Neurocomputing 134 (2014) 53–59

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A comparative study of RPCL and MCE based discriminative training methods for LVCSR

Zaihu Pang [a], Shikui Tu [b], Xihong Wu [a,*], Lei Xu [a,b,**]

[a] Speech and Hearing Research Center, Key Laboratory of Machine Perception (Ministry of Education), Peking University, China
[b] Department of Computer Science and Engineering, The Chinese University of Hong Kong, China

ARTICLE INFO

Article history:
Received 10 June 2012
Received in revised form
15 April 2013
Accepted 9 May 2013
Available online 23 January 2014

Keywords:
Rival penalized competitive learning
Minimum classification error
Discriminative training
Large vocabulary continuous speech recognition

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 focuses 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 the levels of phones and/or hidden Markov states using the same training corpus. The results show that both the MCE and RPCL based methods perform better than the Maximum Likelihood Estimation (MLE) based method. Comparing with the MCE based method, the RPCL based methods have better discriminative and generalizing abilities on both two levels.

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 such as 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 of the same arc (word or phone) in different paths, when comparing with traditional string based competing space [4]. The third ingredient is to adopt the widely used Extended Baum–Welch (EBW) algorithm for parameter estimation. An overview of these methods is referred to [5]. Recently, Rival Penalized Competitive Learning (RPCL) was introduced in [6] to speech recognition with promising results in a comparison with MMIE and MPE. Still, there is a lack of comparison between RPCL and MCE. This paper is motivated for such a comparative study.

MCE criterion 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 [7], the string-level MCE was shown to have similar performance with MMIE based method on small vocabulary tasks. In [8], 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 [4,9] 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 [6], RPCL was implemented on the level of states for a discriminative Hidden Markov Model (HMM) based speech model as shown in Fig. 1. For each input, the winner state which is given by the correct identity state from Viterbi force alignment is enhanced while the most competitive rival state is repelled, which increases the discriminative ability and obtains preferable generalization ability. When applied to LVCSR, it showed better generalization performance than the MMIE and the MPE, especially when the sources of test sets are different from the training set.

This paper follows [6] to present a comparison between RPCL and MCE as discriminative training methods for LVCSR task. To investigate the impact of RPCL and MCE on different levels of

* Corresponding author.
** Corresponding author at: Speech and Hearing Research Center, Key Laboratory of Machine Perception (Ministry of Education), Peking University, China.
E-mail addresses: wxh@cis.pku.edu.cn (X. Wu), lxu@cse.cuhk.edu.hk (L. Xu).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.05.060

a speech recognition system, they are embedded in the levels of phones or hidden Markov states. According to [9] which uses the lattice based competing space as [4], MCE is derived to be implemented at the phone level, and also at the state level. For a fair comparison, RPCL is also extended from the state level in [6] to the phone level. Experiments are conducted on large vocabulary continuous speech recognition tasks: 863-I-Test (matched with the training data) and Hub-4-Test (unmatched with the training data). The results show that the RPCL based methods have better discriminative and generalizing abilities than MCE based methods on both levels, and 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 and then further is extended to phone level by using phone lattice as its competing space. In Section 3, the phone-level MCE and its state-level counterpart are briefly introduced. In Section 4, experimental results of RPCL and MCE on phone level and state level 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, to Sections 3.1 and 3.2 in [13] and particularly its Eqs. (7) and (34) for further details. In the following, we only provide a brief introduction.

In conventional RPCL, 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. Specifically, RPCL learning is simply implemented by

$$\theta_j^{new} - \theta_j^{old} \propto p_{j,t} \nabla_{\theta_j} \varepsilon_t(\theta_j), \tag{1}$$

where the term $\varepsilon_t(\theta_j)(\geq 0)$ measures the error or cost for the j-th learner to describe the current input x_t, the notation ∇_{θ_j} denotes the gradient operator with respect to θ_j, and the winner c_t and the rival r_t (i.e., the second winner) are as follows:

$$p_{j,t} = \begin{cases} 1 & \text{if } j = c_t; \\ -\gamma & \text{if } j = r_t; \\ 0 & \text{otherwise;} \end{cases} \quad \begin{cases} c_t = \arg\min_j \varepsilon_t(\theta_j), \\ r_t = \arg\min_{j \neq c_t} \varepsilon_t(\theta_j), \end{cases} \tag{2}$$

with γ being a small positive number. The rival penalized mechanism makes the boundaries between different learners become more discriminative.

