Discriminative Training of Acoustic Models for System Combination

Tachioka, Y.; Watanabe, S.

TR2013-074 August 2013

Abstract

In discriminative training methods, the objective function is designed to improve the performance of automatic speech recognition with reference to correct labels using a single system. On the other hand, system combination methods, which output refined hypotheses by a majority voting scheme, need to build multiple systems that generate complementary hypotheses. This paper aims to unify the both requirements within a discriminative training framework based on the mutual information criterion. That is, we construct complementary models by optimizing the proposed objective function, which yields to minimize the mutual information with base systems' hypotheses, while maximize that with correct labels, at the same time. We also analyze that this scheme corresponds to weight the training data of a complementary system by considering correct and error tendencies in the base systems, which has close relationship with boosting methods. In addition, the proposed method can practically construct complementary systems by simply extending a lattice-based parameter update algorithm in discriminative training, and can adjust the degree of how much the complementary system outputs are different from base system ones. The experiments on highly noisy speech recognition ('The 2nd CHiME challenge') show the effectiveness of the proposed method, compared with a conventional system combination approach.

Interspeech 2013

Discriminative training of acoustic models for system combination

Yuuki Tachioka[1] *and Shinji Watanabe*[2]

[1]Information Technology R&D Center, Mitsubishi Electric, Kamakura, Japan
[2]Mitsubishi Electric Research Laboratories, Cambridge, US
Tachioka.Yuki@eb.MitsubishiElectric.co.jp, watanabe@merl.com

Abstract

In discriminative training methods, the objective function is designed to improve the performance of automatic speech recognition with reference to correct labels using a single system. On the other hand, system combination methods, which output refined hypotheses by a majority voting scheme, need to build multiple systems that generate complementary hypotheses. This paper aims to unify the both requirements within a discriminative training framework based on the mutual information criterion. That is, we construct complementary models by optimizing the proposed objective function, which yields to minimize the mutual information with base systems' hypotheses, while maximize that with correct labels, at the same time. We also analyze that this scheme corresponds to weight the training data of a complementary system by considering correct and error tendencies in the base systems, which has close relationship with boosting methods. In addition, the proposed method can practically construct complementary systems by simply extending a lattice-based parameter update algorithm in discriminative training, and can adjust the degree of how much the complementary system outputs are different from base system ones. The experiments on highly noisy speech recognition ('The 2$^{\text{nd}}$ CHiME challenge') show the effectiveness of the proposed method, compared with a conventional system combination approach.

Index Terms: discriminative training, margin training, boosting, system combination, MMI

1. Introduction

Over the past decade, the performance of Automatic Speech Recognition (ASR) [1] has been greatly improved, owing significantly to discriminative training methods for acoustic models [2, 3, 4, 5, 6, 7, 8] migrated from Maximum Likelihood (ML) estimation. These approaches improve the performance with reference to correct labels using a single system.

On the other hand, approaches based on combinations of systems (e.g., Recognizer Output Voting Error Reduction (ROVER) [9] and [10, 11]) can obtain refined hypotheses by majority voting of the hypotheses of the base and complementary systems, resulting in higher performance than the base system alone, even if the performance of complementary systems is lower than that of the base system. Because effective system combination relies on a combination of hypotheses with different trends [12], generally, different features or training methods are used to construct complementary systems with different output trends [13, 14, 15, 16]. For example, the random forest approach [13] is a simple realization of constructing complementary systems which builds multiple shared tri-phone trees by randomly changing the topologies of existing trees.

However, system combinations do not necessarily improve the performance when the hypotheses of complementary systems have similar trends or yield too many errors. To address this problem, conventional approaches prepare a number of systems and obtain the optimal system combination by selecting a few systems from among them, based on the performance of a development set. These trial and error approaches may select a combination overly tuned to a specific task that lacks robustness with respect to new data. Therefore, it is desirable that appropriate complementary systems be constructed with some theoretical (training) criteria. For example, the use of a confusion network for constructing complementary systems [14] is a promising direction that utilizes the minimum Bayes risk criterion, although its relationship to training criteria (e.g., discriminative training) is somewhat unclear.

