Front. Electr. Electron. Eng. China 2011, 6(2): 283–290
DOI 10.1007/s11460-011-0152-0

RESEARCH ARTICLE

Zaihu PANG, Shikui TU, Dan SU, Xihong WU, Lei XU

Discriminative training of GMM-HMM acoustic model by RPCL learning

Abstract This paper presents a new discriminative approach for training Gaussian mixture models (GMMs) of hidden Markov models (HMMs) based acoustic model in a large vocabulary continuous speech recognition (LVCSR) system. This approach is featured by embedding a rival penalized competitive learning (RPCL) mechanism on the level of hidden Markov states. For every input, the correct identity state, called winner and obtained by the Viterbi force alignment, is enhanced to describe this input while its most competitive rival is penalized by de-learning, which makes GMMs-based states become more discriminative. Without the extensive computing burden required by typical discriminative learning methods for one-pass recognition of the training set, the new approach saves computing costs considerably. Experiments show that the proposed method has a good convergence with better performances than the classical maximum likelihood estimation (MLE) based method. Comparing with two conventional discriminative methods, the proposed method demonstrates improved generalization ability, especially when the test set is not well matched with the training set.

Keywords discriminative training, hidden Markov model, rival penalized competitive learning, Bayesian Ying-Yang harmony learning, large vocabulary continuous speech recognition

1 Introduction

Parameters of acoustic model in hidden Markov model (HMM) based speech recognition systems are usually estimated using maximum likelihood estimation (MLE) [1,2]. The weakness of MLE lies in that it cannot directly optimize word or phone recognition error rates due to its strong assumptions on sufficient training data and model-correctness [1]. To solve this problem, a variety of discriminative training methods have been extensively investigated to improve automatic speech recognition (ASR) system for decades. Typical ones include maximum mutual information (MMI) estimation [3], minimum classification error (MCE) [4], and minimum word/phone error (MWE/MPE) [5].

Generally, these discriminative methods achieve good performances when acoustic conditions in a testing set match well with those in the training set. However, in most practical conditions, the test speech does not match the training set well. Smoothing techniques have been proposed to improve generalization ability of these methods, such as smoothing sigmoid function in MCE [4], acoustic scaling and weaken language modeling in MMI [3], and I-smoothing in MPE [5]. The Baum-Welch (BW) algorithm is extended to update HMM parameters for implementing these techniques. Though recognition performance can be improved in many large-scale recognition tasks, a large amount of computational cost should be used to make one pass or more of recognition on all the training utterances in order to obtain confusable hypotheses of the training data in implementation of the extended BW algorithm.

In our previous study [6], the Bayesian Ying-Yang (BYY) harmony learning has been introduced into estimating Gaussian mixture model (GMM) components by a two level procedure as shown in Fig. 1. Experiments have shown better performances than ones by not only the standard MLE training but also plus selecting GMM components with the help of the classical BIC and AIC criterions. In Ref. [6], the BYY harmony learning acts on GMM components within the same hidden Markov state and thus does not target at a discriminative training. To enhance discriminative ability, we may make the

Received April 11, 2011; accepted April 28, 2011

Zaihu PANG, Dan SU, Xihong WU (✉), Lei XU (✉)
Speech and Hearing Research Center, Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China
E-mail: wxh@cis.pku.edu.cn; lxu@cse.cuhk.edu.hk

Shikui TU, Lei XU
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China

BYY harmony learning at the levels of states, phones, and words for learning the GMM components across different states, as well as state transfer probabilities, for which we may consider the Ying-Yang alternative learning algorithm given in Sect. 5.3 (especially Fig. 14) of Ref. [7].

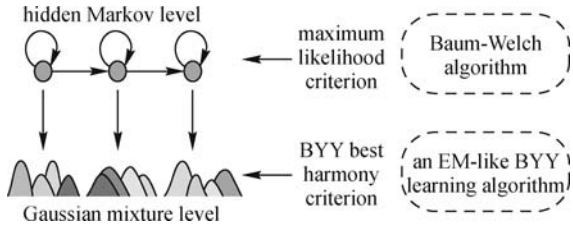


Fig. 1 Training framework of a two-level procedure

This paper attempts to embed the rival penalized competitive learning (RPCL) [8,9] into the HMMs-based acoustic model in a large vocabulary continuous speech recognition (LVCSR) system. First proposed in 1992 [8], 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. Moreover, RPCL can be explained as a simplified approximation of the BYY harmony learning. For a detailed discussion about the link of RPCL to the BYY harmony learning, readers are referred to Ref. [9].

