A GENERALIZED DISCRIMINATIVE TRAINING FRAMEWORK FOR SYSTEM COMBINATION

Yuuki Tachioka

Information Technology R&D Center,
Mitsubishi Electric, Kamakura, Japan

Shinji Watanabe, Jonathan Le Roux, John R. Hershey

Mitsubishi Electric Research Laboratories,
Cambridge, USA

ABSTRACT

This paper proposes a generalized discriminative training framework for system combination, which encompasses acoustic modeling (Gaussian mixture models and deep neural networks) and discriminative feature transformation. To improve the performance by combining base systems with complementary systems, complementary systems should have reasonably good performance while tending to have different outputs compared with the base system. Although it is difficult to balance these two somewhat opposite targets in conventional heuristic combination approaches, our framework provides a new objective function that enables to adjust the balance within a sequential discriminative training criterion. We also describe how the proposed method relates to boosting methods. Experiments on highly noisy middle vocabulary speech recognition task (2nd CHiME challenge track 2) and LVCSR task (Corpus of Spontaneous Japanese) show the effectiveness of the proposed method, compared with a conventional system combination approach.

Index Terms— discriminative training, margin training, boosting, system combination

1. INTRODUCTION

Many researchers have pointed out that combining different systems effectively improves performance (e.g., Recognizer Output Voting Error Reduction (ROVER) [1] and [2, 3]) even if the performance of the complementary systems is lower than that of the base system. Because effective system combination relies on a combination of hypotheses with different trends, generally, different features or training methods are used to construct complementary systems [4, 5, 6, 7]. For example, the random forest approach [4] is a simple way of constructing complementary systems, which builds multiple shared tri-phone trees by randomly changing the topologies of existing trees. Especially for Deep Neural Networks (DNN), to avoid local minimum problems, random initialization and averaging of multiple model parameters are generally used to improve the performance of the original single system. However, system combinations do not necessarily improve the performance when the hypotheses of complementary systems have similar trends or yield too many errors (as we also confirmed in our experiments). Classical system combination approaches require trial-and-error attempts because they do not rely on a general theoretical background such as an objective function in discriminative training [8, 9, 10].

To address this problem, complementary system training algorithm of acoustic models for system combination based on the Minimum Phoneme Error (MPE) criterion has been proposed [11]. This lattice-based approach provides theoretical background for training complementary systems and is promising because conventional discriminative training methods can be easily applied. We also proposed a method to discriminatively train acoustic models based on the Maximum Mutual Information (MMI) criterion in order to clarify the relationship between the reference and hypotheses of the base and complementary system further [12].

In this paper, we extend the above approach and propose a general framework of sequential discriminative training for system combination encompassing various model training methods such as acoustic modeling, here applied to Gaussian Mixture Models (GMM) and DNN, as well as discriminative feature transformation. Our method generalizes the objective function of discriminative training in order to balance the objective function given by correct labels and that given by the hypotheses of the base systems. The advantages of our proposed method are the fact it leads to a simple extension of conventional lattice-based discriminative training and its clear resemblance to a discriminative training method. In addition, because the formulation of our proposed method includes margin-based discriminative training, one can adjust the degree of deviation of the complementary systems' outputs with respect to those of the base systems. Thus, the effectiveness of the proposed approach covers the wide area of discriminative acoustic modeling and feature transformation.

This paper first describes the general discriminative training framework for complementary systems in Section 2. Then, we apply this framework to sequential discriminative training of acoustic models (GMM and DNN) and discriminative feature transformation in Sections 3 and 4, respectively. Sections 5 and 6 show the effectiveness of the proposed approach experimentally.

2. GENERALIZED DISCRIMINATIVE TRAINING FRAMEWORK FOR COMPLEMENTARY SYSTEMS

In this paper, complementary systems are constructed by discriminatively training a model starting from an initial model. The proposed discriminative training method for complementary systems is extended from a discriminative training principle. Assuming Q base systems have already been constructed, the discriminative training objective function $\mathcal{F}$ is generalized to the following proposed objective function $\mathcal{F}^c$, which subtracts from the original objective function involving the correct labels ω_r, the objective functions involving the 1-best hypotheses (lattice) $\omega_{q,1}$ ($q = 1, \ldots, Q$) of the Q base systems:

$$\mathcal{F}^c_\varphi(\omega_r, \omega_{q,1}) = (1+\alpha)\mathcal{F}_\varphi(\omega_r) - \frac{\alpha}{Q}\sum_{q=1}^{Q}\mathcal{F}_\varphi(\omega_{q,1}), \tag{1}$$

where φ is the set of model parameters of a complementary system to be optimized and α is a scaling factor. The 1-best hypotheses can be easily obtained by the lattice rescoring. If α equals zero, this objective function matches that of classical discriminative training. The first term in Eq. (1) promotes good performance according to the discriminative training criterion, whereas the second term makes the target system generate hypotheses that have a different tendency from the original base models. The next sections provide concrete forms of the objective function and model parameters in Eq. (1) for acoustic modeling problems and discriminative feature transformation.