In machine learning, 'boosting' has been widely studied for theoretical support of system combinations. The most popular approach (AdaBoost [17, 18]) combines many 'weak' classifiers that perform slightly higher than random classifiers, and provides performance improvements. This learning algorithm aims to minimize exponential errors [19] and allows successive classifiers to have a different hypothesis trends from previous classifiers by assigning greater weight to training data which is misclassified by classifiers in previous iterations. AdaBoost is effective in simple classification problems (e.g., binary classification), but cannot be applied in a simple manner to ASR problems, which are complex sequential classification problems.

There have been several attempts to apply boosting to ASR problems [15, 16]. For example, the boosting Baum-Welch algorithm [15] is a frame-wise boosting in the Baun-Welch algorithm, which leads to intensive training of the statistics that have a low likelihood at Baum-Welch algorithm iterations. The aim of this method is to refine models by considering the output trend of the base system in Baum-Welch algorithm iterations, but it is not used for constructing complementary systems.

Our proposed method uses a lattice-based discriminative training framework which is extended to construct complementary system models for system combination. Although there are several training methods available [2], this paper focuses on Maximum Mutual Information (MMI) training [7]. Our method proposes to generalize the MMI objective function in order to consider the hypotheses of the base systems by minimizing the mutual information to the hypotheses of the base systems, while maximizing the mutual information to correct labels. The advantages of our proposed method are simple extension of conventional lattice-based discriminative training and clear resemblance to a discriminative training method because it updates parameters from a lattice for discriminative training. The update formulae of model parameters can analytically provide an interesting interpretation similar to boosting that weights training data considering whether the hypotheses of the base systems are correct. In addition, because the formulation of our proposed method includes the margin-based (boosted) MMI, it can adjust the extent of deviation of complementary systems outputs with respect to those of the base systems. This paper describes conventional discriminative training and the proposed system combination in Section 2 and 3, and shows experimental demonstration of the effectiveness of the proposed approach in Section 5.

2. MMI discriminative training

MMI training aims to maximize the following objective function (Eq. (1)) for correct labels in reference to hypotheses in a lattice, which is generated by an initial model (e.g., ML model).

$$\mathcal{F}(\lambda) = \ln \frac{P_\lambda(s_r, \mathbf{x}_t)}{\sum_s P_\lambda(s, \mathbf{x}_t)} = \ln \frac{p_\lambda\left(\mathbf{x}_t|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda\left(\mathbf{x}_t|\mathcal{H}_s\right)^\kappa p_L(s)}, \quad (1)$$

where λ and $\mathbf{x}_t$ are the acoustic model parameters to be optimized and the tth frame feature vector sequence. The summation over utterances is omitted for readability. The product of an acoustic model likelihood p_λ (with acoustic scale κ) and a language model likelihood p_L is denoted by $P_\lambda(s, \mathbf{x}_t)$. The acoustic likelihood is conditioned on $\mathcal{H}_{s_r}$ or $\mathcal{H}_s$, which is the HMM sequence of a correct label s_r or a hypothesis s, respectively.

In the boosted MMI (bMMI)[1] [20], the standard MMI objective function is modified to include a term that enhances (boosts) the effect of hypotheses with low phoneme accuracy:

$$\mathcal{F}^{\mathrm{b}}(\lambda) = \ln \frac{P_\lambda(s_r, \mathbf{x}_t)}{\sum_s P_\lambda(s, \mathbf{x}_t) e^{(-bA(s,s_r))}}, \quad (2)$$

where $A(s, s_r)$ is the frame-wise phoneme accuracy of s for a reference s_r. Update formulae for the mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$ of GMM (HMM state j and Gaussian index m) are given as