In principle, this embedding of the RPCL mechanism may be made on different levels (word, phone, state, Gaussian) of HMMs-based acoustic model. Towards this purpose, we move step by step for solid developments. This paper still adopts the two-level procedure in Fig. 1 with the RPCL mechanism made at the levels of states to enhance the discriminative ability across state. The entire model is trained by the BW algorithm, with the resulted model used as an initialization. Then, two levels are trained alternatively. At the upper level, the transfer probabilities across states are still trained by the BW algorithm, while at the lower level, noticing that an LVCSR task usually involves a large number of states, we adopt an RPCL type learning featured by Eqs. (9) and (34) in Ref. [10] for training GMM components across different hidden Markov states. In implementation, for every input we get the winner state according to the identity of this input obtained by the Viterbi force alignment, while its rival state is sought among a set of candidate competitors. Not only the GMM components associated with this rival state is updated for a downgraded description of the input as the conventional RPCL does, but also the ones associated with the winner state are enhanced for an improved description of the input, which makes the states become more discriminative. Moreover, the strengths of enhancing and de-learning in RPCL utilize the information of the posterior probability of the rival given the input. Experiments on LVCSR tasks show that the proposed method has a good convergence with better performances than the baseline model trained by the classical maximum likelihood (ML) criterion. In comparison with two typical discriminative training methods, namely based on MMI and MPE respectively, the proposed method demonstrates improved generalization ability, especially when the sources of test sets are different from ones of training set.

The rest of the paper is organized as follows. Section 2 presents an introduction on the conventional RPCL and derive one suitable for the task of discriminative learning in speech recognition. Section 3 provides details of implementation and discussions on results. Finally, conclusion is made in Sect. 4.

2 RPCL-based acoustic model training

2.1 RPCL on state levels

First proposed in 1992 [8] 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 Ref. [9] for a systematic review and recent developments. In sequel, we only provide a brief introduction.

Using $\varepsilon_t(\theta_j) \geqslant 0$ to measure the error or cost that the jth learner describes the current input x_t, the learner with $c_t = \arg\min_j \varepsilon_t(\theta_j)$ is called the winner while the second winner $r_t = \arg\min_{j \neq c_t} \varepsilon_t(\theta_j)$ is its rival, the key idea of RPCL is that not only the parameter θ_{c_t} of the winner is learned such that $\varepsilon_t(\theta_{c_t})$ decreases by 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. In general, the winner and rival are decided by Eq. (1) based on the measure $\varepsilon_t(\theta_j)$, while learning is simply implemented by Eq. (2). The rival penalized mechanism makes the boundaries between different learners or models become more discriminative.

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

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

Readers are referred to Sects. 3.1 and 3.2 in Ref. [10] and particularly its Eqs. (9) and (34) for further details.

In speech recognition, we consider to make 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. Particularly, we consider

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

where γ, playing a similar role as in Eq. (1), denotes the de-learning rate. The bigger the γ is, the more strengthen the de-learning is. Noticing that, Eq. (4) is different from Eq. (1) as follows:

1) Not only the Gaussian components associated with the rival state are de-learned, but also learning the Gaussian components on the winner state is enhanced.

2) The strengths of enhancing and de-learning vary as the posterior probability of the rival r_t given the input x_t changes, which may make the discriminative learning more efficiently.

In the following, we show how the above Eq. (4) is obtained by approximately simplifying a general rival penalizing mechanism of the BYY harmony learning. 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 [7]. For a task of learning a mixture of multiple models to which our tasks belongs, without a priori knowledge, the BYY harmony learning is implemented via maximizing $H(p||q, \theta)$ given in Eq. (8) of Ref. [9]. 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 Ref. [9], which is rewritten as