3. DISCRIMINATIVE TRAINING OF ACOUSTIC MODELS

This section applies the MMI criterion to the above-mentioned framework. MMI training aims to maximize the following objective function for a correct label sequence ω_r in reference to hypotheses ω in a lattice, which is generated by an initial model (e.g., ML model):

$$\mathcal{F}_{\lambda}^{\mathrm{MMI}}(\omega_r) = \ln \frac{P_{\lambda}(\omega_r, \mathbf{X})}{\sum_{\omega} P_{\lambda}(\omega, \mathbf{X})}, \tag{2}$$

$$= \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p_{\lambda}(s_r, \mathbf{X})^{\kappa} p_L(\omega_r)}{\sum_{\omega} \sum_{s \in \mathcal{S}_{\omega}} p_{\lambda}(s, \mathbf{X})^{\kappa} p_L(\omega)}, \tag{3}$$

where λ is the set of HMM parameters to be optimized and $\mathbf{X} = \{\mathbf{x}_t | t = 1, \cdots, T\}$ is a feature vector sequence for the T-frame uttarance. The summation over utterances is omitted for readability. The product of the acoustic model score p_{λ} with a HMM state sequence s (and acoustic scale κ) and the language model score p_L is denoted by $P_{\lambda}(\omega, \mathbf{X})$. In Eq. (3), the acoustic score is obtained by the summation over a reference s_r or s, and $\mathcal{S}_{\omega_r}$ or $\mathcal{S}_{\omega}$ is a set of the HMM state sequences, which outputs a correct label ω_r or a hypothesis ω, respectively. For simplicity, the number of base systems Q is taken as one below, and the index q is omitted.

In the MMI criterion, we replace φ by λ_c and $\mathcal{F}$ by $\mathcal{F}^{\mathrm{MMI}}$ in Eq. (1) to obtain:

$$\mathcal{F}_{\lambda_c}^{\mathrm{c}}(\omega_r, \omega_1) = \mathcal{F}_{\lambda_c}^{\mathrm{MMI}}(\omega_r) + \alpha \ln \frac{P_{\lambda_c}(\omega_r, \mathbf{X})}{P_{\lambda_c}(\omega_1, \mathbf{X})}, \tag{4}$$

which is a new objective function for a complementary system within an MMI discriminative training framework, that has an additional log-likelihood ratio term.

In boosted MMI (bMMI) [9], the standard MMI objective function is modified to include a factor that enhances the effect of hypotheses with low accuracies:

$$\mathcal{F}_{\lambda}^{\mathrm{bMMI}}(\omega_r) = \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p_{\lambda}(s_r, \mathbf{X})^{\kappa} p_L(\omega_r)}{\sum_{\omega} \sum_{s \in \mathcal{S}_{\omega}} p_{\lambda}(s, \mathbf{X})^{\kappa} p_L(\omega) e^{-bA(s, s_r)}}, \tag{5}$$

where $A(s, s_r)$ is the state/phoneme/word accuracy calculated from the HMM state sequences of s for a reference s_r, which is computed frame by frame in our implementation. As a simple extension of the Eq. (4), by replacing $\mathcal{F}^{\mathrm{MMI}}$ with $\mathcal{F}^{\mathrm{bMMI}}$, and adding the (reverse sign) boosting factors to the log-likelihood ratio term analogous to Eq. (5), we can introduce the following objective function [1] [2]:

$$\mathcal{F}_{\lambda_c}^{\mathrm{c}}(\omega_r, \omega_1) = \mathcal{F}_{\lambda_c}^{\mathrm{bMMI}}(\omega_r) + \alpha \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p_{\lambda}(s_r, \mathbf{X})^{\kappa} p_L(\omega_r)}{\sum_{s_1 \in \mathcal{S}_{\omega_1}} p_{\lambda}(s_1, \mathbf{X})^{\kappa} p_L(\omega_1) e^{b_1 A(s_1, s_r)}}, \tag{6}$$

where s_1 is an HMM state sequence corresponding to the 1-best hypothesis of the base system ω_1. The reverse sign boosting factor b_1 is discussed in the Section 3.1. This procedure is commonly used to obtain the objective functions of acoustic modeling (GMM and DNN) and discriminative feature transformation in this paper.