$$\begin{aligned}\boldsymbol{\mu}'_{jm} &= \frac{\sum_t \Delta_{jm,t}\mathbf{x}_t + D_{jm}\boldsymbol{\mu}_{jm}}{\sum_t \Delta_{jm,t} + D_{jm}},\\ \mathbf{\Sigma}'_{jm} &= \frac{\sum_t \Delta_{jm,t}\mathbf{x}_t\mathbf{x}_t^T + D_{jm}(\mathbf{\Sigma}_{jm} + \mathbf{U}_{jm})}{\sum_t \Delta_{jm,t} + D_{jm}} - \mathbf{U}'_{jm},\end{aligned} \quad (3)$$

where $\Delta_{jm,t}$ is $\gamma^{num}_{jm,t} - \gamma^{den}_{jm,t}$, and $\gamma^{num}_{jm,t}$ and $\gamma^{den}_{jm,t}$ are the posterior of numerator and denominator of Eq. (1) or (2). $\mathbf{U}_{jm}$ and $\mathbf{U}'_{jm}$ denote $\boldsymbol{\mu}_{jm}\boldsymbol{\mu}^T_{jm}$ and $\boldsymbol{\mu}'_{jm}\boldsymbol{\mu}'^{\,T}_{jm}$, respectively. These update formulae are introduced by approximating the update formulae for discrete HMM optimization [21]. The Gaussian-specific learning-rate constants D_{jm} lead to a positive definite $\mathbf{\Sigma}'_{jm}$. The mixture weights of GMM are also optimized [20]. Algorithm 1 shows the MMI or bMMI algorithm, where i_{eb} is the number of iterations (e.g., four in [20]) of the extended Baum-Welch.

Algorithm 1 Construct MMI or bMMI model

```
Input: ML model $mdl$, numerator ($s_r$ aligned) lattice $\mathcal{A}$ , and
  denominator lattice $\mathcal{L}$ of Eq. (1) or (2)
  for $i = 1$ to $i_{eb}$ do
    Rescore $\mathcal{A}$ and $\mathcal{L}$ with $mdl$
    $\gamma^{num}_{jm,t}$ and $\gamma^{den}_{jm,t}$ $\Leftarrow$posteriors of $\mathcal{A}$ and $\mathcal{L}$, respectively
    $\gamma_{jm,t} \Leftarrow -\gamma^{den}_{jm,t} + \gamma^{num}_{jm,t}$
    $\gamma^{num}_{jm,t}, \gamma^{den}_{jm,t}$ $\Leftarrow$positive and negative parts of $\gamma_{jm,t}$
    $mdl \Leftarrow$ Update $\boldsymbol{\mu}$, $\mathbf{\Sigma}$ by Eq. (3).
  end for
Output: MMI or bMMI model ($mdl$)
```

3. Discriminative training for complementary systems

In this paper, complementary systems are constructed using an initial model (e.g., ML) and base system models (e.g., MMI or bMMI). We propose a discriminative training method for complementary systems by extending (boosted) MMI[2]. This method is relevant to the 'boosting' method for large-scale series data. In order to consider the hypotheses of the base systems, the MMI objective function in Eq. (1) is generalized to the following proposed objective function, which minimizes the mutual information to the hypotheses of the base systems, while maximizing the mutual information to correct labels.

$$\begin{aligned}\mathcal{F}_{\mathrm{c}}(\lambda_c) = &\underbrace{\ln\left(\frac{P_{\lambda_{\mathrm{c}}}(s_r, \mathbf{x}_t)}{\sum_s P_{\lambda_{\mathrm{c}}}(s, \mathbf{x}_t))}\right)^{1+\alpha}}_{\text{MI to the correct labels}}\\ &- \sum_{q=1}^{Q} \underbrace{\ln\left(\frac{P_{\lambda_{\mathrm{c}}}(s_{q,1}, \mathbf{x}_t)}{\sum_s P_{\lambda_{\mathrm{c}}}(s, \mathbf{x}_t))}\right)^{\frac{\alpha}{Q}}}_{\text{MI to the 1-best (q^{th} base system)}},\end{aligned} \quad (4)$$