The state-level RPCL was introduced for speech recognition system in [6] by considering $\varepsilon_t(\theta_j) = -\ln p(x_t|\theta_j)$ across different states $\{j\}$, where $p(x_t|\theta_j)$ is a mixture Gaussian density. In [6], the winner state c_t is determined by the identity of this input by the Viterbi force alignment, that is,

$$p_{j,t} = \begin{cases} 1 + p(r_t|x_t) & \text{if } j = c_t; \\ -p(r_t|x_t)\gamma & \text{if } j = r_t; \\ 0 & \text{otherwise} \end{cases} \quad \begin{cases} c_t = \text{by Viterbi force alignment}, \\ r_t = \arg\min_{j \neq c_t} \varepsilon_t(\theta_j), \end{cases} \tag{3}$$

where p_{jt} was refined to be a simplified approximation to the Bayesian Ying–Yang (BYY) harmony learning for which details are referred to Section 3.1 in [14] and particularly Section 2.1 in [6]. As illustrated in Fig. 2(a), the sample x (red one) is labeled with class A, but it has larger posterior probability for class B, $P(B|x) > P(A|x)$. For the input sample x, the winner is the A, while B is its best rival. Using the learning rule of Eq. (3), $P_A = 1 + p(x|B)$ and $P_B = -\gamma$. The learning of the A is enforced, while the B is de-learned. The class A moves close to the direction of x, while class B moves away from the direction of x. Repeat the learning program using all the samples iteratively until getting a good convergence. After the RPCL learning, the two classes move to a stable place as shown in Fig. 2(b). The BYY best learning provides a favorable new mechanism for model selection and discriminative learning. Readers are referred to papers [15,16] for recent systematic overviews on the fundamentals, the novelties and favorable natures of the BYY harmony learning.

Analogously, RPCL discriminative learning can be made at the phone level for each phone. Suppose the reference phone sequence of the r-th training utterance consists of N_r phones, i.e., $S_r = \{s_r^1, s_r^2, \ldots, s_r^{N_r}\}$. For each reference phone s_r^n, its correct string set $M_{s_r^n}^K$ and incorrect string set $M_{s_r^n}^I$ are defined, respectively, as follows:

$$\forall S \in M_{s_r^n}^K,\ \exists s \in S,\ s \equiv s_r^n, \quad \forall S \in M_{s_r^n}^I,\ \forall s' \in S',\ s' \neq s_r^n, \tag{4}$$

where $s \equiv s_r^n$ means that the phone s has the same phone label with the same time alignment as the reference phone s_r^n, and $s' \neq s_r^n$ means that the phone s' label differs from the reference phone s_r^n but has the same alignments as s_r^n. For the n-th reference phone s_r^n from the r-th utterance, the winner and the rival are defined by its best scored correct phone $s_r^{n,K}$ and incorrect phone

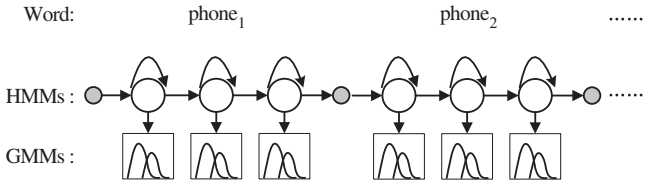

Fig. 1. The hierarchical structure of word in GMM-HMM based speech recognition: word level, phone level (HMM) and state level (GMM).

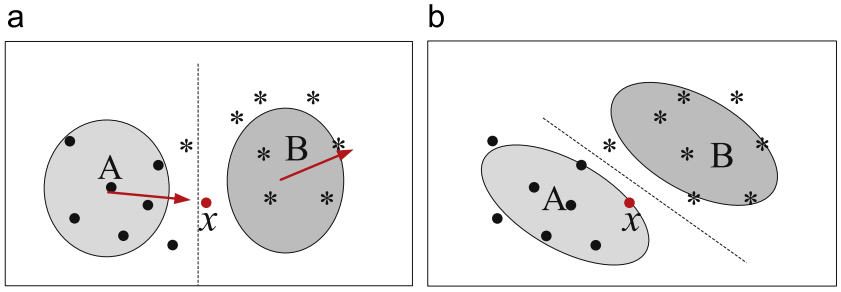

Fig. 2. Two class supervised model training using rival penalized competitive learning: (a) the learning trend of two models for incoming sample x and (b) a stable condition after RPCL iterative learning. (For interpretation of the references to color in this figure the reader is referred to the web version of this paper.)