$$p_{j,t} = \begin{cases} p(c_t|x_t) + \eta_t, & j = c_t = \arg\min_j \varepsilon_t(\theta_j), \\ p(r_t|x_t) - \eta_t, & j = r_t = \arg\min_{j\neq c_t} \varepsilon_t(\theta_j), \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $\eta_t = p(c_t|x_t)p(r_t|x_t)\rho(x_t)$, $\rho(x_t) = \ln \frac{p(x_t|\theta_{c_t})}{p(x_t|\theta_{r_t})}$, with $p(c_t|x_t) = \frac{p(x_t|\theta_{c_t})}{p(x_t|\theta_{c_t})+p(x_t|\theta_{r_t})}$ and $p(r_t|x_t) = 1 - p(c_t|x_t)$. By denoting de-learning rate $\gamma_t = -1 + p(c_t|x_t)\rho(x_t)$, Eq. (5) can be rewritten into

$$p_{j,t} = \begin{cases} p(c_t|x_t) + (1+\gamma_t)p(r_t|x_t), & j = c_t, \\ -p(r_t|x_t)\gamma_t, & j = r_t, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Putting it into Eq. (1), we can 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

$$p_{jk,t} = p_{j,t}p(k|x_t, \theta_j), \tag{7}$$

where $p(k|x_t, \theta_j) = \alpha_{jk}p(x_t|\theta_{jk}) / \left[\sum_{i=1}^{K} \alpha_{ji}p(x_t|\theta_{ji})\right]$. Actually, getting $p_{jk,t}$ by Eq. (7) is another 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, see Fig. 13(b) in Ref. [7]. Still, we make further simplifications. If the winner state c_t is considered reliable, we let $p(c_t|x_t) \approx 1$. Also, the γ_t is considered as a small constant γ. Then, Eq. (6) can be simplified to Eq. (4).

2.2 Parameter re-estimation

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

$$\sum_{t=1}^{T} p_{jk,t}\nabla_{\theta_{jk}}\varepsilon_t(\theta_{jk}) = 0,$$

that we get

$$\begin{aligned} \alpha_{jk}^{\text{new}} &= \frac{\sum_{t=1}^{T} p_{jk,t}}{\sum_{k=1}^{K}\sum_{t=1}^{T} p_{jk,t}}, \\ \mu_{jk}^{\text{new}} &= \frac{\sum_{t=1}^{T} p_{jk,t}x_t}{\sum_{t=1}^{T} p_{jk,t}}, \\ \Sigma_{jk}^{\text{new}} &= \frac{\sum_{t=1}^{T} p_{jk,t}(x_t - \mu_{jk}^{\text{new}})(x_t - \mu_{jk}^{\text{new}})^{\mathrm{T}}}{\sum_{t=1}^{T} p_{jk,t}}. \end{aligned} \tag{8}$$

Together with Eq. (7), we iterate the following steps that implements an RPCL-based acoustic model training:

1) Get RPCL-allocation by Eq. (7);

2) Re-estimate Gaussian components by Eq. (8);

which has a same format as the classical expectation-maximization (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.

3 Experiment

3.1 Implementation details

Summarized in Fig. 2, the general framework for implementing the RPCL learning on the entire GMM-HMM-based acoustic model, with details addressed as follows:

1) Get a candidate rival set $\mathcal{C}_s$ for every state using KL divergence to measure the differences between two states.

2) Align the correct transcription to the training set to get the correct state label c_t for every frame x_t.

3) For every frame x_t, select the most competitive rival r_t of c_t from the candidate set $\mathcal{C}_s$.

4) Get RPCL allocation by Eq. (7) and re-estimate Gaussian component parameters by Eq. (8).

5) Repeat steps 2), 3) and 4) iteratively until getting a good convergence.

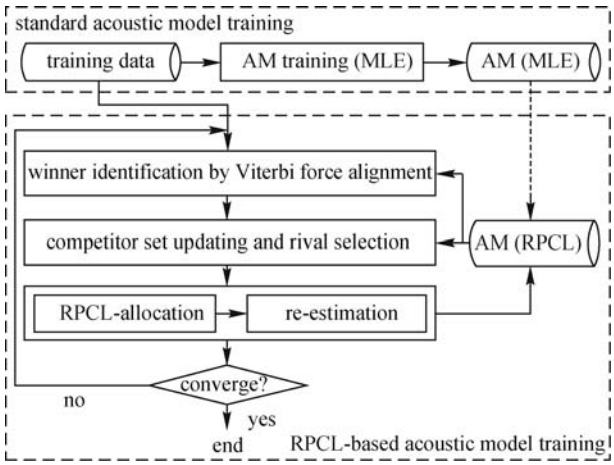


Fig. 2 Framework of discriminative training of GMM-HMM acoustic model by RPCL learning