where λ_c is the acoustic parameter of a complementary system to be optimized, $s_{q,1}$ is a 1-best hypothesis of the qth base system and α is a scaling factor for the complementary system. If α equals zero, this objective function matches that of bMMI. For simplicity, the number of base systems Q is taken as one below, and index q is omitted. With the bMMI extension, two boosting factors (b and b_1) are introduced into Eq. (4) as

$$\mathcal{F}^{\mathrm{b}}_{\mathrm{c}}(\lambda_c) = \mathcal{F}^{\mathrm{b}}(\lambda_{\mathrm{c}}) + \ln\left[\frac{P_{\lambda_c}(s_r, \mathbf{x}_t)}{P_{\lambda_c}(s_1, \mathbf{x}_t)e^{(b_1 A(s_1,s_r))}}\right]^\alpha. \quad (5)$$

Thus, we derived a new objective function for a complementary system within a MMI discriminative training framework.

Now, we discuss the relationship between the proposed objective function and conventional bMMI objective function, in detail. The role of the other factor b_1 is shown as Eq. (6), which is the same to Eq. (5), note that $e^{-b_1 A(s_r,s_r)}$ is not related to the optimization because it is always a constant (e^{-b_1}).

$$\mathcal{F}^{\mathrm{b}}_{\mathrm{c}}(\lambda_c) = \mathcal{F}^{\mathrm{b}}(\lambda_{\mathrm{c}}) + \ln\left[\frac{P_{\lambda_c}(s_r, \mathbf{x}_t)e^{(-b_1 A(s_r,s_r))}}{P_{\lambda_c}(s_1, \mathbf{x}_t)e^{(-b_1(1-A(s_1,s_r)))}}\right]^\alpha. \quad (6)$$

This equation shows that b_1 enhances (boosts) the 1-best hypothesis of a base system with a low error rate $(1 - A(s_1, s_r))$.

In addition, this equation has a close relationship with the original bMMI, which is clearly shown by

$$\mathcal{F}^{\mathrm{b}}_{\mathrm{c}}(\lambda_c) = \ln \frac{P_{\lambda_c}(s_r, \mathbf{x}_t)}{\sum_s P_{\lambda_c}(s_r, \mathbf{x}_t) e^{(-bA(s,s_r)+\delta)}}. \quad (7)$$

A margin shift δ is $\alpha(b_1 A(s_1, s_r) + \ln P_{\lambda_c}(s_1, \mathbf{x}_t) - \ln P_{\lambda_c}(s_r, \mathbf{x}_t))$, which decreases and updates parameters for correct labels for the case that the base system is incorrect or $P_{\lambda_c}(s_1, \mathbf{x}_t)$ is low. This margin shift changes the trends of hypotheses of the complementary systems from both initial and base models.

[1] To avoid the confusion of the term 'boost' in 'bMMI' and 'boosting', this paper basically uses 'boost' in the 'boosting' context.

[2] The MMI discriminative criterion is used in this paper, but this procedure is easily applied for other discriminative criteria.

The update formulae for the mean and covariance of GMM are restored to the original bMMI formulae (Eq. (3)) by simply modifying the variables as

$$
\begin{aligned}
&\Delta_{jm,t} \leftarrow (1+\alpha)\gamma_{jm,t}^{num} - (\gamma_{jm,t}^{den} + \alpha\gamma_{jm,t}^{1}), \\
&\gamma_{jm,t}^{num} \leftarrow \gamma_{jm,t}^{num}, \\
&\gamma_{jm,t}^{den} \leftarrow \frac{1+\alpha\frac{\gamma_{jm,t}^{1}}{\gamma_{jm,t}^{den}}}{1+\alpha}\gamma_{jm,t}^{den} = w_{jm,t}\gamma_{jm,t}^{den}, \\
&D_{jm} \leftarrow \frac{D_{jm}}{1+\alpha}.
\end{aligned}
\tag{8}
$$