3.1. Gaussian Mixture model (GMM)

In GMM training, Eq. (3) is broken down into the update formulae for the mean $\boldsymbol{\mu}_{jm}$ and covariance $\mathbf{\Sigma}_{jm}$ of GMM (HMM state j and Gaussian index m) as

$$\begin{aligned} \boldsymbol{\mu}'_{jm} &= \frac{\sum_t \Delta_{jm,t} \mathbf{x}_t + D_{jm} \boldsymbol{\mu}_{jm}}{\sum_t \Delta_{jm,t} + D_{jm}}, \\ \mathbf{\Sigma}'_{jm} &= \frac{\sum_t \Delta_{jm,t} \mathbf{x}_t \mathbf{x}_t^{\top} + D_{jm}(\mathbf{\Sigma}_{jm} + \mathbf{U}_{jm})}{\sum_t \Delta_{jm,t} + D_{jm}} - \mathbf{U}'_{jm}, \end{aligned} \tag{7}$$

where $\Delta_{jm,t}$ is $\gamma_{jm,t}^{num} - \gamma_{jm,t}^{den}$, $\gamma_{jm,t}^{num}$ and $\gamma_{jm,t}^{den}$ are the numerator and denominator of the posteriors of Eq. (3) or (5), and $\top$ denotes the transpose. $\mathbf{U}_{jm}$ and $\mathbf{U}'_{jm}$ denote $\boldsymbol{\mu}_{jm}\boldsymbol{\mu}_{jm}^{\top}$ and $\boldsymbol{\mu}'_{jm}\boldsymbol{\mu}'^{\top}_{jm}$, respectively. These update formulae are introduced by approximating the update formulae for discrete HMM optimization [13]. The Gaussian-specific learning rate D_{jm} is set to make Σ'_{jm} positive definite. The mixture weights π_{jm} of GMM can be also optimized [9].

We now explain the update equation of the complementary system by using the proposed objective function (6). The update formulae for the mean and covariance of GMM take the same form as the original (b)MMI formulae (7) up to simply modifying the variables as ($\gamma_{jm,t}^{num}$ is unchanged)

$$\begin{aligned} \Delta'_{jm,t} &= (1 + \alpha) \left(\gamma_{jm,t}^{num} - \gamma_{jm,t}^{den\ \prime} \right), \\ \gamma_{jm,t}^{den\ \prime} &= \frac{\gamma_{jm,t}^{den} + \alpha \gamma_{jm,t}^{1}}{1 + \alpha}, \\ D'_{jm} &= \frac{D_{jm}}{1 + \alpha}. \end{aligned} \tag{8}$$

To elucidate the effect of the b_1 term, we first consider for simplicity, the single-frame classification problem of the proposed approach, which is approximated by assuming an utterance has only one frame. In a single frame, because we do not need to consider the HMM states transition, posteriors are proportional to the product of acoustic and language scores multiplied by the boosting factors. In this case, the index t is omitted, and γ_{jm}^{1} can be represented by

$$\gamma_{jm}^{1} = \begin{cases} C_{jm}^{1} e^{b_1} & (j,\ s.t.,\ s_1 = s_r,\ \text{correct}), \\ C_{jm}^{1} & (j,\ s.t.,\ s_1 \neq s_r,\ \text{incorrect}), \end{cases} \tag{9}$$

$$\begin{aligned} C_{jm}^{1} &= \frac{p_{\lambda}(j, m, \mathbf{x})^{\kappa} p_L(\omega_1)}{\sum_{m', j'_1 \in \mathcal{S}_{\omega_1}} p_{\lambda}(j'_1, m', \mathbf{x})^{\kappa} p_L(\omega_1) e^{b_1 A(j'_1, j_r)}}, \\ p_{\lambda}(j, m, \mathbf{x}) &= \pi_{jm} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jm}, \mathbf{\Sigma}_{jm}), \end{aligned} \tag{10}$$

where $\mathcal{N}$ is a probability density of a single Gaussian and j_1 and j_r are the HMM states obtained from the 1-best hypotheses of the base system and the label, respectively. The factor b_1 decreases γ_{jm}^{1} in the case that the base system gives incorrect hypotheses. Because γ_{jm}^{1} is subtracted in Eq. (8), diminishing it increases $\Delta_{jm,t}$ for these hypotheses. This is analogous to boosting algorithms such as AdaBoost [14, 15], which assign larger weights to data points where the base system gives incorrect hypotheses.

[1] There is another derivation obtained by substituting the bMMI criterion Eq. (5) into our generalized form Eq. (1). We will further investigate the relationship in our future work.

[2] Note that because there are multiple HMM state sequences realizing the same phoneme/word sequence, the denominator of the second term in Eq. (6) is obtained by the summation over these multiple sequences, and thus the boosting factor b_1 do affect the optimization.

For the sequential case, it is difficult to show a direct relationship between the posterior and the boosting factors as in the single frame case because, to gather posteriors, the forward-backward algorithm is used and posteriors at the current frame are affected by the previous and future frames. However, similarly to the discussion in the single frame case, because the posterior $\gamma^{1}_{jm,t}$ is an increasing function of the base system's sentence average accuracy, even in the sequential case, the proposed method has a relationship to boosting.

Algorithm 1 shows the proposed algorithm for updating a complementary system model by using the extended Baum-Welch (EBW) algorithm or gradient descent (GD). In this paper, EBW algorithm was used.

