Regular Paper

Prior-based Binary Masking and Discriminative Methods for Reverberant and Noisy Speech Recognition Using Distant Stereo Microphones

Yuuki Tachioka[1,a)] Shinji Watanabe[2] Jonathan Le Roux[2] John R. Hershey[2]

Received: June 2, 2016, Accepted: March 3, 2017

Abstract: Reverberant and noisy automatic speech recognition (ASR) using distant stereo microphones is a very challenging, but desirable scenario for home-environment speech applications. This scenario can often provide prior knowledge such as physical information about the sound sources and the environment in advance, which may then be used to reduce the influence of the interference. We propose a method to enhance the binary masking algorithm by using prior distributions of the time difference of arrival. This paper also validates state-of-the-art ASR techniques that include various discriminative training and feature transformation methods. Furthermore, we develop an efficient method to combine discriminative language modeling and minimum Bayes risk decoding in the ASR post-processing stage. We also investigate the effectiveness of this method when used for reverberated and noisy ASR of deep neural networks (DNNs) as well when used in systems that combine multiple DNNs using different features. Experiments on the medium vocabulary sub-task of the second CHiME challenge show that the system submitted to the challenge achieved a 26.86% word error rate (WER), moreover, the DNN system with the discriminative training, speaker adaptation and system combination achieves a 20.40% WER.

Keywords: CHiME challenge, noise-robust ASR, prior-based binary masking, discriminative methods, feature transformation, deep neural networks, system combination

1. Introduction

Automatic speech recognition (ASR) is a fundamental component of various speech interfaces; ASR has many applications in environments such as in the home or in the car. In such scenarios, close-talking input is often impractical or unsafe, and, while very challenging, allowing the speaker to be far from the microphone is highly desirable. To validate the effectiveness of state-of-the-art speech enhancement and ASR techniques in distant-talking conditions, several challenges have been organized [4], [22], [44]. Among these, the Computational Hearing in Multisource Environments (CHiME) challenges recently introduced noise-robust speech processing tasks with a small number of microphones [4], [44]. The goal of these tasks is to recognize speech from a distant target speaker that was binaurally recorded in a domestic environment. Whereas the first CHiME challenge is a simple keyword recognition task [4], the second CHiME challenge contains a medium vocabulary recognition task (track 2). In particular, track 2 contains simulated speech samples that are taken from the Wall Street Journal (WSJ0) 5k vocabulary read speech corpus, convolved with binaural room impulse responses, and then mixed with binaural recordings of a noisy domestic environment [44]. The second challenge is much more complex and difficult from a speech recognition point of view. To overcome this challenging task, we propose a system involving state-of-the-art and newly-proposed components, including a noise suppression method as well as various discriminative training and feature transformation methods.

We propose a binary masking (BM) method based on the estimated time difference of arrival (TDOA) that is to be used for noise suppression; this method takes advantage of the availability of the binaural training data provided by the challenge. If many microphones are available, linear noise suppression techniques are effective and generate little distortion [20]. When only two microphones are used, however, one can expect SNR improvements of up to only 3 dB when using techniques such as standard delay-and-sum beamforming. Therefore, one needs to resort to non-linear methods for better performance. One such non-linear method is a BM technique based on the TDOA that has been shown to be simple and effective for a small number of microphones [37]. However, the TDOA estimation accuracy can be severely degraded in the presence of reverberation and noise [40]. To compensate for the influence of reverberation and noise, we propose to use the training data to generate a prior distribution of the discrepancy between the instantaneous inter-microphone phase difference and the expected phase difference of sound emanating from the target speaker location. That prior distribution is then used when building the binary mask. We refer to this approach as prior-based BM.

In this paper, the goal is not only to improve the baseline recognition systems by using source-separation-based approaches, but also to understand to what extent performance can be improved by using the discriminative training ASR approach; this allows researchers who may not be experts in ASR to better evaluate

[1] Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura, Kanagawa 247–8501, Japan
[2] Mitsubishi Electric Research Laboratories, Cambridge, MA, US
[a)] Tachioka.Yuki@eb.MitsubishiElectric.co.jp