Weights $w_{jm,t}$ depend on the relationship between the posterior of the 1-best hypothesis of the base system $\gamma_{jm,t}^{1}$ and $\gamma_{jm,t}^{den}$ as

$$
w_{jm,t} \begin{cases} >1 & \text{if } \gamma_{jm,t}^{1} > \gamma_{jm,t}^{den}, \\ =1 & \text{if } \gamma_{jm,t}^{1} = \gamma_{jm,t}^{den}, \\ <1 & \text{if } \gamma_{jm,t}^{1} < \gamma_{jm,t}^{den}. \end{cases}
\tag{9}
$$

When $\gamma_{jm,t}^{1}$ is large, $w_{jm,t}$ is large and Δ is small, in which case the parameters are changed only slightly, and the generated hypotheses are similar to those of the base model. Contrarily, when $\gamma_{jm,t}^{1}$ is small, parameters are changed drastically, and the generated hypotheses are different from those of base model. Therefore, when $\gamma_{jm,t}^{1}$ is small, it is highly likely that the hypotheses of the base systems are incorrect and the statistics are exaggerated. This shows that the weight $w_{jm,t}$ is interpreted as that of frame-wise 'boosting' to construct complementary systems. Moreover, factors (b and b_1) give smaller weight $w_{jm,t}$ for the case that the base system gives a incorrect hypothesis or the complementary system gives an correct hypothesis. This is because posteriors $\gamma_{jm,t}^{1}$ and $\gamma_{jm,t}^{den}$ are increasing functions of the base system accuracy ($e^{b_1 A(s_1)}$) and the complementary system error rate ($e^{-bA(s)}$), respectively. This mechanism is the same as that based on the exponential error in 'boosting' [17, 19].

Thus, the proposed method can be interpreted as a practical realization of 'boosting' for sequential and large-scale ASR problems, in addition to a general extension of MMI discriminative training. Algorithm 2 shows the proposed algorithm for updating a complementary system model by using the extended Baum-Welch algorithm.

Algorithm 2 Construct complementary system model

```
Input: ML model $mdl$, base system models $mdl_q$, numerator ($s_r$ aligned) lattice $\mathcal{A}$, and denominator lattice $\mathcal{L}$ of Eq. (1) or (2)
  for $i = 1$ to $i_{eb}$ do
    Rescore $\mathcal{A}$ and $\mathcal{L}$ with $mdl$
    $\gamma_{jm,t}^{num}$ and $\gamma_{jm,t}^{den}$ $\Leftarrow$posteriors of $\mathcal{A}$ and $\mathcal{L}$, respectively
    $\gamma_{jm,t} \Leftarrow -\gamma_{jm,t}^{den} + (1+\alpha)\gamma_{jm,t}^{num}$
    for $q = 1$ to $Q$ do
      Rescore $\mathcal{L}$ with $mdl_q$
      $\mathcal{L}_1$ $\Leftarrow$best path of $\mathcal{L}$
      Rescore $\mathcal{L}_1$ with $mdl$
      $\gamma_{jm,t}^{1} \Leftarrow$ posterior of $\mathcal{L}_1$
      $\gamma_{jm,t} \Leftarrow -\frac{\alpha}{Q}\gamma_{jm,t}^{1} + \gamma_{jm,t}$
    end for
    $\gamma_{jm,t}^{num}, \gamma_{jm,t}^{den}$ $\Leftarrow$positive and negative parts of $\gamma_{jm,t}$
    $mdl \Leftarrow$ Update $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ by Eq. (3) with Eq. (8).
  end for
Output: Complementary system model ($mdl$)
```

4. Experimental setup

We evaluated the performance improvement provided by these system combination techniques by using 2nd CHiME challenge Track 2, which is designed for evaluating the word error rate (WER) of a medium vocabulary task (Wall Street Journal (WSJ0)) under reverberated and non-stationary noisy environments [22]. The language model size was 5 k (basic). We used the Kaldi toolkit [23]. The training data set (si_tr_s) contained 7138 utterances from 83 speakers (si84), the evaluation data set (si_et_05) contained 330 utterances from 12 speakers (Nov'92), and the development set (si_dt_05) contained 409 utterances from 10 speakers. Acoustic models were trained using si_tr_s and the acoustic scale κ was tuned using si_dt_05. These data simulate realistic environments. Noise is non-stationary, such as other speakers' utterances, household noise, or music and is added to 'isolated' speech at SNR = $\{-6, -3, 0, 3, 6, 9\}$dB. Although the database provides two-channel data, we used noise-suppressed single-channel data by the prior-based binary masking [24].

