A Comparative Study of RPCL and MCE Based Discriminative Training Methods for LVCSR

Zaihu Pang[1], Xihong Wu[1], and Lei Xu[1,2]

[1] Speech and Hearing Research Center, Key Laboratory of Machine Perception, (Ministry of Education), Peking University
{pangzh,wxh}@cis.pku.edu.cn, lxu@cse.cuhk.edu.hk
[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong

Abstract. This paper presents a comparative study of two discriminative methods, i.e., Rival Penalized Competitive Learning (RPCL) and Minimum Classification Error (MCE), for the tasks of Large Vocabulary Continuous Speech Recognition (LVCSR). MCE aims at minimizing a smoothed sentence error on training data, while RPCL focus on avoiding misclassification through enforcing the learning of correct class and de-learning its best rival class. For a fair comparison, both the two discriminative mechanisms are implemented at state level. The LVCSR results show that both MCE and RPCL perform better than Maximum Likelihood Estimation (MLE), while RPCL has better discriminative and generative abilities than MCE.

Keywords: Rival penalized competitive learning, minimum classification error, discriminative training, large vocabulary continuous speech recognition.

1 Introduction

In recent years, Discriminative Training (DT) methods significantly improve the performance of speech recognition. The success of DT methods for large-scale tasks relies on three key ingredients. The first one is the formulation of a DT criterion. The most widely used DT criteria include Maximum Mutual Information (MMI) [1], and a class of error minimizing discriminative training criteria: Minimum Classification Error (MCE) [2] and Minimum Word/Phone Error (MWE/MPE)[3]. The second ingredient is the use of lattice-based competing space, which provides more competing paths and avoids reduplicative computation on the same word in different strings, when comparing with traditional string based competing space. The third ingredient is to adopt the widely used Extended Baum-Welch(EBW) algorithm for parameter estimation. An overview of these methods are referred to [4]. Recently, Rival Penalized Competitive Learning (RPCL) was introduced in [5] to speech recognition with promising results in a comparison with MMIE and MPE. However, there is still a lack of comparison between RPCL and MCE. This paper is motivated for such a comparative study.

Y. Zhang et al. (Eds.): IScIDE 2011, LNCS 7202, pp. 27–34, 2012.

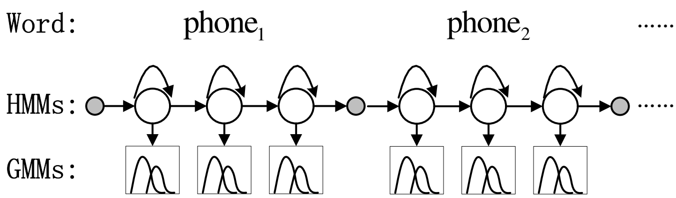

Fig. 1. The hierarchical structure of word in GMM-HMM based speech recognition: word level, phone level(HMM) and state level(GMM)

MCE criteria was first proposed in [2], which aims at minimizing the expectation of a smoothed string error on training data. The MCE discriminant function can be generalized to model word strings, phones, and other levels in speech recognition. In an early study [6], the string-level MCE was shown to have similar performance with MMIE based method on small vocabulary tasks. In [7], phone-level based MCE was used for the acoustic model training of a continuous phoneme recognition task, which turned out to be more effective than string-level based MCE. Moreover, studies in recent years [8,9] investigated lattice-based MCE methods, which have comparative performance with MPE based method on the large vocabulary tasks.

First proposed in 1992 [10,11], RPCL is a further development of competitive learning on a task of multiple classes or models that compete to learn samples. For each sample, the winner learns while its rival (i.e., the second winner) is repelled a little bit from the sample, which reduces a duplicated sample allocation such that the boundaries between models become more discriminative. In [5], RPCL was implemented on hidden Markov states for a discriminative Hidden Markov Model (HMM) based speech model as shown in Fig.1. Only the best rival state of the correct state is repelled, which increases its discriminative ability and obtains preferable generalization ability. When applied to LVCSR, it showed improved generalization performance than the MMIE and MPE, especially when the sources of test sets are different from the training set.