the benefit of these methods. Recent improvements in ASR techniques have led to high-accuracy speech recognition systems [3], [36]. Over the past 20 years in particular, model training techniques have gradually migrated from maximum-likelihood (ML) estimation approaches to discriminative training techniques [2], [26], [29], [38]. In addition, various types of feature transformations have been proposed [1], [13], [15], [16], [17], [28]. While such state-of-the-art ASR techniques have been shown to be very effective in clean speech conditions, further investigation is needed in order to improve the effectiveness of ASR techniques in challenging conditions such as in the presence of environmental reverberation and noise. This paper proposes an approach to overcome these challenges by evaluating discriminative training and feature transformation techniques based on the samples provided in the second CHiME challenge. As the conditions between the training data and the test data are matched, it is reasonable to expect that discriminative training methods will lead to significant performance improvements even in reverberant and noisy conditions. In particular, we investigate the performance change when using maximum mutual information (MMI) and boosted MMI (bMMI) training. We also investigate several feature transformation approaches that include linear discriminant analysis (LDA) [16], maximum likelihood linear transformation (MLLT) [13], [15], and adaptation techniques such as speaker adaptive training (SAT) [1] and feature-space maximum likelihood linear regression (fMLLR) [12]. Other discriminative non-linear feature transformations such as feature-space boosted MMI (f-bMMI) were also investigated. LDA makes use of long context features across a few contiguous frames (e.g., nine frames) to exploit feature dynamics, which reduces the influence of non-stationary noises and reverberation. MLLT finds a linear transformation to reduce state-conditional feature correlations; it performs a joint optimization of feature transformation matrices and acoustic model parameters. Speaker adaptation methods such as SAT and fMLLR were originally developed for decreasing the variation between speakers, but they are also known to improve the ASR accuracy in noisy environments by adapting to unknown and changing noise conditions in effect, performing noise adaptive training [12], [25], [39]. Discriminative non-linear feature transformations can provide yet further gains in performance, because the feature transformation is optimized to reduce directly the error rates of the decoder [33].

Whereas the aforementioned conventional acoustic modeling techniques are mainly used within the Gaussian mixture model (GMM) framework, this paper also investigates their use within the commonly used hybrid DNN-HMM (hidden Markov model) approach [17]. The study includes all of the previously mentioned discriminative training and linear feature transformation techniques, but excludes f-bMMI [*1]. The experimental evaluation shows that these techniques still continue to effectively improve performance when used with a DNN. This DNN study is one of the primary new contributions of this paper when compared to our original challenge workshop publication [41].

In the ASR post-processing step, we propose to use a rescoring technique based on a simple combination of discriminative language modeling (DLM) [9], [27], [34] and minimum Bayes risk (MBR) decoding [5], [14], [24], [45]. In contrast with [24], which performs DLM with the MBR criterion, our work combines DLM and MBR *decoding* in a cascade form; we simply use the re-ranked 1-best obtained through DLM to initialize the MBR decoding. As a final step, system combination e.g., recognizer output voting error reduction (ROVER) [11] and its variants [10], [18], [42] can be used to obtain refined hypotheses by majority voting of the hypotheses of different systems; this results in higher performance than each base system can achieve individually. In order to create systems with complementary hypotheses, this work constructs two systems based on Mel-Frequency Cepstral Coefficient (MFCC) features as well as Perceptual Linear Prediction (PLP) features.

In summary, the goal of this paper is to evaluate the effectiveness of various state-of-the-art and novel techniques for ASR in reverberant and noisy environments by using the second CHiME challenge medium vocabulary task. Although the primary novelty of our approach lies in the combination of multiple components, additional novelties exist in the techniques themselves that are presented here. In particular, the techniques providing additional novel approaches are the prior-based BM (Section 3), and the combination of DLM and MBR (Section 4.6). The primary novel contributions of this work when compared to our previous work [41] are the introduction of a DNN system with speaker adaptation and the ROVER system combination.

2. System Overview

Figure 1 is a schematic diagram of the proposed system, which consists of three components. First is the noise suppression step, which is a prior-based BM that suppresses directional interferences (Section 3). Second is the feature transformation step, including feature-level transformations (LDA and MLLT with/without fMLLR, which are conventional and thus not explained in detail here) as well as discriminative feature transformations (feature-space techniques, presented in Section 4.3) [41]. Third is the ASR decoding step; it uses an acoustic model (GMM/DNN) with sequence discriminative training (Sections 4.1 and 4.2). Decoding results are re-ranked using DLM (Section 4.4), and MBR is performed based on the DLM output (Section 4.5). The best results were obtained by the ROVER combination of the hypotheses of two DNN systems using different features (MFCC and PLP).

3. Prior-based Binary Masking (BM)

In the CHiME challenge, two-channel recordings are provided and the target speaker is in a fixed frontal position with respect to the microphones [*2]. Binary masking based on the TDOA has

[*1] This is a reasonable exclusion because the lower layers of the deep neural networks (DNNs) already serve as an effective non-linear feature transformation.

[*2] This is a reasonable setting suitable for many applications, in which the users are either in a frontal position (such as when using home appliances), or in a fixed position (such as when using car navigation systems). In a situation where speakers are able to move freely, our prior-based BM approach could be modified to allow for multiple priors according to the speaker direction; this direction could be estimated by another method such as the cross-spectrum phase method [23].