We describe the settings of acoustic feature and feature transformation [25]. The baseline acoustic features were MFCC and PLP (1-13 order MFCCs (PLPs) + Δ + $\Delta\Delta$). In addition to this, feature transformation techniques (Linear Discriminant Analysis (LDA) [26], Maximum Likelihood Linear Transformation (MLLT) [27, 28], and Speaker Adaptive Training (SAT) [29]) and speaker adaptation technique (feature space Maximum Likelihood Linear Regression (fMLLR) [30]) were used.

The procedure of training acoustic models and the setup of feature transformations are described in [24, 25]. The number of the context-dependent HMM states was 2500 and the total number of Gaussians was 15000. Tree structures were different between MFCC and PLP features, the latter of which considers a random forest-like effect. Parameters α and b_1 were 0.75 and 0.3, which were optimized by using development set.

5. Results and Discussion

Table 1 shows the WER of the development set using MFCC and PLP features. The upper, upper middle, lower middle, and lower sections correspond to conventional single systems (S1-S4), ROVER among conventional multiple systems (R1-R4), proposed systems (P1,P2), and ROVER including proposed systems (RP1-RP6), respectively. The WER using PLP is lower than that using MFCC, but the combination of bMMI (MFCC and PLP) improves the WER by 1.88% (S2→R2). This shows the importance of combining different hypotheses as well as the effectiveness of using different features. The performance of the complementary system model (bMMI$_c$) is lower than that of ML (MFCC) or between the performances of ML and bMMI (PLP), but the combination of bMMI and bMMI$_c$ improves the performance of bMMI by 0.34% (MFCC and PLP, S2→RP1 / S4→RP2). Moreover, the combination of ML, bMMI, and bMMI$_c$ improves the WER of the combination of ML and bMMI by 0.27% (R1→RP3). The addition of bMMI$_c$ to the combination of ML and bMMI with MFCC and PLP improves the WER by 0.34% and obtains the best WER (R4→RP6). This shows the effectiveness of the proposed method.

Table 2 shows the WER using MFCC and PLP features with the feature transformation of LDA+MLLT+SAT+fMLLR. The trends are almost the same to those above. In this case, because the performance of ML is notably lower than that of bMMI, the combination with the ML model is not effective for ROVER. In this case, the performance of the combination of ML and bMMI

Table 1: Average WER[%] for isolated speech (**si_dt_05**) with noise suppression by prior-based binary masking. (MFCC and PLP) (upper: conventional Single systems (S), upper middle: ROVER among conventional multiple systems (R), lower middle: single Proposed systems (P), and lower: ROVER including Proposed system (RP))

ID	MFCC			PLP			WER
	ML	bMMI	bMMI$_c$	ML	bMMI	bMMI$_c$	
S1	✓						46.88
S2		✓					45.59
S3				✓			48.23
S4					✓		46.65
R1	✓	✓					45.11
R2		✓			✓		43.71
R3	✓	✓			✓		43.56
R4	✓	✓		✓	✓		**43.39**
P1			✓				46.95
P2						✓	47.53
RP1		✓	✓				45.25
RP2					✓	✓	46.31
RP3	✓	✓	✓				44.84
RP4	✓	✓	✓		✓		43.79
RP5	✓	✓	✓		✓	✓	43.17
RP6	✓	✓	✓	✓	✓	✓	**43.05**