Algorithm 1 Construct complementary system model for GMM

```
Input: Initial model $\lambda$ (e.g., ML), base system models $\lambda_q$, numerator ($\omega_r$ aligned) lattice $\mathcal{A}$, and denominator lattice $\mathcal{L}$ of Eq. (2) or (5)
  for $i = 1$ to $i_{eb}$ do
    Rescore $\mathcal{A}$ and $\mathcal{L}$ with $\lambda$
    $\gamma^{num}_{jm,t}$ and $\gamma^{den}_{jm,t}$ $\Leftarrow$posteriors are gathered on $\mathcal{A}$ and $\mathcal{L}$, respectively
    $\gamma_{jm,t} \Leftarrow -\gamma^{den}_{jm,t} + (1+\alpha)\gamma^{num}_{jm,t}$
    for $q = 1$ to $Q$ do
      Rescore $\mathcal{L}$ with $\lambda_q$
      $\mathcal{L}_1$ $\Leftarrow$best path of $\mathcal{L}$
      Rescore $\mathcal{L}_1$ with $\lambda$
      $\gamma^{1}_{jm,t}$ $\Leftarrow$ posteriors are gathered on $\mathcal{L}_1$
      $\gamma_{jm,t} \Leftarrow -\frac{\alpha}{Q}\gamma^{1}_{jm,t} + \gamma_{jm,t}$
    end for
    $\gamma^{num}_{jm,t}, \gamma^{den}_{jm,t}$ $\Leftarrow$positive and negative parts of $\gamma_{jm,t}$
    $\lambda \Leftarrow$ Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by EBW or GD (Eq. (7))
  end for
Output: Complementary system model ($\lambda_c \leftarrow \lambda$)
```

3.2. Deep Neural Networks

In a DNN-HMM hybrid system, sequential discriminative training according to the MMI criterion (2) has been proposed [16, 17, 18]. The DNN provides posterior probabilities for the HMM state j. Acoustic likelihood p_θ is replaced by pseudo likelihood as

$$p_\theta\left(\mathbf{x}_t|j\right) = \frac{p_\theta\left(j|\mathbf{x}_t\right)}{p_0\left(j\right)}, \tag{11}$$

where $p_0\left(j\right)$ is the prior probability calculated from the training data. For each HMM state, the model θ includes a softmax activation function p_θ:

$$p_\theta(j|\mathbf{x}_t) = \frac{\exp a(j|\mathbf{x}_t)}{\sum_{j'}\exp a(j'|\mathbf{x}_t)}, \tag{12}$$

where a is the activation at the output layer. These activations are trained discriminatively according to MMI criterion. The MMI objective function is the same as Eq. (6), simply replacing λ by θ. The update rule for activation a is obtained by differentiating the objective function:

$$\begin{aligned}\frac{\partial\mathcal{F}^{\mathrm{bMMI}}}{\partial a(j)} &= \sum_{j'}\frac{\partial\mathcal{F}^{\mathrm{bMMI}}}{\partial\log p_\theta\left(\mathbf{x}_t|j'\right)}\frac{\partial\log p_\theta\left(\mathbf{x}_t|j'\right)}{\partial a(j)}, \\ &= \kappa(\gamma^{num}_{j,t} - \gamma^{den}_{j,t}).\end{aligned} \tag{13}$$

For the proposed method, the denominator posterior is modified by Eq. (8) as in the GMM case. The gradients for all the DNN parameters are derived from Eq. (13) based on the back-propagation procedure.

Algorithm 2 shows that the method for constructing complementary system models for DNN is similar to the GMM case. This versatility is one of the advantages of the proposed generalized framework.

Algorithm 2 Construct complementary system model for DNN

```
Input: Initial model $\theta$, base system models $\theta_q$, numerator ($\omega_r$ aligned) lattice $\mathcal{A}$, and denominator lattice $\mathcal{L}$ of Eq. (2) or (5)
  for $i = 1$ to $i_{eb}$ do
    Rescore $\mathcal{A}$ and $\mathcal{L}$ with $\theta$
    $\gamma^{num}_{j,t}$ and $\gamma^{den}_{j,t}$ $\Leftarrow$posteriors are gathered on $\mathcal{A}$ and $\mathcal{L}$, respectively
    $\gamma_{j,t} \Leftarrow -\gamma^{den}_{j,t} + (1+\alpha)\gamma^{num}_{j,t}$
    for $q = 1$ to $Q$ do
      Rescore $\mathcal{L}$ with $\theta_q$
      $\mathcal{L}_1$ $\Leftarrow$best path of $\mathcal{L}$
      Rescore $\mathcal{L}_1$ with $\theta$
      $\gamma^{1}_{j,t}$ $\Leftarrow$ posteriors are gathered on $\mathcal{L}_1$
      $\gamma_{j,t} \Leftarrow -\frac{\alpha}{Q}\gamma^{1}_{j,t} + \gamma_{j,t}$
    end for
    $\gamma^{num}_{j,t}, \gamma^{den}_{j,t}$ $\Leftarrow$positive and negative parts of $\gamma_{j,t}$
    $\theta \Leftarrow$ Update $a$ by EBW or GD (Eq. (13))
  end for
Output: Complementary system model ($\theta_c \leftarrow \theta$)
```