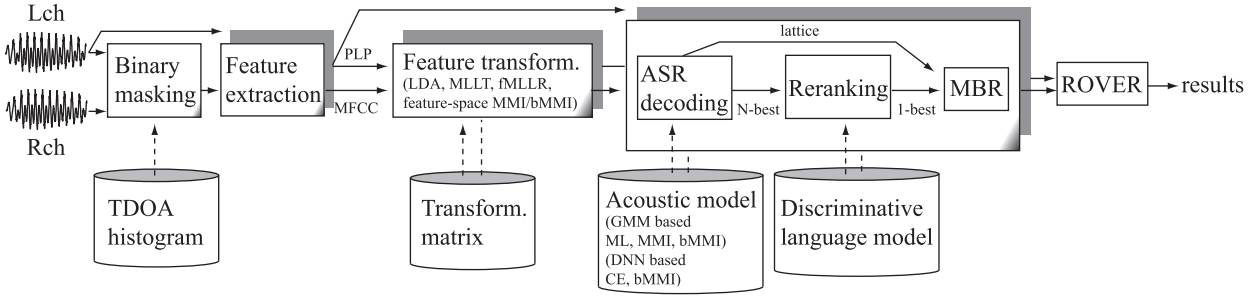


Fig. 1 Schematic diagram of the proposed system.

been shown to be more effective when used for ASR with a small number of microphones than simple delay-and-sum beamforming [37]. Consequently, we investigate the usage of this BM technique in our system.

When the receiver is in a frontal position and there is little reverberation and noise, the TDOA for signals coming from the target speaker should be close to zero. Hence, time-frequency bins for which the inter-microphone phase difference is not close to zero are unlikely to contain energy from the target speaker. However, in the presence of reverberation, the phase differences of the sound waves from a frontal target source may be non-zero. **Figure 2** shows the phase difference histograms at 250 Hz and 1 kHz in the "reverberated" (i.e., no noise) speech of the CHiME challenge training set. At 250 Hz, the histogram is almost symmetrical and the variance is small; at 1 kHz, however, the mean has drifted and the variance is large. The extent to which the phase difference is affected by reverberation and noise varies significantly for each frequency bin. Thus, a simple binary mask using only physical information will not be effective; indeed, preliminary experiments showed that this type of binary mask led to a slight improvement of word error rate (WER). As in Ref. [7], a statistical model is needed. In order to account for the offset of the phase difference when compared to the anechoic case as well as its variance, a prior-based BM is proposed. The phase difference $\theta_{t,\omega}$ at time frame t and frequency bin ω is calculated for each time-frequency bin as $\theta_{t,\omega} = \angle\left(X^L_{t,\omega}/X^R_{t,\omega}\right) \in (-\pi, \pi]$, where $X^L_{t,\omega}$ and $X^R_{t,\omega}$ are the complex short-time Fourier spectra for the left and right channels, respectively, and $\angle$ denotes the argument operator of a complex number.

In classical BM, a time-varying masking vector $\boldsymbol{W}_t = [W_{t,1}, \ldots, W_{t,\omega}, \ldots, W_{t,\Omega}]^\top \in \mathbb{R}^\Omega$ (where $\top$ denotes transposition) is designed using the following thresholding function:

$$W_{t,\omega} = \begin{cases} \epsilon & \text{if } |\theta_{t,\omega}| > \theta_c, \\ 1 & \text{if } |\theta_{t,\omega}| \le \theta_c, \end{cases} \tag{1}$$

where ϵ is a very small constant for spectral smoothing, and θ_c is a threshold determined in advance. Noise suppressed spectra $\boldsymbol{Y}_t \in \mathbb{C}^\Omega$ are obtained as $\boldsymbol{Y}_t = \boldsymbol{W}_t \odot \left(\boldsymbol{X}^L_t + \boldsymbol{X}^R_t\right)/2$, where $\boldsymbol{X}^L_t, \boldsymbol{X}^R_t \in \mathbb{C}^\Omega$, and $\odot$ denotes the element-wise multiplication of two vectors.

In our prior-based BM approach, a time-varying masking vector $\boldsymbol{W}'_t$ is determined using a frequency-dependent prior probability $q_\omega(\theta)$ of the phase difference θ. This prior probability is obtained from a phase difference histogram computed on the training data, renormalized to sum to unity. Denoting the peak of the

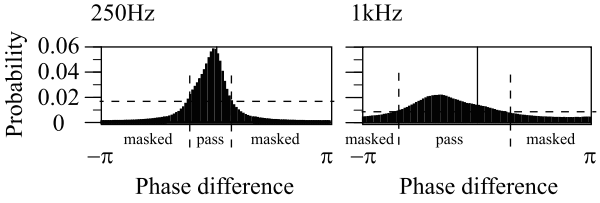


Fig. 2 Histogram of phase differences for two frequency bins.