$s_r^{n,J}$ in the denominator lattice as follows:

$$s_r^{n,K} = \arg\min_{s \in S_r^{den}, s \equiv s_r^n} \{\varepsilon_t(\theta_s)\}, \quad s_r^{n,J} = \arg\min_{s \in S_r^{den}, s \neq s_r^n} \{\varepsilon_t(\theta_s)\},$$

$$\varepsilon_t(\theta_s) = -\ln\left[\frac{1}{T_s} p(X|\theta_s)\right], \tag{5}$$

where S_r^{den} is the decoding space from the denominator lattice of r-th utterance. For comparing competing ability fairly, the likelihood of every phone is normalized by the corresponding length T_s.

The posterior probability of every phone s is computed by

$$\Gamma_s^{RPCL} = \Gamma_s^r \cdot \xi_s, \quad \xi_s = \begin{cases} 1+\delta_s & \text{if } s = s_r^{n,K} \\ -\delta_s \gamma & \text{if } s = s_r^{n,J} \end{cases} \tag{6}$$

where Γ_s^r is the posterior probability of the phone s of the r-th utterance, which is collected from the lattice using the forward-backward algorithm, and ξ_s is the weight of the phone s and δ_s is the counterpart of $p(r_t|x_t)$ of Eq. (2)(b) to represent the degree of the competition:

$$\delta_{s_r^n} = \frac{\dfrac{1}{T_{s_r^{n,J}} p(X_{s_r^{n,J}}|\theta_{s_r^{n,J}})}}{\dfrac{1}{T_{s_r^{n,K}}} p(X_{s_r^{n,K}}|\theta_{s_r^{n,K}}) + \dfrac{1}{T_{s_r^{n,J}}} p(X_{s_r^{n,J}}|\theta_{s_r^{n,J}})}. \tag{7}$$

In Eq. (6), γ is the de-learning rate. The bigger the γ is, the stronger the de-learning is. For one reference phone s_r^n, the learning of the winner phone $s_r^{n,K}$ is enhanced, while its rival phone $s_r^{n,J}$ 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 phone, which makes the phones more discriminative.

Accordingly, the parameters of each Gaussian mixture component are updated according to the following modification of the BW algorithm:

$$\alpha_{jm}^{new'} = \Gamma_{jm} / \sum_{m'=1}^{K_j} \Gamma_{jm'}, \quad \Gamma_{jm} = \sum_s \sum_{t=1}^{t_s} \Gamma_{sjm}(t) \Gamma_s^{RPCL},$$

$$\mu_{jm}^{new'} = \frac{1}{\Gamma_{jm}} \sum_s \sum_{t=1}^{t_s} \Gamma_{sjm}(t) \Gamma_s^{RPCL} \cdot x_t,$$

$$\Sigma_{jm}^{new'} = \frac{1}{\Gamma_{jm}} \sum_s \sum_{t=1}^{t_s} \Gamma_{sjm}(t) \Gamma_s^{RPCL} \cdot [(x_t - \mu_{jk})(x_t - \mu_{jk})^T], \tag{8}$$

where Γ_{sjm} denotes the posterior probability of phone s, state j and Gaussian component m, and K_j is the number of Gaussian component of state j. Eq. (8) differs from the BW algorithm and the EBW algorithm for the lattice based MCE in the role of Γ_s^{RPCL} as introduced above.

The above estimate $\theta^{new'}$ for each $\theta \in \{\alpha_{jm}, \mu_{jm}, \Sigma_{jm}\}$ specifies a direction in which θ^{old} may be updated along with. However, a direct use of $\theta^{new'}$ indicates a move with a too large learning step along the direction $\theta^{new'} - \theta^{old}$. Similar to Box-3 in Fig. 7 in [14] and the Ying step at the end of [12], we consider the following linear interpolation:

$$\theta^{new} = (1-\lambda)\theta^{old} + \lambda\theta^{new'} \tag{9}$$

where λ indicates an appropriate step-size in which the update θ^{new} approaches to $\theta^{new'}$, with $0 < \lambda \leq 1$.