Since a practical speech recognition purpose involves several thousands of tied states, considering all the states as the candidate rivals is not only computational not feasible but also unnecessary. Instead, we only consider a subset of states, called the candidate rival set $\mathcal{C}_s$ consisting of top-N nearest states that compete the correct state c_t of the input x_t. Also, the states mapped to the same monophone as the state c_t are excluded from $\mathcal{C}_s$. The size of $\mathcal{C}_s$ is controlled to be much smaller than the number of all states. Moreover, only the most competitive rival is used in the iterative learning by Eqs. (7) and (8), which will significantly reduce computational complexity.

There could be different ways for judging whether one state competes with the state c_t. We adopt the KL divergence $\mathrm{KL}(f||g)$ via an efficient computed approximation [11]:

$$\mathrm{KL}(f||g) \approx \sum_{i=1}^{K} \alpha_i \cdot \min_{j=1,2,\ldots,K} \mathrm{KL}(f_i||g_j), \quad (9)$$

which is based on a matching function between each component of the GMM under the states f and g. In Eq. (9), α_i is the mixture weight of Gaussian f_i. Theoretically, the candidate rival set needs to be updated iteratively. Practically, as long as the set size is big enough, the elements of the rival set remain virtually unchanged in the iteratively learning. To save computing cost, the candidate set is prepared with a size $N = 100$ before the iteratively learning in the following experiments. A dynamic candidate rival set will be studied in future research.

The correct state c_t per input x_t is obtained by the Viterbi force alignment. At the beginning of learning, the state transient probability and all the other parameters in the HMM were initially estimated by the classic BW algorithm from which c_t is also initially obtained via the Viterbi force alignment. During the RPCL learning by Eqs. (7) and (8), Gaussian components are updated iteratively and the updated parameters may be used together with the state transient probabilities to get c_t via the Viterbi force alignment. After a certain period of RPCL learning, we may also re-estimate the state transient probabilities by the BW algorithm.

For the RPCL-based acoustic model training, in addition to the allocation by Eq. (4), we also consider another two choices (a) and (b) as follows:

$$p_{j,t} = \begin{cases} 1 + (1+\gamma_t)p(r_t|x_t), & j = c_t, \\ -p(r_t|x_t)\gamma_t, & j = r_t, \\ 0, & \text{others}, \end{cases}$$
$$\gamma_t = \begin{cases} 1 - p(c_t|x_t)\rho(x_t), & \text{choice (a)}, \\ 1 - \rho(x_t), & \text{choice (b)}, \\ \text{small } \gamma > 0, & \text{choice (c)}, \end{cases} \quad (10)$$

where $p_{j,t}$ is computed by approximately letting $p(c_t|x_t) \approx 1$ in Eq. (6), and γ_t is computed by a further stepwise use of $p(c_t|x_t) \approx 1$ in different levels. It should be noted that choice (c) equivalently implements Eq. (4) if considering $1 + \gamma_t \approx 1$.

During the learning process, we consider the following average posteriori probability of the correct states on the training data:

$$F_{\mathrm{RPCL}}(\theta) = \frac{1}{T} \sum_{t=1}^{T} p(c_t|x_t), \quad (11)$$

where c_t is the winner state that corresponds to the identity of the input x_t by the Viterbi force alignment. F_{RPCL} is used to control the iterative learning procedure to avoid over-training. The learning procedure stops while

it is not increasing or just gets little increment.

3.2 Experimental setup

The speech corpus employed in this paper is the continuous Mandarin speech corpora 863-I, which was provided by Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development. It 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 the proposed model, we also test the proposed model 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 using 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 ML criterion and the BW update formulas. Then, the proposed state-level RPCL-based model was trained according to Fig. 2 with details given in Sect. 3.1. For comparing with the traditional discriminative method, the lattice-based MMIE and MPE based acoustic model were also trained and tested on the two test sets. The lattice-based MMIE and MPE methods are implemented by the HTK 3.4 toolkit. I-smoothing is used in MMIE and MPE based method and their configuration is set to the recommended values in tutorial example [12]. The language model is a word-based trigram built on 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. The performance evaluation metric used in Mandarin speech recognition experiments is the Chinese character error rate (CER).