histogram for frequency ω as $\bar{q}_\omega = \max_\theta q_\omega(\theta)$, we define the masking vector as

$$W'_{t,\omega} = \begin{cases} \epsilon & \text{if } q_\omega(\theta_{t,\omega})/\bar{q}_\omega < q_c, \\ \left(q_\omega(\theta_{t,\omega})/\bar{q}_\omega\right)^\alpha & \text{otherwise,} \end{cases} \tag{2}$$

where q_c is a threshold that determines the relative height with respect to the peak above which a time-frequency bin is passed. α is a warping parameter that can set the behavior of the mask from soft to binary. Both q_c and α are tuned manually in the development set. Whereas in classical BM, thresholding is based on a constant tolerance angle between the reference and the observation, our thresholding function takes the shape of the histogram into account. For histograms with a pronounced peak, such as the one corresponding to the 250 Hz frequency bin in Fig. 2, the tolerance angle is small, and only time-frequency bins for which the phase difference is very close to the peak are passed by the mask. On the other hand, the tolerance angle is large for flatter histograms such as the one corresponding to the 1 kHz bin in Fig. 2; in the latter case, phase differences farther from the peak are passed as well.

4. Discriminative Training Methods for Acoustic Modeling and Feature Transformation

4.1 MMI Discriminative Training of Acoustic Models

The goal of discriminative training algorithms is to obtain models that minimize the empirical risk computed from the correct labels and recognition hypotheses. Several training criteria have been introduced [19], [38], such as MMI [2], minimum classification error [26], or minimum phone error (MPE) [29]. We focus on MMI in this work, because MMI is the most widely used criterion and because it is the starting point for the more advanced bMMI, which we use below.

The goal of MMI training is to maximize the mutual information between correct labels and recognition hypotheses, based on the following objective function:

$$\mathcal{F}_{\mathrm{MMI}}(\lambda) = \sum_r \log \frac{p_\lambda\left(\boldsymbol{x}^{(r)}|\mathcal{H}_{s^{(r)}}\right)^\kappa p_L(s^{(r)})}{\sum_s p_\lambda\left(\boldsymbol{x}^{(r)}|\mathcal{H}_s\right)^\kappa p_L(s)}, \tag{3}$$

where $\boldsymbol{x}^{(r)} = (\boldsymbol{x}_1^{(r)}, \ldots, \boldsymbol{x}_t^{(r)}, \ldots, \boldsymbol{x}_{T_r}^{(r)})$ is the r-th utterance's feature sequence of length T_r; λ denotes the GMM-based acoustic model parameters composed of mixture weights, mean vectors, and (diagonal) covariance matrices; these parameters are optimized using the extended Baum-Welch algorithm; $\mathcal{H}_{s^{(r)}}$ and $\mathcal{H}_s$ are the HMM sequences that represent the correct label $s^{(r)}$ and a recognition result s, respectively; p_λ is the acoustic model likelihood, κ is the acoustic scale, and p_L is the language model likelihood.

While MMI is effective, performance can be further improved by giving more weight to the training data that is improperly recognized, as proposed in the bMMI framework [31]. The above objective function is extended to a boosted version as follows:

$$\mathcal{F}_{\mathrm{bMMI}}(\lambda) = \sum_r \log \frac{p_\lambda\left(\boldsymbol{x}^{(r)}|\mathcal{H}_{s^{(r)}}\right)^\kappa p_L(s^{(r)})}{\sum_s p_\lambda\left(\boldsymbol{x}^{(r)}|\mathcal{H}_s\right)^\kappa p_L(s)\mathrm{e}^{-bA(s,s^{(r)})}}, \tag{4}$$

where $A(s, s^{(r)})$ is the phoneme accuracy of hypothesis s for a reference $s^{(r)}$, and $b \geq 0$ is a boosting factor that controls the phoneme accuracy dependent weight. In this paper, we study the performances of bMMI for noisy speech ASR, comparing them to the performance of ML.

4.2 MMI Discriminative Training of Deep Neural Networks

GMM-HMM systems have constituted the mainstream architecture for decades, but DNN-HMM hybrid systems have outperformed them in recent years when used in clean speech conditions. In this paper, we investigate the effectiveness of DNN-HMM hybrid systems in noisy and reverberant speech conditions, and we show that these systems can bring further improvements compared to our challenge submission system [41]. In particular, we explore the benefits of sequence-level discriminative training methods for DNNs. DNNs are already discriminative at the frame level, because they are constructed based on discriminative criteria such as cross entropy (CE). Sequence-level discriminative training goes further in that it attempts to minimize the risk on the whole sequence instead of independently on each single frame; this type of training has been shown to improve performance over simple cross-entropy training [21], [43].