This paper follows [5] to present a comparison between RPCL and MCE as discriminative training methods using state level competing space for LVCSR task. Referring to [5], RPCL is implemented at the state level. For a fair comparison, according to [9], MCE is also derived to be implemented at state level. Experiments are conducted on large vocabulary continuous speech recognition task: 863-I-Test (matched with training data) and Hub-4-Test (unmatched with training data). The results show that the RPCL based method has better discriminative and generative abilities than MCE based method on the test data either matched or unmatched with train data.

The rest of this paper is organized as follows: In Section 2, state-level RPCL is reviewed. In Section 3, MCE using state level competing space is briefly introduced. In Section 4, experimental results of RPCL and MCE are presented. Finally, conclusions are made in Section 5.

2 Rival Penalized Competitive Learning

First proposed in 1992 [10,11] and further developed subsequently, RPCL is a competitive-learning-featured, general problem-solving framework, for multi-learners or multi-agents with each to be allocated to learn one of multiple structures underlying observations. Readers are referred to [12] for a systematic review and recent developments. In the following, we only provide a brief introduction.

We measure the error or cost for the j-th learner to describe the current input x_t by $\varepsilon_t(\theta_j) \geq 0$. The winner and rival are decided by Eq.(1), where c_t is called winner, and the second winner r_t is its rival. RPCL learning can be simply implemented by Eq.(2). Not only the parameter θ_{c_t} of the winner is learned such that $\varepsilon_t(\theta_{c_t})$ decreases to some extent, but also the parameter θ_{r_t} of the rival is de-learned such that $\varepsilon_t(\theta_{r_t})$ increases by a little bit. The rival penalized mechanism makes the boundaries between different learners or models become more discriminative. Readers are referred to Sect.3.1 and Sect.3.2 in [13] and particularly its Eq.(9)& Eq.(34) for further details.

$$p_{j,t} = \begin{cases} 1, & \text{if } j = c_t \\ -\gamma, & \text{if } j = r_t \\ 0, & \text{otherwise} \end{cases} \quad \begin{cases} c_t = \arg\min_j \varepsilon_t(\theta_j) \\ r_t = \arg\min_{j \neq c_t} \varepsilon_t(\theta_j) \end{cases} \tag{1}$$

$$\theta_j^{new} - \theta_j^{old} \propto p_{j,t} \nabla_{\theta_j} \varepsilon_t(\theta_j). \tag{2}$$

Making RPCL based discriminative learning on $p(x_t|\theta_j)$ across different hidden Markov states, for each state j, we have

$$\varepsilon_t(\theta_j) = -\ln p(x_t|\theta_j) \tag{3}$$

where $p(x_t|\theta_j) = \sum_{k=1}^{K} \alpha_{jk} \mathcal{N}(x_t|\mu_{jk}, \Sigma_{jk})$ is a mixture of Gaussian distributions $\mathcal{N}(x_t|\mu_{jk}, \Sigma_{jk})$ with mean μ_{jk} and covariance matrix Σ_{jk}. For every input, instead of getting $c_t = \arg\min_j \varepsilon_t(\theta_j)$, the state that corresponds to the identity of this input by the Viterbi force alignment is regarded as the winner state c_t. Still, we get the rival by $r_t = \arg\min_{j \neq c_t} \varepsilon_t(\theta_j)$. After the initialization of all the parameters $\{\theta_j\}$ by the MLE based BW algorithm, the parameter θ_j can be iteratively optimized by Eq.(2).

