria in word level (e.g., least word competing tokens).

6. ACKNOWLEDGEMENTS

We appreciate Frank Seide, Li-Sheng Fan, Gang Guo in speech group, MSRA for their helps in experimentation. We also thank Chao-Jun Liu at York University for the tool to collect tokens.

REFERENCES

[1] L.R. Bahl, P.F. Brown, P.V. De Souza and R.L. Mercer, ''Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. of ICASSP-86*, pp.49-52, Tokyo, Japan, 1986.

[2] H. Jiang, F. Soong and C.-H. Lee, ''A dynamic in-search data selection method with its applications to acoustic modeling and utterance verification," *to appear in IEEE Trans. on Speech and Audio Processing*, 2004.

[3] H. Jiang, O. Siohan, F. Soong and C.-H. Lee, "A dynamic in-search discriminative training approach for large vocabulary speech recognition," *Proc. of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002)*, pp.I-113-116, Orlando, Florida, May 2002.

[4] B.-H. Juang and S. Katagiri, ``Discriminative Learning for Minimum Error Training," *IEEE Trans. on Acoustic, Speech, Signal Processing*, Vol. 40, pp.3043-3054, No. 12, Dec. 1992.

[5] B.-H. Juang, W. Chou and C.-H. Lee, ``Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, pp.257-265, Vol.5, No.3, May 1997.

[6] Y. Normandin, R. Cardin and R. Demori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, Apr. 1994.

[7] D. Povey and P.C. Woodland, ``Minimum Phone Error and I-smoothing for improved discriminative training," Proc. of ICASSP 2002, pp.I105-I108.

[8] P.C. Woodland and D. Povey, ``Large Scale Discriminative Training of hidden Markov models for speech recognition," *Computer Speech & Langauge*, pp.25-47, Vol. 16, No. 1, January 2002.

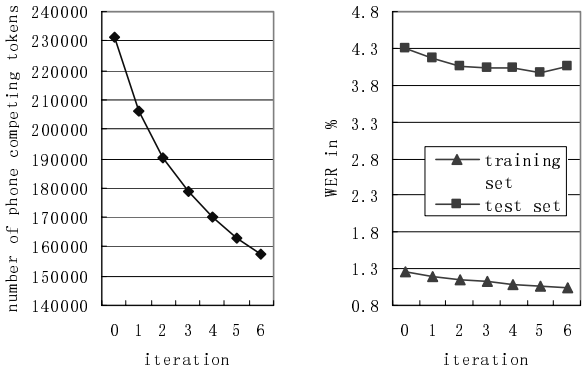

Figure 1: Objective function and WER curves on RM

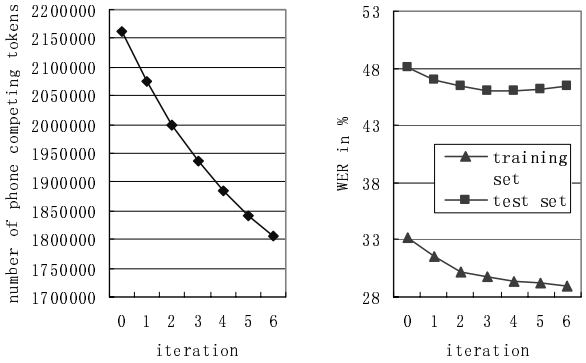

Figure 2: Objective function and WER curves on the Switchboard task