A DNN model with parameters θ outputs posterior probabilities $p_\theta(j|\boldsymbol{x}_t^{(r)})$ for each HMM state j at frame t. These probabilities are computed using a softmax layer applied to the top layer of the DNN:

$$p_\theta(j|\boldsymbol{x}_t^{(r)}) = \frac{\exp a_\theta(j|\boldsymbol{x}_t^{(r)})}{\sum_{j'} \exp a_\theta(j'|\boldsymbol{x}_t^{(r)})}, \tag{5}$$

where a_θ is the output of the top layer. Each layer of the DNN transforms the outputs of the previous layer through an affine transform, whose parameters are a subset of θ, followed by a nonlinear operation such as a sigmoid.

In order to use the classical HMM-based decoding framework, hybrid DNN-HMM systems replace the acoustic likelihood of GMMs by a pseudo-likelihood $p_\theta\left(\boldsymbol{x}_t^{(r)}|j\right)$ obtained as

$$p_\theta\left(\boldsymbol{x}_t^{(r)}|j\right) \propto p_\theta\left(j|\boldsymbol{x}_t^{(r)}\right)/p_0(j), \tag{6}$$

where $p_0(j)$ is the prior probability calculated from the count of states in the training data.

The values of the parameters θ are trained discriminatively according to the MMI criterion. The (boosted) MMI objective function is similar to that shown in Eqs. (3) and (4); the only difference is that the GMM likelihoods $p_\lambda\left(\boldsymbol{x}^{(r)}|\mathcal{H}_s\right)$ are replaced for the whole sequence by the equivalent DNN pseudo-likelihoods $p_\theta\left(\boldsymbol{x}^{(r)}|\mathcal{H}_s\right)$:

$$\mathcal{F}_{\mathrm{bMMI}}(\theta) = \sum_r \log \frac{p_\theta\left(\boldsymbol{x}^{(r)}|\mathcal{H}_{s^{(r)}}\right)^\kappa p_L(s^{(r)})}{\sum_s p_\theta\left(\boldsymbol{x}^{(r)}|\mathcal{H}_s\right)^\kappa p_L(s)\mathrm{e}^{-bA(s,s^{(r)})}}. \tag{7}$$

The gradient of the objective function with respect to the top layer output a_θ can be obtained by the chain rule as:

$$\begin{aligned}\frac{\partial \mathcal{F}_{\mathrm{bMMI}}(\theta)}{\partial a_\theta(j)} &= \sum_{j'} \frac{\partial \mathcal{F}_{\mathrm{bMMI}}}{\partial \log p_\theta\left(\boldsymbol{x}^{(r)}|j'\right)} \frac{\partial \log p_\theta\left(\boldsymbol{x}^{(r)}|j'\right)}{\partial a_\theta(j)}, \\ &= \kappa(\gamma_{j,t}^{\mathrm{num}} - \gamma_{j,t}^{\mathrm{den}}),\end{aligned} \tag{8}$$

where $\gamma_{j,t}^{\mathrm{num}}$ and $\gamma_{j,t}^{\mathrm{den}}$ are the posteriors of state j at frame t in the numerator and denominator of (7). The efficient calculation of these quantities is a classical step of MMI and MPE derivations for GMM systems and is described in detail in Refs. [29], [43]. All of the DNN parameters are estimated using the back-propagation procedure that begins with Eq. (8).

4.3 Feature-space MMI Discriminative Training

In addition to the acoustic model, sequence discriminative training can also be used to derive a feature transformation. This is referred to as feature-space discriminative training [28]. In this section, the I-dimensional vector $\boldsymbol{x}_t \in \mathbb{R}^I$ denotes the original static features without dynamic features (that is $\boldsymbol{x}_t$ does not include Δ and $\Delta\Delta$; this is unlike the previous sections). The transformed features $\boldsymbol{y}_t \in \mathbb{R}^I$ are obtained by adding $\boldsymbol{x}_t$ to an offset determined by applying a linear transformation $\boldsymbol{M}$ to a high-dimensional feature vector $\boldsymbol{h}_t \in \mathbb{R}^J$, where $\boldsymbol{M}$ is estimated using sequence discriminative training: $\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{M}\boldsymbol{h}_t$. The dimension J of $\boldsymbol{h}_t$ is assumed to be much larger than the dimension I of the original features $\boldsymbol{x}_t$ (i.e., $J \gg I$), and the role of the $I \times J$ matrix $\boldsymbol{M}$ is to project these rich high-dimensional features back down to the low-dimensional space containing the original features. The high-dimensional features $\boldsymbol{h}_t$ are obtained from $\boldsymbol{x}_t$ based on a universal background model (UBM) represented by a GMM, which we now describe in more detail. We denote the concatenation of $\boldsymbol{x}_t$ with its Δ and $\Delta\Delta$ features, $\boldsymbol{x}_t^* \in \mathbb{R}^{3I}$, as $\boldsymbol{x}_t^* = [\boldsymbol{x}_t{}^\top, \Delta\boldsymbol{x}_t{}^\top, \Delta\Delta\boldsymbol{x}_t{}^\top]^\top$. A diagonal-covariance GMM for $\boldsymbol{x}_t^*$ is learned from the training data; the number of Gaussian components is denoted as N_g, and their mean and variance in dimension i are denoted as $\mu_{n,i}$ and $\sigma_{n,i}$, respectively. Using this GMM, the high-dimensional features, $\boldsymbol{h}_t = \left[\boldsymbol{h}_{t,1}^\top, \ldots, \boldsymbol{h}_{t,N_g}^\top\right]^\top$, are computed from $\boldsymbol{x}_t^*$ as follows:

$$\boldsymbol{h}_{t,n} = p_{\mathrm{G}}(n|\boldsymbol{x}_t^*)\left[\frac{x_{t,1}^* - \mu_{n,1}}{\sigma_{n,1}}, \ldots, \frac{x_{t,3I}^* - \mu_{n,3I}}{\sigma_{n,3I}}, \xi\right]^\top, \tag{9}$$

where $p_{\mathrm{G}}(n|\boldsymbol{x}_t^*)$ is the posterior probability of the mixture component n at frame t, and ξ is a scaling factor for the bias term. Each

sub-vector $\boldsymbol{h}_{t,n} \in \mathbb{R}^{3I+1}$ is a normalized and reweighted version of the feature vector based on the parameters and posterior of the n-th component. Although the number of total dimensions of feature $\boldsymbol{h}_t$ becomes very large in this setup, $\boldsymbol{h}_t$ is sparsified by setting to zero all but a given number of sub-vectors corresponding to the Gaussian components with the highest posterior probabilities $p_{\mathrm{G}}(n|\boldsymbol{x}_t^*)$.

The objective function with respect to the matrix $\boldsymbol{M}$ is obtained similarly to the previous sections by replacing $\boldsymbol{x}$ in the bMMI objective functions (Eq. (4)) with the transformed feature $\boldsymbol{y}$, as follows:

$$\mathcal{F}_{\text{f-bMMI}}(\boldsymbol{M}) = \sum_r \log \frac{p_\lambda\left(\boldsymbol{y}^{(r)}|\mathcal{H}_{s^{(r)}}\right)^\kappa p_L(s^{(r)})}{\sum_s p_\lambda\left(\boldsymbol{y}^{(r)}|\mathcal{H}_s\right)^\kappa p_L(s)\mathrm{e}^{-bA(s,s^{(r)})}}. \quad (10)$$

In our GMM systems, f-bMMI training with respect to $\boldsymbol{M}$ and b-MMI training with respect to the GMM parameter λ are iteratively performed to optimize both parameters [*3].

4.4 Discriminative Language Modeling

DLM [9], [27], [34] learns patterns of errors in the N-best hypotheses output by a speech recognizer, and adjusts the hypotheses' scores so that the one with the least errors is selected. The score can be modified simply using the inner product of a feature vector $\boldsymbol{\phi}(s)$ extracted from a hypothesis s and a weight vector $\boldsymbol{w}$. The re-scored best hypothesis $\hat{s}^{(r)}$ is then obtained as:

$$\hat{s}^{(r)} = \arg\max_{s \in \mathcal{S}^{(r)}} \left[w^{(0)} \cdot p_\lambda\left(\boldsymbol{y}^{(r)}|\mathcal{H}_s\right)^\kappa p_L(s) + \boldsymbol{w}^\top \phi(s) \right], \quad (11)$$

where $w^{(0)}$ is the weight for the original acoustic and language model score, and $\mathcal{S}^{(r)}$ is the set of N-best hypotheses for utterance r. Features are usually N-gram counts. During training, separate weight vectors $\boldsymbol{w}^{(r)}$ for each speech utterance r are estimated by using an on-line training algorithm, which employs the following rule:

$$\boldsymbol{w}^{(r)} \leftarrow \boldsymbol{w}^{(r-1)} + (\boldsymbol{\phi}(s^{(r)}) - \boldsymbol{\phi}(\hat{s}^{(r)})). \quad (12)$$

To increase the generalization ability, the weight vector used at test time is obtained by averaging the weight vectors for all training utterances [6]. In our paper, instead of using the reference as $s^{(r)}$ in Eq. (12), we select $s^{(r)}$ within the N-best list as the hypothesis with lowest WER with respect to the reference.