Table 2: Average WER[%] for isolated speech (**si_dt_05**). (MFCC and PLP with LDA+MLLT+SAT+fMLLR)

ID	MFCC			PLP			WER
	ML	bMMI	bMMI$_c$	ML	bMMI	bMMI$_c$	
S1	✓						38.15
S2		✓					35.86
S3				✓			38.10
S4					✓		36.43
R1	✓	✓					36.06
R2		✓			✓		**34.65**
R3	✓	✓			✓		34.95
R4	✓	✓		✓	✓		34.97
P1			✓				36.21
P2						✓	36.72
RP1		✓	✓				35.67
RP2					✓	✓	36.21
RP3	✓	✓	✓				35.56
RP4	✓	✓	✓		✓		34.95
RP5	✓	✓	✓		✓	✓	**34.54**
RP6	✓	✓	✓	✓	✓	✓	34.55

(R1) is lower than that of the combination of bMMI and bMMI$_c$ (RP1), even though the numbers of systems are the same (two) for both cases. In addition to this, the proposed system improves the WER by 0.11% (R2→RP5) because the performance of bMMI$_c$ is moderate, which makes system combination effective. This is an advantage of the performance adjustability of the proposed method.

Tables 3 and 4 show the WER of the evaluation set. The proposed method is also effective for the evaluation set and improves the WER by 0.34% (base feature) and 0.51% (transformed feature) (R4→RP6). Tables 5 and 6 show the WER in terms of SNR by comparing R1 with RP3 and R4 with RP6. For almost all cases (except some cases of R1→RP3 in Table 5), the proposed method improves the WER, especially for low SNR cases (1% maximum). Thus, the performance improvements are stable and robust in different environments.

Table 3: Average WER[%] for isolated speech (**si_et_05**) with noise suppression by prior-based binary masking. (MFCC and PLP)

ID	MFCC			PLP			WER
	ML	bMMI	bMMI$_c$	ML	bMMI	bMMI$_c$	
S1	✓						42.45
S2		✓					40.74
S3				✓			44.40
S4					✓		42.10
R1	✓	✓					40.03
R4	✓	✓		✓	✓		**38.64**
P1			✓				42.94
P2						✓	43.69
RP1		✓	✓				40.57
RP3	✓	✓	✓				40.23
RP6	✓	✓	✓	✓	✓	✓	**38.30**

Table 4: Average WER[%] for isolated speech (**si_et_05**). (MFCC and PLP with LDA+MLLT+SAT+fMLLR)

ID	MFCC			PLP			WER
	ML	bMMI	bMMI$_c$	ML	bMMI	bMMI$_c$	
S1	✓						32.20
S2		✓					29.46
S3				✓			32.23
S4					✓		29.98
R1	✓	✓					29.26
R4	✓	✓		✓	✓		**28.00**
P1			✓				30.09
P2						✓	30.46
RP1		✓	✓				28.80
RP3	✓	✓	✓				28.81
RP6	✓	✓	✓	✓	✓	✓	**27.49**

Table 5: WER[%] in terms of SNR[dB] for isolated speech (**si_et_05**). (MFCC and PLP)

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
R1	56.16	49.22	42.56	34.69	30.92	26.60	40.03
R4	55.76	47.86	40.78	33.44	28.97	**25.01**	38.64
RP3	55.82	48.96	42.59	35.16	31.46	27.37	40.23
RP6	**54.75**	**47.36**	**40.31**	**33.16**	**28.92**	25.29	**38.30**

Table 6: WER[%] in terms of SNR[dB] for isolated speech (**si_et_05**). (MFCC and PLP with LDA+MLLT+SAT+fMLLR)

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
R1	47.08	38.11	30.99	23.69	19.09	16.59	29.26
R4	45.86	36.63	29.16	22.36	18.59	15.39	28.00
RP3	46.39	38.00	30.32	23.20	18.77	16.18	28.81
RP6	**44.80**	**35.79**	**28.86**	**22.34**	**18.05**	**15.09**	**27.49**

6. Conclusions

We proposed a method of discriminative training of acoustic models for system combination. The proposed method can construct complementary systems in the framework of discriminative training methods, and it is capable of improving the WER on reverberated and highly noisy speech. In future work, the proposed method will be combined with some other discriminative techniques [8] (e.g., feature-space discriminative training and discriminative language modeling).