3.3 Experimental results

Table 1 shows the CER results of the proposed RPCL-embedded acoustic model training algorithm in its first 10 iterations. The three choices by Eq. (10) are implemented respectively, where the choice (c) is implemented by Eq. (4) under different de-learning strengths $\gamma = 0.1$, $\gamma = 0.2$, and $\gamma = 0.3$. The results show that the choice (c) outperforms the choices (a) and (b) on both two test sets, while the choice (c) with $\gamma = 0.3$ is the best for 863-I-Test and the one with $\gamma = 0.2$ is the best for Hub-4-Test data. On the contrary, in Fig. 3 the values of the criterion F_{RPCL} by Eq. (11) under choices (a) and (b) are higher than that under choice (c). This observation is understandable, because F_{RPCL} indicates the accuracy of state alignment per frame x_t from the training data, whereas CER reflects the accuracy of joint state alignments per character represented by a sequence of frames $\{x_t\}$ from the test data. Also, as the de-learning strength γ grows, the performance of the choice (c) becomes closer to those of (a) and (b) while the stability of the choice (c) declines. Moreover, the choices (a) and (b) given in Eq. (10) are more close to the BYY harmony learning. It suggests that considering learning and de-learning directly on the phone or word level may make the BYY harmony learning further improve the CER performance.

Moreover, we summarize the results of the RPCL approach in Table 2, in comparisons with MLE, as well as other two traditional discriminative training (DT) methods. It can be observed that all methods perform better on the 863-I-Test which well matches the training set than on the Hub-4-Test from an unmatched corpus, and

Table 1 CER results of RPCL learning on 863-I-Test and Hub-4-Test varies with time

	863-I-Test/%					Hub-4-Test/%				
	a	b	c			a	b	c		
			$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$			$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$
1	13.78	13.71	13.45	13.26	13.37	26.34	26.43	25.72	25.76	25.72
2	13.78	13.47	13.39	13.49	13.54	25.97	25.97	25.71	25.60	25.54
3	13.36	13.49	13.36	13.28	13.28	**25.95**	**25.90**	25.66	25.43	25.41
4	13.24	13.28	13.32	13.26	13.08	26.02	26.03	25.67	25.33	25.54
5	13.32	**13.13**	13.23	13.12	13.02	26.16	26.16	25.71	25.48	25.51
6	**13.23**	13.15	13.34	13.02	12.89	26.44	26.34	25.61	25.54	25.43
7	13.45	13.26	13.30	13.08	13.10	26.41	26.50	25.61	25.77	25.64
8	13.41	13.26	13.21	13.08	**12.87**	26.35	26.17	25.65	25.32	25.27
9	13.43	13.41	**13.06**	**13.02**	12.97	26.18	26.12	**25.28**	25.36	25.42
10	13.41	13.34	13.08	13.08	12.99	25.96	25.92	25.58	**25.17**	**25.24**

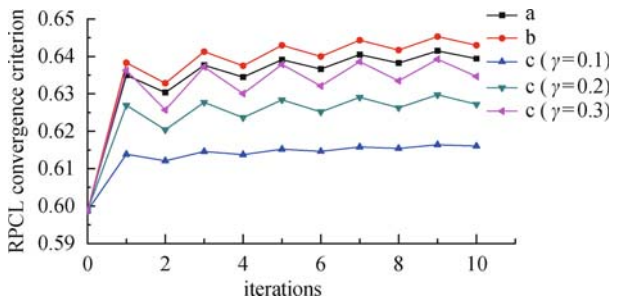


Fig. 3 Convergence criterion of RPCL learning varies with time

all DT methods outperform the classical MLE on both two test sets. To further compare the relative strengths of different DT methods, we calculate the relative reduction (RR) percentages of each DT method with respective the MLE which serves as a baseline. As shown in Table 2, the observations are as follows: 1) MPE has an advantage in the matched test set but deteriorates rapidly when encountering an unmatched test set; 2) The RPCL implementations obtain evident improvements over MLE and also is better than MMIE for the matched test set, and become superior (especially the choice (c)) for the unmatched test set, which indicates that our proposed RPCL approach exhibits a better generalization ability.