4.5 Minimum Bayes Risk Decoding

MBR decoding is another re-scoring technique that attempts to approximately minimize the Bayes risk obtained from the WER [5], [14], [45]. The algorithm modifies the 1-best word sequence $s_1^{(r)}$ by word-by-word replacements to obtain a modified word sequence $\tilde{s}^{(r)}$ that minimizes the expected edit distance $L(\tilde{s}^{(r)}, s')$ to other word sequences s' in the hypothesis lattice [*4] $\mathcal{L}^{(r)}$. The edit distance L is approximately computed based on the forward-backward algorithm [45] and this procedure repeats until no symbols are replaced.

[*3] Note that f-bMMI training is undertaken only for the GMM-based acoustic models, because the DNN acoustic models in Section 4.2 already include (non-linear) discriminative feature transformations in their deep networks.

[*4] N-best lists can be used instead of lattices.

4.6 Combination of Minimum Bayes Risk Decoding with Discriminative Language Modeling

In the previous section, conventional MBR decoding starts from the 1-best word sequence of the lattice and then by forming alignments of the rest of hypotheses. The iteration above can reach local minimum, similar to the ML training in acoustic modeling. Our approach improves the initial point by replacing the conventional 1-best word sequence $s_1^{(r)}$ with the 1-best word sequence $\hat{s}$ in an N-best list re-scored by DLM [*5] to efficiently combine minimum Bayes risk decoding with DLM-based N-best re-scoring.

4.7 System Combination

A combination of multiple systems, even if some of the systems have significantly lower performance, may outperform the best single system, in particular when the systems tend to display different patterns in their errors. Many system combination methods, such as Refs. [8], [10], [11], [18], [32], have been proposed. Here, we use ROVER [11], which is the simplest approach, because system combination is a complementary component of this paper. ROVER combines the 1-best results outputs of multiple systems which mainly differ by their input features, MFCC and PLP.

5. Experimental Setup

5.1 Task Description

We validated the effectiveness of our proposed approach for reverberated and noisy speech on track 2 of the second CHiME challenge [44], which is a medium-vocabulary task whose speech utterances are taken from the *Wall Street Journal* database (WSJ0). **Table 1** presents detailed information about the training (**si_tr_s**), development (**si_dt_05**), and evaluation (**si_et_05**) datasets. **Table 2** shows the settings for the ASR systems.

Acoustic models were trained using the **si_tr_s** and some of the parameters (e.g., language model weights) were tuned using the

Table 1 Number of utterances and speakers in each dataset. Development and evaluation datasets were provided for each SNR.

dataset	# utterances	# speakers
Training dataset (**si_tr_s**)	7,138	83
Development dataset (**si_dt_05**)	409	10
Evaluation dataset (**si_et_05**)	330	12

Table 2 Setup for the ASR systems.

Sampling frequency	16 kHz
Window length	25 ms
Window shift	10 ms
Feature 1	0th~12th MFCCs/PLPs + Δ + ΔΔ
Feature 2	(0th~12th MFCCs/PLPs × 9 frames) + LDA+MLLT (→40 dim.)
Feature 3	0th~22th filter banks (FBANK) + Δ + ΔΔ
HMM state	2,500 shared triphone states
Number of Gaussians	15,000
Hidden layer of DNN	3
Vocabulary size	5,000

[*5] The accurate assignment probability can be obtained by converting the estimated DLM weights to arc weights in a lattice. However, the conversion is not trivial since DLM would include unseen n-gram features or wide-span features, and the corresponding DLM weights cannot be converted to those of lattice arcs, in a straightforward manner.

WERs on the **si_dt_05**. This database simulates realistic environments. There are two types of data, "reverberated" and "isolated." The "reverberated" data were created by convolving clean speech with binaural room impulse responses corresponding to a frontal position at a distance of 2 m from the stereo microphones in a family living room. The "isolated" data were created by adding real-world noises recorded in the same room to the "reverberated" data, and then adding noise excerpts selected to obtain signal-to-noise ratio (SNR) ranges of −6, −3, 0, 3, 6, and 9 dB without rescaling. Added noise sources are typically non-stationary (e.g., other speakers' utterances, home noises, or music). We used Kaldi toolkit [30] for the experiments.

5.2 Feature Extraction and Transformation

We now describe the settings of the feature extraction and the feature transformation. The baseline acoustic features were MFCCs. In addition to these, PLP features were used for the final system combination, as described in Section 4.7. In this paper, the LDA classes are taken as the tri-phone HMM states. We concatenate 13-order static MFCCs in nine contiguous frames to consider the influence of long context, instead of using conventional delta features. This results in a total of 117-dimensional features, which are compressed into 40 dimensions. We use diagonal-covariance models, together with MLLT feature space transformation to decrease correlations between features.