4. DISCRIMINATIVE FEATURE TRANSFORMATION

In addition to discriminative training of acoustic models, feature transformation based on the discriminative training criterion can be used [19]. This method estimates a matrix $\mathbf{M}$ that projects from high-dimensional (L-dimensional) non-linear features to low-dimensional (K-dimensional) transformed features as:

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{M}\mathbf{h}_t, \tag{14}$$

where $\mathbf{h}_t$ is the nonlinear feature, and $\mathbf{y}_t$ is the transformed feature. The matrix $\mathbf{M}$ is $K \times L$-dimensional and trained on the MMI criterion, which is then called feature-space MMI (f-MMI) and can also be extended to a boosted version (f-bMMI). After substituting $\mathbf{y}$ of Eq. (14) for $\mathbf{x}$ into Eq. (5), we obtain the objective function for f-bMMI:

$$\mathcal{F}^{\text{f-bMMI}}_{\mathbf{M}}(\omega_r) = \ln\frac{\sum_{s_r\in\mathcal{S}_{\omega_r}} p_{\mathbf{M}}\left(s_r, \mathbf{Y}\right)^{\kappa} p_L(\omega_r)}{\sum_{\omega}\sum_{s\in\mathcal{S}_\omega} p_{\mathbf{M}}\left(s, \mathbf{Y}\right)^{\kappa} p_L(\omega) e^{-bA(s,s_r)}}, \tag{15}$$

where $\mathbf{Y}$ is a feature vector $\{\mathbf{y}_t|t = 1,\cdots,T\}$. Differentiating the objective function $\mathcal{F}$ by $\mathbf{M}$ as

$$\frac{\partial\mathcal{F}}{\partial\mathbf{M}} = \left[\begin{array}{ccc}\frac{\partial\mathcal{F}}{\mathbf{y}_1} & \cdots & \frac{\partial\mathcal{F}}{\mathbf{y}_{T_f}}\end{array}\right]\left[\begin{array}{ccc}\mathbf{h}_1 & \cdots & \mathbf{h}_{T_f}\end{array}\right]^{\top}, \tag{16}$$

where T_f is the total number of frames of training data. The optimized matrix $\mathbf{M}$ is obtained by gradient descent using the (b)MMI statistics. Indirect differential of the objective function is given by

$$\frac{\partial\mathcal{F}}{\partial\mathbf{y}_t} = \sum_j\sum_m\frac{\gamma^{ML}_{jm,t}}{\sum_t\gamma^{ML}_{jm,t}}\left[\frac{\partial\mathcal{F}}{\partial\boldsymbol{\mu}_{jm,t}} + 2\frac{\partial\mathcal{F}}{\partial\boldsymbol{\Sigma}_{jm,t}}(\mathbf{y}_t - \boldsymbol{\mu}_{jm,t})\right], \tag{17}$$

where $\gamma_{jm,t}^{ML}$ is the ML model posterior and $\frac{\partial \mathcal{F}}{\partial \boldsymbol{\mu}_{jm,t}}$ and $\frac{\partial \mathcal{F}}{\partial \boldsymbol{\Sigma}_{jm,t}}$ have been already obtained by the (b)MMI discriminative training of acoustic models [19]. To form the features, N components of the GMM are obtained by clustering the Gaussians in the initial triphone acoustic models into N components and re-estimating their parameters. The non-linear feature $\mathbf{h}_t$ [20] is calculated as

$$\mathbf{h}_{t,n} = \left[p_{t,n} \frac{x_{t,1} - \mu_{n,1}}{\sigma_{n,1}}, \cdots, p_{t,n} \frac{x_{t,K} - \mu_{n,K}}{\sigma_{n,K}}, \beta p_{t,n} \right]^{\top}, \tag{18}$$

where $\mu_{n,k}$ and $\sigma_{n,k}$ are k^{th} dimensional mean and standard deviation parameters of the n^{th} Gaussian component. β is the scaling factor. $p_{t,n}$ are Gaussian component posteriors computed for each frame, which are approximated such that all but the N_1-best posteriors are set to zero. This approximation is undertaken in order to reduce computational cost by ensuring that $\mathbf{h}_t$ is sparse.