Table 2 Performance comparison on 863-I-Test and Hub-4-Test for different discriminative methods

	863-I-Test		Hub-4-Test	
	CER/%	RR/%	CER/%	RR/%
MLE	13.67	—	26.61	—
RPCL(a)	13.23	3.21	25.95	2.48
RPCL(b)	13.13	3.95	25.90	2.67
RPCL(c) $\gamma = 0.1$	13.06	4.46	25.28	5.00
RPCL(c) $\gamma = 0.2$	13.02	4.75	25.17	5.41
RPCL(c) $\gamma = 0.3$	12.87	5.85	25.24	5.15
MMIE	13.28	2.85	26.11	1.88
MPE	12.38	9.44	26.14	1.77

3.4 Discussion

In the past few years, extensions of MPE have been developed and achieve good performances on English and Arabic speech recognition systems [13]. However, when evaluated on Mandarin broadcast news in term of CER, MPE outperforms its two popular extensions, namely minimum phone frame error (MPFE) and physical-state level version of minimum Bayes risk (sMBR) based methods [14]. Therefore, we consider MPE for Mandarin speech recognition in this paper. However, MPE was shown in Ref. [5] to have poor performance on the not-well matched test data set, which is also confirmed by our results in Table 2.

In our experiment in Table 2, MMIE only obtained 2.85% and 1.88% relative reductions on 863-I-Test and Hub-4-Test respectively. This result is reasonable, because the performance of MMIE was shown in Ref. [15] to deteriorate as the number k of Gaussian components of GMM becomes large, and a large $k = 32$ is adopted in this paper. Selecting an appropriate k is a model selection problem, and there was a recent effort in Ref. [6] towards this goal for the HMM-based LVCSR system. Moreover, as in Ref. [5], the performance of MMIE also deteriorates when we proceed from a matched test set to an unmatched one, which is confirmed by the results in Table 2.

In the literature, MCE is another discriminative learning approach for speech recognition task. In an early study [16], the string-level MCE was shown to has similar performance with MMIE-based method on small vocabulary tasks. Moreover, studies in recent years [17–19] investigated lattice-based MCE methods, which have comparative performance with MPE-based method on the large vocabulary tasks. Therefore, although we did not include MCE's results in Table 2, the comparisons of the RPCL approach with MMIE and MPE still indicate that the RPCL discriminative learning is promising in speech recognition, especially that the RPCL approach has a very good generalization performance on the unmatched test set which makes the task of speech recognition very difficult.

Figure 3 and Table 1 imply that the current per-frame state-level discriminative learning scheme may not be the best for a CER result. Extending our method on other levels (word, phone, Gaussian) and using more prior knowledge (language model, speaker information) for updating candidate rival set, may further improve the performance.

4 Conclusion

We propose an RPCL-based discriminative acoustic model training method for LVCSR system. The winner state is enhanced by learning while its rival is penalized by de-learning, which makes GMMs-based hidden states become more discriminative. Experiments on LVCSR tasks show that the proposed method has a good convergence with better performances than the baseline model trained by the classical ML criterion. In comparison with two typical discriminative training methods based on MMI and MPE, respectively, the proposed method demonstrates improved generalization ability, especially when the sources of test sets are different from ones of training set.

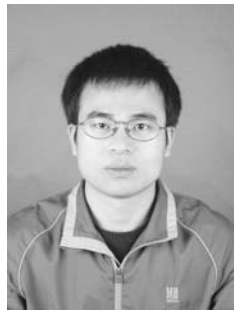
The method is currently implemented on the state-level acoustic models, and it is planned to be used on other levels (word, phone, Gaussian). Moreover, the RPCL mechanism may be replaced by a general BYY harmony learning for a further improvement.