3. Minimum classification error discriminant function

Using lattice as its competing space, the phone-level MCE based DT method [4,9] considers the following discriminant function for each string set:

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

and

$$g_J(\theta) = \log\left[\frac{1}{|M_{s_r^n}^J|} \sum_{S \in M_{s_r^n}^J} p_\theta^\beta(X_r|S) p^\beta(S)\right]^{1/\beta} \tag{11}$$

where β is the weighting exponent from which the phone-level MCE criterion in consideration is written as

$$F_{MCE} = \sum_{r=1}^{R} \sum_{n=1}^{N_r} f(d_{s_r^n}), \quad d_{s_r^n} = -g_K(\theta) + g_J(\theta), \tag{12}$$

where $f(z) = -1/(1+e^{2\rho z})$, and $d_{s_r^n}$ is the misclassification measure related to the reference phone s_r^n.

To compare with the state-level RPCL method [6], the MCE is also considered at the state level. The competing space of state-level MCE is the same as that of state-level RPCL. The discriminant function and the loss function of the state-level MCE are in the same format as the phone-level MCE. The difference comes from the discriminative unit and its discriminative state sequences.

The reference state sequence is obtained by the Viterbi force alignment, and it is kept to be the same for all frames. For every frame t, the candidate competing state set is selected according to the KL distance measure in the same way as the one used in [6]. For every reference state $s_{t,r}$, its correct state sequence set $M_{s_{t,r}}^K$ contains only the best alignment state sequence, while the incorrect state sequence set $M_{s_{t,r}}^J$ contains those different from the correct one only at the time t.

The above implementation of MCE is based on Section 3.1 of [9], which extends the original MCE in [2] for the LVCSR task with the use of lattices to compactly represent competing space.

On the whole, though both the RPCL and the MCE enforce learning of correct class and de-learning its best rival, they have difference at the allocation mechanism. In RPCL, the enforce learning and the de-learning are controlled by the posterior probability of the de-learning rate. While in MCE, the enforce learning and the de-learning are controlled by the smoothed sequence error. Also, from the form the sequence learning, the mechanism of RPCL is inclined to the local error, while the MCE focuses on the long sequence error.

4. Experiments and results

The speech corpus employed in this paper is the continuous Mandarin speech corpora 863-I, which contains about 120 h, 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 triphone models built by decision-tree state clustering. After clustering, the resulted HMM had 4517 tied states with 32 Gaussian mixtures per state. The acoustic models were first trained using the MLE criterion and the BW update formulas. Using this acoustic model, two sets of lattices named numerator and denominator are generated using HTK toolkit [17]. Both the phone-level MCE and RPCL methods share the same training

lattices. To improve generalization, a syllable based unigram language model is trained to generate phone lattices. Referring to [4,9], both the phone and state level MCE based methods are implemented with $\beta=1/15$ and $\rho=0.04$. For investigating the effect of the different de-learning rate, both the phone and state level RPCL based methods are implemented with different de-learning rates $\gamma=0.2$, 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 experimental results were obtained through a single pass recognition on test speech.

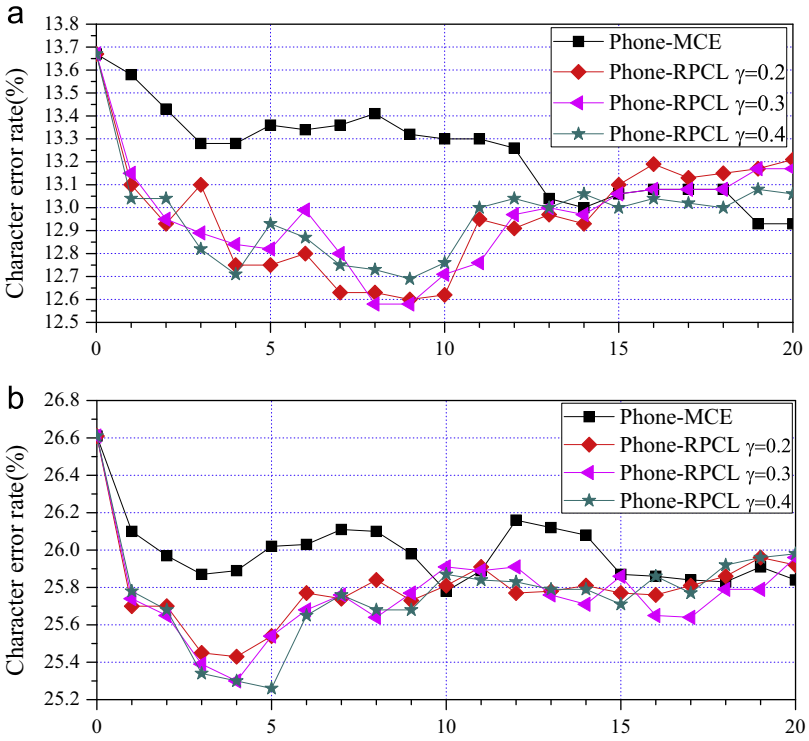

Fig. 3. Character error rates (CER) (%) for each iteration on (a) 863-I-Test (matched with training set) and (b) Hub-4-Test (unmatched with training set) using phone-level MCE and RPCL methods.