As in the GMM case, the objective function for complementary systems is introduced from Eq. (1) by replacing φ by $\mathbf{M}_c$ and $\mathcal{F}$ by $\mathcal{F}^{\text{f-MMI}}$ ($b = 0$ for Eq. (15)) as

$$\mathcal{F}_{\mathbf{M}_c}^{\text{c}}(\omega_r, \omega_1) = \mathcal{F}_{\mathbf{M}_c}^{\text{f-MMI}}(\omega_r) + \alpha \ln \frac{P_{\mathbf{M}_c}(\omega_r, \mathbf{Y})}{P_{\mathbf{M}_c}(\omega_1, \mathbf{Y})}, \tag{19}$$

and, in the same procedure from Eq. (4) to Eq. (6), the boosted version of Eq. (19) is given by

$$\begin{aligned} \mathcal{F}_{\mathbf{M}_c}^{\text{c}}(\omega_r, \omega_1) &= \mathcal{F}_{\mathbf{M}_c}^{\text{f-bMMI}}(\omega_r) \\ &+ \alpha \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p_{\mathbf{M}_c}(s_r, \mathbf{Y})^{\kappa} p_L(\omega_r)}{\sum_{s_1 \in \mathcal{S}_{\omega_1}} p_{\mathbf{M}_c}(s_1, \mathbf{Y})^{\kappa} p_L(\omega_1) e^{-b_1 A(s_1, s_r)}}. \end{aligned} \tag{20}$$

Thus the proposed framework can be applied to the discriminative feature transformation for a complementary system starting from the generalized objective function.

Algorithm 3 shows the proposed algorithm for updating a complementary system model by using the gradient descent algorithm.

Algorithm 3 Construct complementary system model for f-MMI

```
Input: Acoustic model $\lambda$, initial matrix $\mathbf{M}$, base system matrix $\mathbf{M}_q$,
  numerator ($\omega_r$ aligned) lattice $\mathcal{A}$, and denominator lattice $\mathcal{L}$ of
  Eq. (15)
  for $i = 1$ to $i_{eb}$ do
    Rescore $\mathcal{A}$ and $\mathcal{L}$ with $\lambda$ using $\mathbf{y}_t$ $(= \mathbf{x}_t + \mathbf{M}\mathbf{h}_t)$
    $\gamma_{jm,t}^{num}$ 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 $\lambda$ using $\mathbf{y}_t$ $(= \mathbf{x}_t + \mathbf{M}_q\mathbf{h}_t)$
      $\mathcal{L}_1$ $\Leftarrow$best path of $\mathcal{L}$
      Rescore $\mathcal{L}_1$ with $\lambda$
      $\gamma_{jm,t}^{1} \Leftarrow$ 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 $\gamma_{jm,t}$
    $\mathbf{M} \Leftarrow$ Update elements in $\mathbf{M}$ by calculating the indirect differential in Eq. (17)
  end for
Output: Complementary system matrix ($\mathbf{M}_c \leftarrow \mathbf{M}$)
```

5. EXPERIMENTAL SETUP

We evaluated the performance improvement provided by these system combination techniques on two corpus: the 2$^{\text{nd}}$ CHiME challenge Track 2 and Corpus of Spontaneous Japanese. The former aimed to validate the performance of the proposed method for acoustic modeling (GMM and DNN) and discriminative feature transformation and the effectiveness of our proposed generalized framework experimentally. The latter aimed to show that the proposed method is effective for other tasks and the performance improvement does not depend on tasks. The 2$^{\text{nd}}$ CHiME challenge Track 2 was designed for evaluating the word error rate (WER) of a medium vocabulary task (Wall Street Journal (WSJ0)) under reverberated and non-stationary noisy environments [21]. The language model size was 5 k (basic). 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 simulated realistic environments. Noise was non-stationary, such as other speakers' utterances, household noise, or music and was added to 'isolated' speech at SNR $= \{-6, -3, 0, 3, 6, 9\}$dB. Although the database provided two-channel data, we used noise-suppressed single-channel data obtained by prior-based binary masking [22].

The settings of acoustic feature and feature transformation was as follows [23]. We used the Kaldi toolkit [24]. The baseline features were MFCC and PLP (1-13 order MFCCs/PLPs + Δ + $\Delta\Delta$). Feature transformation techniques (Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transformation (MLLT)) and speaker adaptation technique (Speaker Adaptive Training (SAT) and feature space Maximum Likelihood Linear Regression (fMLLR)) were used.

The procedure of training acoustic models and the setup of feature transformations are described in [22, 23]. The number of the context-dependent HMM states was 2,500 and the total number of Gaussians was 15,000. Tree structures were different between MFCC and PLP features, the latter of which also considered a random forest-like effect. For the DNN, we used a CPU version of neural network training implemented in Kaldi with 3 hidden layers and 1,000,000 parameters. The initial learning rate was 0.01 and was decreased to 0.001 at the end of training. In discriminative feature transformation, 400 Gaussians were used and offset features were calculated for each of the 40 dimensional features with context expansion (9 frames). The dimension of the feature vector $\mathbf{h}_t$ was 400 × 40 × 9. Features with the top 2 posteriors were selected and all other features were ignored. β was set to 5. For the proposed method, parameters α and b_1 were 0.75 and 0.3, which were optimized by using the development set.