One problem for the RPCL learning is how to determine an appropriate penalizing strength γ that is usually set in a heuristic way. Favorably, the Bayesian Ying-Yang (BYY) harmony learning includes a mechanism of penalizing rivals. RPCL can be regarded as a rough approximation of the BYY harmony learning for learning a mixture of multiple models, while the BYY harmony learning provides a top-down guidance for choosing penalizing strength [14]. For a task of learning a mixture of multiple models to which our tasks belongs, the BYY harmony learning is implemented via maximizing $H(p||q, \theta)$ given in Eq.(8) of [12]. From its implementation by the flow $\nabla_\theta H(p||q, \theta)$, we observe its link to RPCL learning. Particularly, we consider the one given by Eq.(13) in [12], which is rewritten as

$$p_{j,t} = \begin{cases} p(c_t|x_t) + (1+\gamma_t)p(r_t|x_t), & j = c_t \\ -p(r_t|x_t)\gamma_t, & j = r_t \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Putting it into Eq.(2), we make a gradient based iterative implementation of this RPCL simplified BYY harmony learning. Moreover, it follows from Eq.(3) that we get $p_{j,t}\nabla_{\theta_{jk}}\varepsilon_t(\theta_j) = -p_{jk,t}\nabla_{\theta_{jk}} \ln p(x_t|\theta_{jk})$ with

$$p_{jk,t} = p_{j,t}p(k|x_t, \theta_j) \tag{5}$$

where $p(k|x_t, \theta_j) = \alpha_{jk}p(x_t|\theta_{jk})/[\sum_{i=1}^{K} \alpha_{ji}p(x_t|\theta_{ji})]$.

Alternatively, we may also make a batch way updating with the whole training set used. Particularly, considering a Gaussian $p(x_t|\theta_{jk}) = \mathcal{N}(x_t, \mu_{jk}, \Sigma_{jk})$, it follows from solving $\sum_{\mathrm{t=1}}^{\mathrm{T}} p_{jk,t}\nabla_{\theta_{jk}}\varepsilon_t(\theta_{jk}) = 0$ that we get

$$\begin{aligned} \alpha_{jk}^{new} &= \frac{\sum_{\mathrm{t=1}}^{\mathrm{T}} p_{jk,t}}{\sum_k \sum_{\mathrm{t=1}}^{\mathrm{T}} p_{jk,t}}, \mu_{jk}^{new} = \frac{\sum_{\mathrm{t=1}}^{\mathrm{T}} p_{jk,t}x_t}{\sum_{\mathrm{t=1}}^{\mathrm{T}} p_{jk,t}}, \\ \Sigma_{jk}^{new} &= \frac{\sum_{\mathrm{t=1}}^{\mathrm{T}} p_{jk,t}(x_t - \mu_{jk}^{new})(x_t - \mu_{jk}^{new})^T}{\sum_{\mathrm{t=1}}^{\mathrm{T}} p_{jk,t}}. \end{aligned} \tag{6}$$

Together with Eq.(5), we iterate the following Ying-Yang alternation that implements a RPCL type BYY harmony learning:

Yang-step: get RPCL-allocation by Eq.(5)
Ying-step: re-estimate Gaussian components by Eq.(6),

which has a same format as the classical EM algorithm and thus shares a similar computing complexity. The difference comes from the weights $p_{jk,t}$ via which the rival penalized mechanism is embedded.

While implementing the RPCL type BYY harmony learning by Eq.(4), getting $p_{jk,t}$ by Eq.(5) is already an approximation of the BYY harmony learning that leads to $p_{jk,t} = p_{j,t}p(k|x_t, \theta_j) + \delta_{jk,t}p(j|x_t)$ with $\delta_{jk,t}$ considering the winner enhancing and rival penalizing mechanism among the Gaussian components under the same state j. If the winner state c_t is considered reliable, we let $p(c_t|x_t) \approx 1$. The γ_t is considered as a small constant, $\gamma_t = \gamma$ and $\gamma < 0$. Then, Eq.(4) can be simplified to

$$p_{j,t} = \begin{cases} 1 + p(r_t|x_t), & j = c_t \\ -p(r_t|x_t)\gamma, & j = r_t \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where γ, playing a similar role as in Eq.(1), denotes the de-learning rate. The bigger the γ is , the stronger the de-learning is. The learning of the winner state c_t is enhanced, while its rival state r_t is de-learned with a de-learning rate γ. The strengths of enhancing and de-learning vary as the degree of the competition, namely the posterior probability of the rival state, which make the states more discriminative.