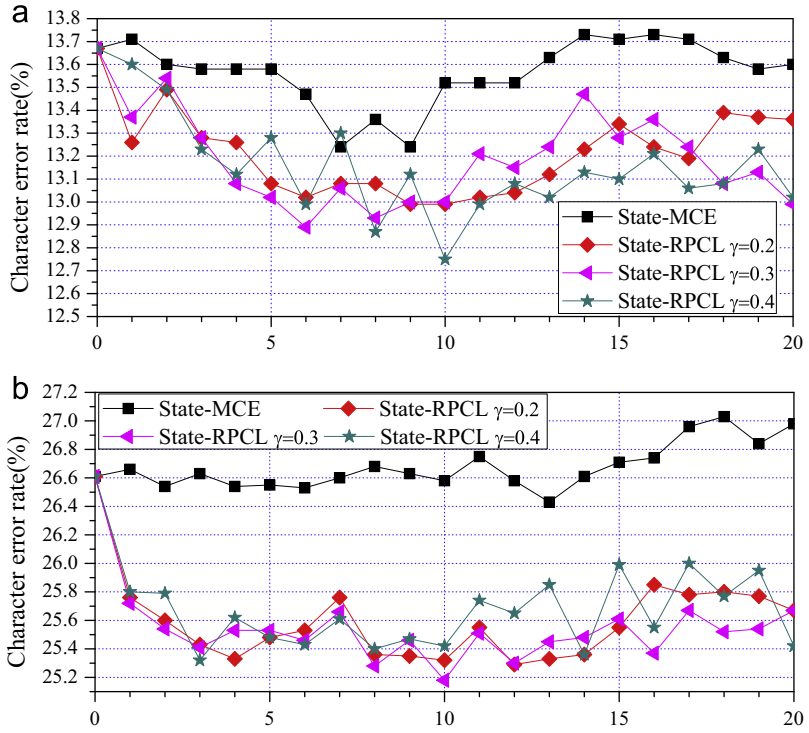

Fig. 4. Character error rate (%) for each iteration on (a) 863-I-Test and (b) Hub-4-Test using state-level MCE and RPCL methods.

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 863-I-Test and 26.61% on Hub-4-Test, that is, the performance tested on the matched test data is much better than that tested on not-well-matched test data.

4.1. RPCL vs MCE: with $\lambda = 1$ in Eq. (9) for RPCL

Based on the experimental results, we have the following observations:

- At the phone-level, CER of each iteration for two methods is shown in Fig. 3. Comparing with the MLE based method, both DT methods get improved recognition performance on the two test sets. As shown in Fig. 3, for both the matched and unmatched sets, the CER of RPCL first decreases to be smaller than that of MCE and then increases with the gap vanishing gradually. This is a typical phenomenon that is usually called "overtraining", which indicates that learning regularization is needed. In other words, learning by Eq. (9) with $\lambda = 1$ has a too aggressive learning step size, which will be reduced in the experiments shown in Fig. 5.
- At the state-level in Fig. 4, RPCL consistently outperforms MCE, especially for the unmatched set in Fig. 4(b) where RPCL stably improves MLE a lot but MCE does not show obvious improvement over MLE. Although there are still slight fluctuations, the state-level implementation of RPCL is stabilized even by the updating Eq. (9) with $\lambda = 1$.
- The best recognition performances of each method at different de-learning rates are given in Table 1:
  - the phone-level MCE outperforms the state-level MCE on both 863-I-Test and Hub-4-Test;
  - RPCL has a larger improvement over MLE on the phone-level implementation than the state-level on the 863-I-Test. Moreover the state-level RPCL slightly outperforms the phone-level one on the Hub-4-Test;
  - Among all results, the phone-level RPCL with γ=0.3 is the best on the 863-I-Test, while the state-level RPCL with γ=0.2 gets the best result on Hub-4-Test.

Table 1
Performance comparison based on Figs. 3 and 4.