Corpus of Spontaneous Japanese (CSJ) is a lecture-style LVCSR speech recognition task [25]. Test set 1 contained about 10-15 minutes lecture by 10 different male speakers. The ASR settings were similar to the CHiME challenge, but the language model size was about 70k and the number of the context-dependent HMM states was 3,500 and the total number of Gaussians was 96,000. The parameters for the proposed method are the same to those for the CHiME challenge.

We used ROVER for combining output hypotheses from multiple systems. Certainly, especially for two systems, confusion network combination (CNC) is better than ROVER, however, ROVER is more simple and can be applied for many systems.

6. RESULTS AND DISCUSSION

6.1. CHiME challenge (Noise robust ASR)

For the GMM system, although detailed results are shown in [12], we briefly describe the results for comparison with the other approaches. Table 1 shows the WER using MFCC and PLP features with the feature transformation of LDA+MLLT and SAT+fMLLR. The upper, upper middle, lower middle, and lower sections correspond to conventional single systems (S1-S4), ROVER among conventional multiple systems (R1,R2), proposed complementary systems (P1,P2), and ROVER including proposed complementary systems (RP1,RP2), respectively. The performances of proposed complementary systems (P1,P2) were in between that of ML(S1,S3) and that of bMMI(S2,S4). Because the performance of ML was much lower than that of bMMI, the combination with the ML model was not effective for ROVER (S2→R1). In this case, even though the numbers of systems were the same (two) for both cases, the performance of the combination of bMMI and bMMI$_\mathrm{c}$ (RP1) was higher than that of the combination of ML and bMMI (R1) because the performance of bMMI$_\mathrm{c}$ was moderate, which made the system combination effective. This is an advantage of the performance adjustability of the proposed method. Adding two systems to the conventional ROVER using four systems further improved the WER by 0.42%(dt) and 0.51%(et) (R2→RP2). Because the hypotheses of MFCC systems are quite different from those of PLP systems, alternative update of the complementary system for both feature systems could not improve the performance.

In addition, we validated the discriminative feature transformation and DNN on the development set. Table 2 (left column) shows the WER using discriminative feature space transformation on top of MFCC features with the feature transformation of LDA+MLLT and SAT+fMLLR. f-bMMI is usually combined with discriminative training of GMM (i.e., bMMI). In this case, we constructed complementary systems in two ways: for both f-bMMI and bMMI, the objective functions were modified (i.e., f-bMMI$_\mathrm{c}$ + bMMI$_\mathrm{c}$ using Eqs. (20) and (6)) or only for f-bMMI, the objective function was modified (i.e., f-bMMI$_\mathrm{c}$ + bMMI using Eqs. (20) and (5)). The performance of the combination of bMMI and f-bMMI (R3) was lower than that of f-bMMI only, but the combination with the proposed complementary systems (RP3 and RP4) improved the accuracy. There was no significant difference between f-bMMI$_\mathrm{c}$ + bMMI and f-bMMI$_\mathrm{c}$ + bMMI$_\mathrm{c}$. Table 3 shows the WER using DNN on top of MFCC and PLP features with the feature transformation of LDA+MLLT. Discriminative training improved the accuracy by 4.19% (S7→S8) significantly. Combination with the proposed method also improved the accuracy further (RP5).

Table 1. Average WER[%] for isolated speech (**si_dt_05** and **si_et_05**) on acoustic modeling (GMM). (MFCC and PLP with LDA+MLLT+SAT+fMLLR) (upper: conventional Single systems (S), upper middle: ROVER among conventional multiple systems (R), lower middle: single Proposed complimentary systems (P), and lower: ROVER including Proposed complementary system (RP))

ID	MFCC			PLP			WER	
	ML	bMMI	bMMI$_\mathrm{c}$	ML	bMMI	bMMI$_\mathrm{c}$	(dt)	(et)
S1	✓						38.15	32.20
S2		✓					35.86	29.46
S3				✓			38.10	32.23
S4					✓		36.43	29.98
R1	✓	✓					36.06	29.26
R2	✓	✓		✓	✓		34.97	28.00
P1			✓				36.21	30.09
P2						✓	36.72	30.46
RP1		✓	✓				35.67	28.80
RP2	✓	✓	✓	✓	✓	✓	34.55	27.49

Table 2. Average WER[%] for isolated speech (**si_dt_05**, **si_et_05**) on discriminative feature transformation. (MFCC with LDA+MLLT and SAT+fMLLR)

ID	bMMI	f-bMMI	f-bMMI$_\mathrm{c}$ + bMMI$_\mathrm{c}$	f-bMMI$_\mathrm{c}$ +bMMI	WER	
					(dt)	(et)
S5	✓				35.86	29.46
S6		✓			**33.19**	**27.00**
R3	✓	✓			33.80	27.15
P3			✓		35.38	28.27
P4				✓	33.88	27.86
RP3		✓	✓		32.75	**26.60**
RP4		✓		✓	**32.67**	26.62

Table 3. Average WER[%] for isolated speech (**si_dt_05**, **si_et_05**) on acoustic modeling (DNN). (MFCC with LDA+MLLT)

ID	DNN	bMMI	bMMI$_\mathrm{c}$	WER	
				(dt)	(et)
S7	✓			36.59	30.84
S8		✓		**32.40**	**26.91**
P5			✓	33.09	27.97
RP5		✓	✓	**31.38**	**26.48**

Table 4. WER[%] in terms of SNR[dB] for isolated speech (**si_et_05**) on f-bMMI (S6→RP3) and DNN (S8→RP5).