3 Minimum Classification Error

To compare with the state-level RPCL method [5], the MCE is also implemented at the state level. The competing space of state-level MCE is the same as that of state-level RPCL. The state-level MCE based method is derived by following [8,9]. Its loss function and discriminant function are introduced as follows.

Suppose the reference state sequence of the r_{th} training utterance consists of N_r states, i.e., $S_r = \{s_r^1, s_r^2, \cdots, s_r^{N_r}\}$. For each reference state s_r^n, its correct string set $M_{s_r^n}^K$ and incorrect string set $M_{s_r^n}^J$ are defined, respectively, as:

$$\begin{aligned} &\forall S \in M_{s_r^n}^K, \exists \mathrm{s} \in \mathrm{S}, s \equiv s_r^n, \\ &\forall S \in M_{s_r^n}^J, \forall \mathrm{s}' \in \mathrm{S}', s' \neq s_r^n \end{aligned} \tag{8}$$

In Eq.(8), $s \equiv s_r^n$ means that the state s has the same state label and same time alignment as the reference phone s_r^n. Through, defined strictly, the boundary constraints are slightly loosed in practice to allow certain degree of differences in time alignment.

The reference state sequence is the state sequence obtained by the Viterbi force alignment, therefore the reference state sequence keeps to be same for all frames. For every frame t, the different between the incorrect and correct state sequences is the state at time t, which is from the candidate competing state set that was selected from all state by using KL distance measure in away same as [5]. For every reference state $s_{t,r}$, its correct state sequence set $M_{s_{t,r}}^K$ contains only one sequence, namely the state sequence obtained by the Viterbi force alignment. The incorrect state sequence set $M_{s_{t,r}}^J$ are the state sequences that contain the competing state at frame t and same states as the correct state sequence at other frames.

The discriminant function for each string set can be formulated as:

$$g_K(\theta) = \log \left[\frac{1}{|M_{s_r^n}^K|} \sum_{S \in M_{s_r^n}^K} p_\theta^\alpha(X_r|S) p^\alpha(S) \right]^{1/\alpha} \tag{9}$$

and

$$g_J(\theta) = \log \left[\frac{1}{|M_{s_r^n}^J|} \sum_{S \in M_{s_r^n}^J} p_\theta^\alpha(X_r|S) p^\alpha(S) \right]^{1/\alpha}. \tag{10}$$

The misclassification measure related to the reference state s_r^n can be written as:

$$d_{s_r^n} = -g_K(\theta) + g_J(\theta). \tag{11}$$

Consequently, the state-level MCE criteria can be written as:

$$F_{MCE} = \sum_{r=1}^{R} \sum_{n=1}^{N_r} f(d_{s_r^n}) \tag{12}$$

where $f(z) = -1/(1 + e^{2\rho z})$.

4 Experiments and Results

4.1 Experimental Setup

The speech corpus employed in this paper is the continuous Mandarin speech corpora 863-I, which contains about 120 hours, including 166 speakers, 83 male speakers and 83 female speakers. The training set consists speech of 73 male speakers and 73 female speakers. The test set (863-I-Test) was selected from the remainder 20 speakers, 20 utterances each. From the same corpus with the training set, this test set is well matched with the training set. For investigating the generalization ability of different models, we also test the models on a not-well-matched test set, the 1997 HUB-4 Mandarin broadcast news evaluation (Hub-4-Test), which consists of 654 utterances, including 230 for male speakers and 424 for female speakers.