	863-I-Test		Hub-4-Test	
	CER (%)	RR (%)	CER (%)	RR (%)
MLE	13.67	–	26.61	–
Phone-MCE	12.93	5.41	25.78	3.12
Phone-RPCL γ=0.2	12.60	7.83	25.43	4.43
Phone-RPCL γ=0.3	12.58	7.97	25.30	4.92
Phone-RPCL γ=0.4	12.59	7.90	25.26	5.07
State-MCE	13.24	3.15	26.43	0.68
State-RPCL γ=0.2	13.02	4.75	25.17	5.41
State-RPCL γ=0.3	12.87	5.85	25.24	5.15
State-RPCL γ=0.4	12.75	6.73	25.32	4.85

4.2. RPCL vs MCE: with different λ in Eq. (9) for RPCL

As shown in the Fig. 3, the performance of RPCL fluctuates as the training proceeds. To obtain more stable performances, we

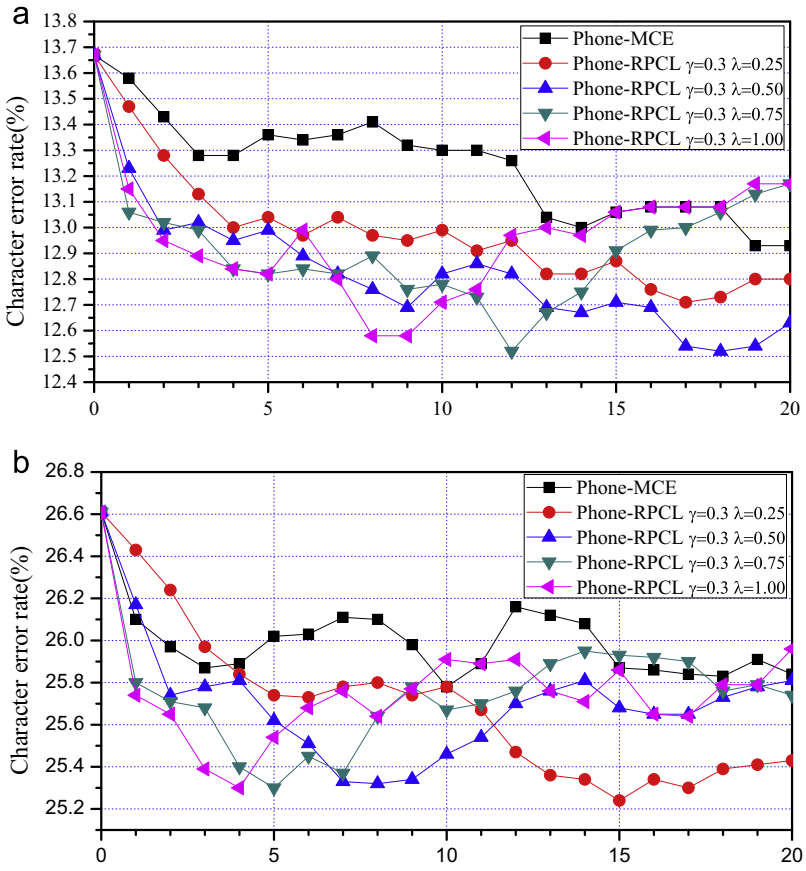

Fig. 5. Character error rate (%) for each iteration on (a) 863-I-Test and (b) Hub-4-Test using phone-level RPCL methods with γ=0.3 and λ=0.25,…,1.0. The results of Phone-MCE are taken from Fig. 3.