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
S6	44.14	35.42	28.56	21.46	17.41	**14.98**	27.00
S8	43.86	33.36	28.13	22.01	17.75	16.36	26.91
RP3	43.21	34.24	28.25	21.58	**17.17**	15.13	26.60
RP5	**42.85**	**32.43**	**27.91**	**21.56**	17.75	16.40	**26.48**

Table 5. Average WER[%] (CSJ, test set 1) on acoustic modeling (GMM). (MFCC)

ID	ML	bMMI	bMMI$_\mathrm{c}$	WER
S1	✓			21.00
S2		✓		**18.64**
R1	✓	✓		18.69
P1			✓	18.81
RP1		✓	✓	18.52
RP2	✓	✓	✓	**18.28**

We also validated the performance on the evaluation set, and confirmed the similar experimental tendencies. Table 4 further investigates the WER in terms of SNR by comparing S6 with RP3 (f-bMMI case) and S8 with RP5 (DNN case). For almost all the cases, the proposed method improved the WER, especially for the low SNR cases (1.2% maximum). Thus, the performance improvements were stable and robust in different environments.

In conclusion, the experimental results confirmed the effectiveness of the proposed approach for a wide range of sequential discriminative training methods for acoustic modeling and feature transformation.

6.2. CSJ (LVCSR)

The performance was evaluated on a second corpus (CSJ). This task did not include noises but it was composed of spontaneous speech and the vocabulary size was much larger than the CHiME challenge (WSJ0). Table 5 shows the WER for the test set 1 by using the proposed GMM training. In this case, conventional ROVER (R1) decreased the performance from the single system (S2), however, the proposed method using two or three systems improved the accuracy by 0.36%. Thus, the proposed approach was also effective for large-scale spontaneous speech recognition.

7. CONCLUSIONS

We proposed a general discriminative training framework 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 as well as large vocabulary spontaneous speech recognition tasks. Moreover, it is effective for discriminative training of acoustic models (GMM and DNN) and discriminative feature transformation. In future work, the proposed method will be combined with other discriminative techniques, such as acoustic modeling with other discriminative criteria and discriminative language modeling [8].

8. REFERENCES

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

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

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

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

[5] C. Breslin and M.J.F. Gales, "Generating complementary systems for speech recognition," *in Proc. INTERSPEECH*, pp. 525-528, 2006.

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

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

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

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

[10] B. Roark, M. Saraçlar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," *in Proc. ACL*, pp. 47–54, 2004.

[11] F. Diehl and P.C. Woodland, "Complementary Phone Error Training," *in Proc. INTERSPEECH*, 2012.

[12] Y. Tachioka and S. Watanabe, "Discriminative training of acoustic models for system combination," *in Proc. INTERSPEECH*, pp. 2355–2359, 2013.

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

[14] Y. Freund and R.E. 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.

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

[16] B. Kingsbury, T.N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," *in Proc. INTERSPEECH*, pp. 485–488, 2012.

[17] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," *in Proc. INTERSPEECH*, pp. 2345–2349, 2013.

[18] Y. Kubo, T. Hori, and A. Nakamura, "Large vocabulary continuous speech recognition based on WFST structured classifiers and deep bottleneck features," *in Proc. ICASSP*, pp. 7629–7633, 2013.

[19] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," *in Proc. ICASSP*, pp. 961–964, 2005.

[20] D. Povey, "Improvements to fMPE for discriminative training of features," *in Proc. INTERSPEECH*, pp. 2977–2980, 2005.

[21] 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 Proc. ICASSP*, pp. 126–130, 2013.

[22] Y. Tachioka, S. Watanabe, J. Le Roux, and J.R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," *in Proc. The 2nd International Workshop on Machine Listening in Multisource Environments*, pp. 19–24, 2013.

[23] Y. Tachioka, S. Watanabe, and J.R. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," *in Proc. ICASSP*, pp. 6935–6939, 2013.

[24] 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 Proc. ASRU*, pp. 1–4, 2011.

[25] K. Maekawa, H. Koiso, S. Furui, H. Isahara, "Spontaneous speech corpus of Japanese," *in Proc. LREC2000*, vol. 2, pp. 947–952, 2000.