Acknowledgements The work was supported in part by the National Natural Science Foundation of China (Grant No. 90920302), the National Key Basic Research Program of China (No. 2009CB825404), the HGJ Grant (No. 2011ZX01042-001-001), a research program from Microsoft China, and by a GRF grant from the Research Grant Council of Hong Kong SAR (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. Brown P. The acoustic-modeling problem in automatic speech recognition. Dissertation for the Doctoral Degree. Pittsburgh: Carnegie Mellon University, 1987
2. Gales M, Young S. The application of hidden Markov models in speech recognition. Foundations and Trends in Signal Processing, 2008, 1(3): 195–304
3. Bahl L, Brown P, De Souza P, Mercer R. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: Proceedings of 1986 IEEE International Conference on Acoustics, Speech, and Signal Processing. 1986, 11: 49–52
4. Juang B H, Chou W, Lee C H. Minimum classification error rate methods for speech recognition. IEEE Transactions on Speech and Audio Processing, 1997, 5(3): 257–265
5. Povey D, Woodland P C. Minimum phone error and I-smothing for improved discriminative training. In: Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. 2002, 1: 105–108
6. Su D, Wu X H, Xu L. GMM-HMM acoustic model training by a two level procedure with Gaussian components determined by automatic model selection. In: Proceedings of 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing. 2010, 4890–4893
7. Xu L. Bayesian Ying-Yang system, best harmony learning, and five action circling. Frontiers of Electrical and Electronic Engineering in China, 2010, 5(3): 281–328
8. Xu L, Krzyzak A, Oja E. Unsupervised and supervised classifications by rival penalized competitive learning. In: Proceedings of the 11th International Conference on Pattern Recognition. 1992, II: 672–675
9. Xu L. Rival penalized competitive learning. Scholarpedia, 2007, 2(8): 1810
10. Xu L. A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving. Pattern Recognition, 2007, 40(8): 2129–2153
11. Kuhn H W. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, 1955, 2(1–2): 83–97
12. Young S, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P. The HTK Book Version 3.4. Cambridge: Cambridge University Press, 2006
13. Povey D, Kingsbury B. Evaluation of proposed modifications to MPE for large scale discriminative training. In: Proceedings of 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing. 2007, 4: IV-321–IV-324
14. Cheng Y J, Lin C K, Lee L S. Evaluation and analysis of minimum phone error training and its modified versions for large vocabulary Mandarin speech recognition. In: Proceedings of 2008 IEEE International Symposium on Chinese Spoken Language Processing. 2008, 1: 157–160
15. Valtchev V, Odell J J, Woodland P C, Young S J. MMIE training of large vocabulary recognition systems. Speech Communication, 1997, 22(4): 303–314
16. McDermott E, Katagiri S. String-level MCE for continuous phoneme recognition. In: Proceedings of EuroSpeech 1997. 1997, 123–126
17. Macherey W, Haferkamp L, Schluter R, Ney H. Investigations on error minimizing training criteria for discriminative training in acoustic speech recognition. In: Proceedings of EuroSpeech 2005. 2005, 2133–2136
18. Schlter R, Macherey W, Mller B, Ney H. Comparison of discriminative training criteria and optimization methods for speech recognition. Speech Communication, 2001, 34(3): 287–310
19. Fu Q, He X, Deng L. Phone-discriminating minimum classification error (P-MCE) training criteria for phonetic recognition. In: Proceedings of InterSpeech 2007. 2007, 2073–2076

Zaihu PANG received the B.S. degree in Computer Science and Technology from the Jilin University, Changchun, China, in 2006. He is currently pursuing the Ph.D degree at the Speech and Hearing Research Center, Peking University, Beijing, China. His principal interest is in 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. He obtained his Bachelor degree from School of Mathematical Science, Peking University, in 2006. His research interests include statistical learning, pattern recognition, and bioinformatics.

Dan SU received the B.S. degree in Electrical Engineering from the Shandong University, Jinan, China, in 2004, the Ph.D degree from the Speech and Hearing Research Center, Peking University, Beijing, China, in 2010. His research interest is manly in large vocabulary speech recognition.

Xihong WU received the B.S. degree from the Jilin University, Changchun, China, in 1989, the M.S. degree from the Institute of Harbin Shipbuilding Engineering in China in 1992, and the Ph.D degree from the Department of Radio Electronics, Peking University, Beijing, China, in 1995. From 1995 to 1997, he was a Postdoctoral Researcher in the National Laboratory on Machine Perception, Peking University. Then, he joined the Peking University. In 1999, he became an Associate Professor, and in 2004, he became a Full Professor at Peking University. He has been elected a senior member of IEEE in 2009. Currently, he is the Director of the Speech and Hearing Research Center, and the Deputy Dean of Key Laboratory of Machine Perception (Ministry of Education) at Peking University. His areas of research focus include computational auditory models and auditory scene analysis, auditory psychophysics, speech signal processing, and natural language processing.

Lei XU, 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, China, and an Honorary Professor with Xidian University (See Front. Electr. Electron. Eng. China, 2011, 6(1): 119 for a detailed introduction).