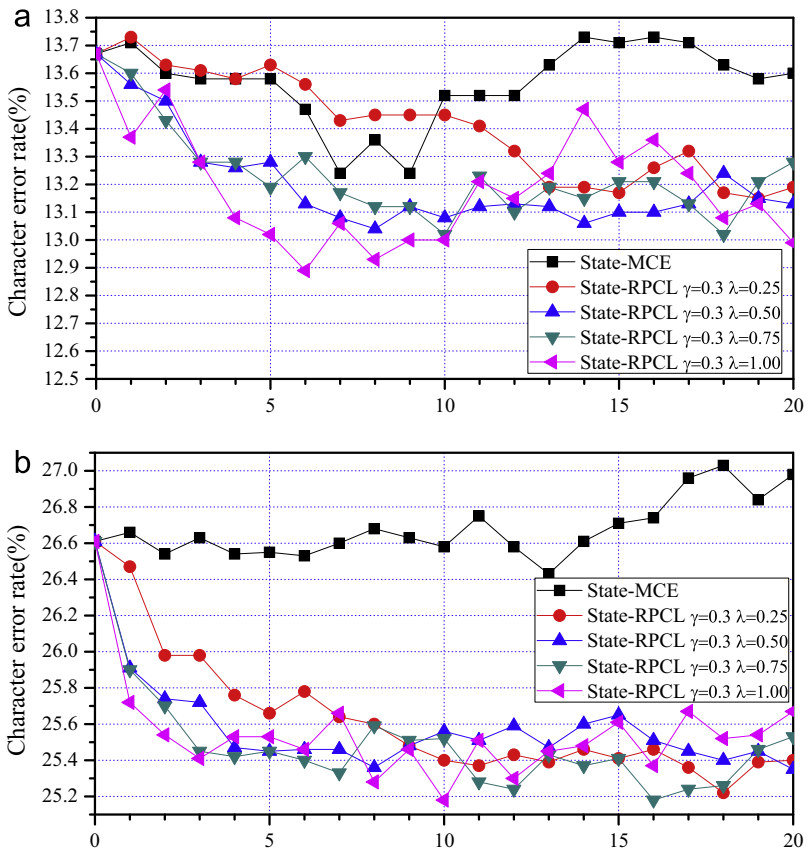

Fig. 6. Character error rate (%) for each iteration on (a) 863-I-Test and (b) Hub-4-Test using state-level RPCL methods with $\gamma=0.3$ and $\lambda=0.25,\ldots,1.0$. The results of State-MCE are taken from Fig. 4.

Table 2
Performance comparison based on Figs. 5 and 6. The best results of "Phone-MCE" and "State-RPCL $\gamma=0.3$ $\lambda=1.00$" are the same as their corresponding ones in Table 1.

	863-I-Test		Hub-4-Test	
	CER (%)	RR (%)	CER (%)	RR (%)
MLE	13.67	–	26.61	–
Phone-MCE	12.93	5.41	25.78	3.12
Phone-RPCL $\gamma=0.3$ $\lambda=0.25$	12.71	7.02	25.24	5.15
Phone-RPCL $\gamma=0.3$ $\lambda=0.50$	12.52	8.41	25.32	4.85
Phone-RPCL $\gamma=0.3$ $\lambda=0.75$	12.52	8.41	25.30	4.92
Phone-RPCL $\gamma=0.3$ $\lambda=1.00$	12.58	7.97	25.30	4.92
State-MCE	13.24	3.15	26.43	0.68
State-RPCL $\gamma=0.3$ $\lambda=0.25$	13.15	3.80	25.36	4.70
State-RPCL $\gamma=0.3$ $\lambda=0.50$	13.04	4.61	25.35	4.74
State-RPCL $\gamma=0.3$ $\lambda=0.75$	13.02	4.75	25.18	5.37
State-RPCL $\gamma=0.3$ $\lambda=1.00$	12.89	5.71	25.18	5.37

implement Eq. (9) by decreasing λ from $\lambda=1$ to $\lambda=0.75$, 0.5, 0.25, which actually decreases the learning step size from large to small. We demonstrate the performances of RPCL with varying λ at a de-learning rate $\gamma=0.3$.

- It can be observed in Fig. 5 that the fluctuations in CER of RPCL become weak as the step size λ decreases, and the RPCL with $\lambda=0.25$ is generally the best and consistently outperforms the phone-level MCE. Comparing Fig. 5 with Fig. 3 implies that an appropriate step size λ is important for phone-level RPCL.
- Although the state-level RPCL in Fig. 4 is already stable, adjusting step size λ in Fig. 6 leads to a further improved relative reduction on 863-I-Test from the best one 7.97 in Table 1 to 8.41 in Table 2.

5. Conclusions

This paper has provided a comparison of MCE and RPCL in discriminative training for LVCSR systems. The two methods are both implemented at phone and hidden Markov state levels, and tested on the data sets that are matched or unmatched with the training data set. Experimental results show that RPCL consistently performs better than MCE at both phone and state levels on both matched and unmatched test data sets. All the results indicate that RPCL is a promising method for the task of LVCSR.

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China (Nos. 91120001 and 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 a Chang Jiang Chair Professor in Peking University.