The acoustic models chosen for speech recognition were cross-word triphones models built by decision-tree state clustering. After clustering, the resulted HMM had 4,517 tied states with 32 Gaussian mixtures per state. The acoustic models were first trained using the ML criterion and the BW update formulas. Referring to [8,9], the state level MCE based methods is implemented with $\alpha = 1/15$ and $\rho = 0.04$. For investigating the effect of the different de-learning rate, the state level RPCL based methods are implemented with different de-learning rates γ =0.2, 0.3 and 0.4.

The language model for recognition evaluation is a word-based trigram built from a vocabulary of 57K entries. The input speech data is made up of Mel-frequency cepstral coefficients (MFCCs), with 13 cepstral coefficients including the logarithmic energy and their first and second-order differentials. All experiment results were obtained through a single pass recognition on test speech.

4.2 Experimental Results

The performance evaluation metric used in Mandarin speech recognition experiments is the Chinese character error rate (CER). The MLE based acoustic model yields a CER of 13.67% on test data 863-I-Test and 26.61% on test Hub-4-Test data, that is, the performance tested on the matched test data is much better than that tested on not-well-matched test data.

Character error rate of each iteration for two methods are shown in Fig.2. As shown in the figure, both the two methods get improved recognition performance on the two test sets.

The recognition performance of each method is given in Table 1. For RPCL based methods, an appropriate value for γ can further improve the performance. The RPCL with γ = 0.4 and 0.2 obtain best performance on 863-I-Test and Hub-4-Test respectively. The RPCL based methods with all de-learning rate performs better than the MCE based method, which shows the RPCL has better discriminative ability than the MCE. Tested on the unmatched set, the MCE based method is only slightly better than the MLE based method while the RPCL based method also obtain large improvement as that on the matched test set, which demonstrates the RPCL has better generative ability than the MCE.

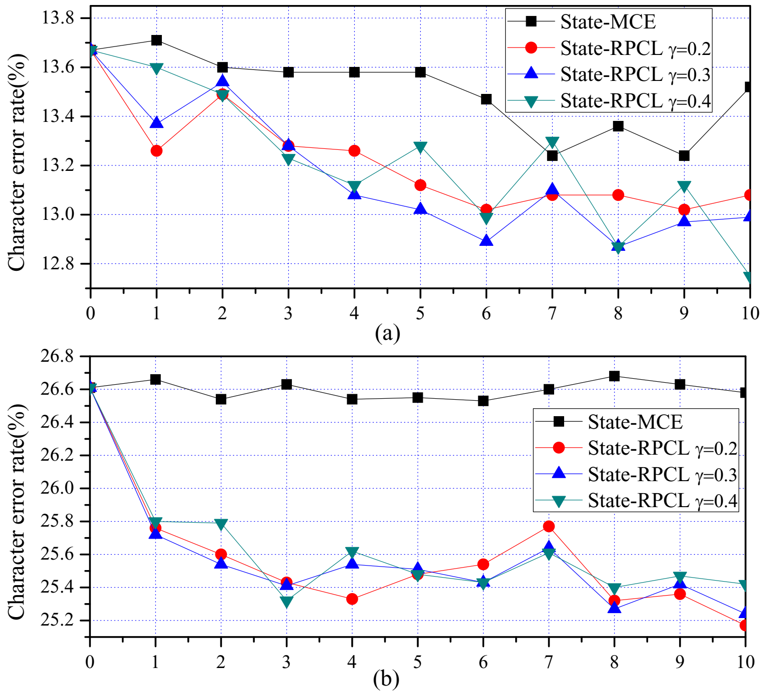

Fig. 2. Character error rate(%) varies with time on 863-I-Test(a) and Hub-4-Test(b)

Table 1. Performance comparison on 863-I-Test and Hub-4-Test for different methods

	863-I-Test		Hub-4-Test	
	CER(%)	RR(%)	CER(%)	RR(%)
MLE	13.67	-	26.61	-
State-MCE	13.24	3.15	26.53	0.30
State-RPCL $\gamma = 0.2$	13.02	4.75	25.17	5.41
State-RPCL $\gamma = 0.3$	12.87	5.85	25.24	5.15
State-RPCL $\gamma = 0.4$	12.75	6.73	25.32	4.85