7. References

[1] J. Baker, L. Deng, J. Glass, S. Khudanpur, C. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding part 1," *IEEE Signal Processing Magazine*, vol. 26, pp. 75–80, 2009. 5.

[2] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition," *IEEE Signal Processing Magazine*, vol. 25, pp. 14–36, 2008. 9.

[3] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *in Proceedings ICASSP*, vol. 11, pp. 49–52, 1986.

[4] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," *in Proceedings ICASSP*, vol. I, pp. 105–108, 2002.

[5] E. McDermott, T. J. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 203–223, 2007. 1.

[6] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative training based on an integrated view of MPE and MMI in margin and error space," *in Proceedings ICASSP*, pp. 4894–4897, 2010.

[7] G. Heigold, H. J. Ney, R. Schlüter, and S. Wiesler, "Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance," *IEEE Signal Processing Magazine*, vol. 29, pp. 58–69, 2012. 11.

[8] M. Gales, S. Watanabe, and E. Fosler-Lussier, "Structured discriminative models for speech recognition: an overview," *IEEE Signal Processing Magazine*, vol. 29, pp. 70-81, 2012. 11.

[9] J. Fiscus, "A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER)," *in Proceedings ASRU*, pp. 347–354, 1997.

[10] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," *in Proceedings NIST Speech Transcription Workshop*, 2000.

[11] B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney, "Frame based system combination and a comparison with weighted ROVER and CNC," *in Proceedings ICSLP*, pp. 537–540, 2006.

[12] T. Shinozaki and S. Furui, "Strategies for model training and adaptation based on data dependency control," *APSIPA Overview*, 2011.

[13] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," *in Proceedings ICASSP*, pp. 197–200, 2005.

[14] C. Breslin and M. Gales, "Generating complementary systems for speech recognition," *in Proceedings ICASSP*, pp. 337–340, 2007.

[15] H. Tang, M. Hasegawa-Johnson, and T. S. Huang, "Toward robust learning of the Gaussian mixture state emission densities for hidden Markov models," *in Proceedings ICASSP*, pp. 5242–5245, 2010.

[16] G. Saon and H. Soltau, "Boosting systems for LVCSR," *in Proceedings INTERSPEECH*, pp. 1341–1344, 2010.

[17] Y. Freund and R. Schapire, "A dicision-theoretic generalisation of online learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, pp. 119–139, 1997. 8.

[18] P. Viola and M. Jones, "Robust real-time object detection," *in Proceedings Second International Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling*, pp. 1–25, 2001. 7.

[19] J. Friedman, T. Hestie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, vol. 28, pp. 337–407, 2000.

[20] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," *in Proceedings ICASSP*, pp. 4057–4060, 2008.

[21] Y. Normandin and S. D. Morgera, "An improved MMIE training algorithm for speaker-independent, small vocabulary, continuous speech recognition," *in Proceedings ICASSP*, vol. 1, pp. 537–540, 1991.

[22] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," *in Proceedings ICASSP*, 2013.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," *in Proceedings ASRU*, pp. 1–4, 2011.

[24] Y. Tachioka, S. Watanabe, J. Le Roux, and J. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," *The 2nd International Workshop on Machine Listening in Multisource Environments*, 2013.

[25] Y. Tachioka, S. Watanabe, and J. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," *in Proceedings ICASSP*, 2013.

[26] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," *in Proceedings ICASSP*, pp. 13–16, 1992.

[27] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," *in Proceedings ICASSP*, pp. 661–664, 1998.

[28] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272–281, 1999. 3.

[29] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," *in Proceedings ICSLP*, pp. 1137–1140, 1996.

[30] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.