References

[1] L. Bahl, P. Brown, P. de Souza, R. Mercer, Maximum mutual information estimation of hidden Markov model parameters for speech recognition, in: Proceedings of the ICASSP, 1986, pp. 49–52.
[2] B.H. Juang, S. Katagiri, Discriminative learning for minimum error classification, IEEE Trans. Signal Process. 40 (1992) 3043–3054.
[3] D. Povey, P.C. Woodland, Minimum phone error and I-smoothing for improved discriminativetraining, in: Proceedings of the ICASSP, 2002, pp. 105–108.

[4] W. Macherey, L. Haferkamp, R. Schlüter, H. Ney, Investigations on error minimizing training criteria for discriminative training in acoustic speech recognition, in: Proceedings of the EuroSpeech, 2005, pp. 2133–2136.
[5] H. Jiang, Discriminative training of HMMs for automatic speech recognition: a survey, Comput. Speech Lang. 24 (2010) 589–608.
[6] Z.H. Pang, S.K. Tu, D. Su, X.H. Wu, L. Xu, Discriminative training of GMM-HMM acoustic model by RPCL learning, Front. Electr. Electron. Eng. China 6 (2011) 283–290 (A special issue on Machine Learning and Intelligence Science: IScIDE2010 (B)).
[7] R. Schlüter, W. Macherey, B. Müller, H. Ney, Comparison of discriminative training criteria and optimization methods for speech recognition, Speech Commun. 34 (2011) 287–310.
[8] Q. Fu, X.D. He, L. Deng, Phone-discriminating minimum classification error (P-MCE) training criteria for phonetic recognition, in: Proceedings of the Inter-Speech, 2007, pp. 2073–2076.
[9] Z.J. Yan, B. Zhu, Y. Hu, R.H. Wang, Minmum word classification error training of HMMs for automatic speech recognition, in: Proceedings of the ICASSP, 2008, pp. 4521–4524.
[10] L. Xu, A. Krzyzak, E. Oja, Unsupervised and supervised classifications by rival penalized competitive learning, in: Proceedings of the ICPR, 1992, pp. 672–675.
[11] L. Xu, A. Krzyzak, E. Oja, Rival penalized competitive learning for clustering analysis, RBF net, and curve detection, IEEE Trans. Neural Netw. 4 (1993) 636–649.
[12] L. Xu, Rival penalized competitive learning, Scholarpedia 2 (8) (2007) 1810.
[13] L. Xu, A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving, Pattern Recognit. 40 (2007) 2129–2153.
[14] L. Xu, Bayesian Ying–Yang system, best harmony learning, and five action circling, Front. Electr. Electron. Eng. China 5 (2010) 281–328.
[15] L. Xu, Codimensional matrix pairing perspective of BYY harmony learning: hierarchy of bilinear systems, joint decomposition of data-covariance, and applications of network biology, Front. Electr. Electron. Eng. China 6 (2011) 86–119 (A special issue on Machine Learning and Intelligence Science: IScIDE2010 (A)).
[16] L. Xu, On essential topics of BYY harmony learning: current status, challenging issues, and gene analysis applications, Front. Electr. Electron. Eng. 7 (2012) 147–196 (A special issue on Machine Learning and Intelligence Science: IScIDE2010 (C)).
[17] S. Young, G. Evermann, M. Gales, et al., The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2006.

Zaihu Pang is currently a Ph.D. candidate at the Speech and Hearing Research Center, Peking University, PR China. He received the B.S. degree from College of Computer Science and Technology, Jilin University, PR China, in 2006. His research interests include speech recognition and statistical learning.

Shikui Tu is a Ph.D. candidate of the Department of Computer Science and Engineering, The Chinese University of Hong Kong, PR China. He received the B.S. degree from School of Mathematical Science, Peking University, PR China, in 2006. His research interests include statistical learning, pattern recognition, and bioinformatics.

Xihong Wu received the B.S. degree from Jilin University, PR China, in 1989, the M.S. degree from the Institute of Harbin Shipbuilding Engineering in PR China, in 1992, and the Ph.D. degree from the Department of Radio Electronics, Peking University, PR China, in 1995. He is currently a professor and supervisor of Ph.D. candidates with Peking University. He has been elected a senior member of IEEE, in 2009. His areas of research focus include computational auditory models and auditory scene analysis, auditory psychophysics, speech signal processing, and natural language processing.

Lei Xu is a IEEE Fellow (2001–) and Fellow of International Association for Pattern Recognition (2002–), and Academician of European Academy of Sciences (2002–); a Chair Professor with the Chinese University of Hong Kong, a Chang Jiang Chair Professor with Peking University and an Honorary Professor with Xidian University, PR China.