5 Conclusions

This paper provides a preliminary comparison between MCE and RPCL in discriminative training of HMM based acoustic model in a large vocabulary continuous speech recognition (LVCSR) system. Both the two methods are implemented at the level of hidden states, and are tested on the data sets that are matched or unmatched with the training data set. The experimental results show that RPCL consistently performs better than MCE, on either matched or unmatched test data sets. All the results lead to such a conclusion that RPCL is a promising method with good theoretical basis and practical utility for the task of LVSCR. In the future, we will further implement and compare RPCL with MCE at the phone level or other levels.

Acknowledgments. The work was supported in part by the National Natural Science Foundation of China (No.90920302), a HGJ Grant of China (No. 2011ZX01042-001-001), a research program from Microsoft China and by a GRF grant from the Research Grant Council of Hong Kong SAR (Project CUHK 4180/10E). Lei Xu is also supported by Chang Jiang Scholars Program, Chinese Ministry of Education for Chang Jiang Chair Professorship in Peking University.

References

1. Bahl, L., Brown, P., De Souza, P., Mercer, R.: Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: Proc. of 1986 IEEE Intl. Conf. on ICASSP, pp. 49–52 (1986)
2. Juang, B.H., Katagiri, S.: Discriminative learning for Minimum Error Classification. IEEE Trans. on Signal Processing 40(12), 3043–3054 (1992)
3. Povey, D., Woodland, P.C.: Minimum phone error and I-smothing for improved discriminative training. In: Proc. of 2002 IEEE Intl. Conf. on ICASSP, vol. 1, pp. 105–108 (2002)
4. Jiang, H.: Discriminative training of HMMs for automatic speech recognition: A survey. Computer Speech and Language 24, 589–608 (2010)
5. Pang, Z.H., Tu, S.K., Su, D., Wu, X.H., Xu, L.: Discriminative training of GMM-HMM acoustic model by RPCL learning. Frontiers of Electrical and Electronic Engineering in China 6(2), 283–290 (2011)
6. Schlüter, R., Macherey, W., Müller, B., Ney, H.: Comparison of discriminative training criteria and optimization methods for speech recognition. Speech Communication 34(3), 287–310 (2001)
7. Fu, Q., He, X., Deng, L.: Phone-discriminating minimum classification error(P-MCE) training criteria for phonetic recognition. In: Proceedings of InterSpeech 2007, pp. 2073–2076 (2007)
8. Macherey, W., Haferkamp, L., Schlüter, R., Ney, H.: Investigations on error minimizing training criteria for discriminative training in acoustic speech recognition. In: Proceedings of EuroSpeech 2005, Lisbon, Portugal, pp. 2133–2136 (2005)
9. Yan, Z.J., Zhu, B., Hu, Y., Wang, R.H.: Minmum word classification error training of HMMs for automatic speech recognition. In: Proc. of 2008 IEEE Intl. Conf. on ICASSP, pp. 4521–4524 (2008)
10. Xu, L., Krzyzak, A., Oja, E.: Unsupervised and supervised classifications by rival penalized competitive learning. In: Proc. 11th Intl. Conf. on Pattern Recognition, Netherlands, vol. I, pp. 672–675 (1992)
11. Xu, L., Krzyzak, A., Oja, E.: Rival penalized competitive learning for clustering analysis, RBF net, and curve detection. IEEE Trans. on Neural Networks 4(4), 636–649 (1993)
12. Xu, L.: Rival penalized competitive learning. Scholarpedia 2(8), 1810 (2007)
13. Xu, L.: A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving. Pattern Recognition 40(8), 2129–2153 (2007)
14. Xu, L.: Bayesian Ying-Yang system, best harmony learning, and five action circling. FEEE in China 5(3), 281–328 (2010)