For DNN, mel filter bank (FBANK) features tend to lead to better performance than MFCC features. We validate the effectiveness of FBANK features in addition to MFCC features and MFCC + LDA+MLLT features. For further noise robustness, we also investigate the use of SAT and global fMLLR.

In discriminative feature transformation (Section 4.3), the UBM is constructed using $N_g(= 400)$ Gaussians. Offset features are calculated for each of $K_g(= 39)$-dimensional MFCC features including Δ and $\Delta\Delta$s, and the posterior probabilities are expanded using nine contiguous frames. The total dimension of the feature vector $\boldsymbol{h}_t$ is 144k (400 [Gaussians] × (39 + 1) [dimensions/Gaussian/frame] × 9 [frames]). Features with the top two posteriors are selected and all other features are set to zero.

5.3 Acoustic Models

We summarize the experimental procedure based on the above setup as follows: First, a clean acoustic model was trained. The number of mono-phones was 40, including silence ("sil"). Second, reverberated acoustic models were trained using the "reverberated" dataset. Third, noisy acoustic models were trained multi-conditionally using the "isolated" dataset without noise suppression. Finally, from this ML model, the effectiveness of the discriminative training and feature transformation for the "isolated" dataset was validated. The parameters used in our experiments were set to be those described in the WSJ tutorial attached to the Kaldi toolkit.

For the DNN, we used the nnet2 of neural network training implemented in the Kaldi toolkit with three hidden layers whose activation functions were sigmoid. Stacking hidden layers layer by layer, the DNN was constructed instead of using the restricted Boltzmann machine. The learning rate η was decreased from the initial learning rate η_0 (0.01) to the final learning rate η_e (0.001) at the end of training as $\eta = \eta_0 \exp(i \log(\eta_e/\eta_0)/i_{\max})$ where i is an iteration number. The number of iterations $i_{\max}$ was 43 and the minibatch size was 128. Nine concatenated frames were input and the number of hidden layer nodes was 309.

5.4 Discriminative Language Modeling

Weights $\boldsymbol{w}$ in Eq. (11) of a DLM were learned on the training data set using 100-best recognition candidates, where the weight w_0 associated with the original score was set to 20. Using these weights, results were re-ranked, with w_0 set to 13. Weights were obtained by averaged perceptron at three iterations. Features were counts of uni-, bi-, and tri-grams.

5.5 System Combination

System combination techniques are effective for the case in which the hypotheses of the respective systems are different but the performance of the systems is similar. The most promising approach is to use additional features; thus, after generation of the best hypotheses of the DNN-HMM system for MFCC and PLP feature with regard to the time alignment and the confidence measure, these hypotheses were combined using ROVER.

6. Results and Discussion

6.1 Discriminative Training

With regard to the MFCC features, discriminative training improved the WER from the ML baseline as shown in **Table 3** (upper) [*6]. The mixture of speech and noise increases the likelihood of detecting erroneous phonemes and leads to incorrect recognition especially when the noise source is other people's utterances. These errors could be modified by discriminative training. The boosting factor in Eqs. (4), (7), and (10), b, was set to 0.1 because the preliminary experiments show that the performance did not heavily depend on the boosting factors and that the optimized values of the boosting factor were approximately 0.1–0.2. The denominator lattices for discriminative training were generated using the ML model. The boosted MMI improved the WER by

Table 3 WER[%] of GMM-HMM for **si_dt_05** without noise suppression. MFCC features (upper), MFCC + LDA+MLLT (middle), MFCC + LDA+MLLT + SAT+fMLLR (lower).

∘MFCC + Δ + ΔΔ							
	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
ML	74.20	66.57	58.24	51.84	46.73	40.64	56.37
bMMI	72.78	64.71	55.69	50.83	44.00	40.27	54.71
f-bMMI	68.64	61.56	53.11	47.65	41.73	36.98	51.61
∘MFCC + LDA+MLLT							
ML	70.95	62.62	53.98	47.37	40.27	34.84	51.67
f-bMMI	66.65	57.46	48.25	42.99	35.71	31.07	47.02
∘MFCC + LDA+MLLT + SAT+fMLLR							
ML	68.36	58.30	48.80	40.73	35.09	28.54	46.64
f-bMMI	**62.43**	**52.23**	**42.17**	**35.31**	**29.84**	**24.72**	**41.12**

[*6] The MMI and f-MMI results were omitted, because the performance of those was lower than those of the bMMI and f-bMMI and recently, the results of GMM were less meaningful than at the time of the second CHiME challenge. The detailed evaluations are found in Ref. [41].

Table 4 WER[%] of GMM-HMM for **si_dt_05** with noise suppression by conventional binary masking (BM) and the proposed prior-based BM. MFCC features were used.

○MFCC + Δ + ΔΔ							
	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
conventional BM	73.98	66.90	57.93	52.35	46.38	40.54	56.35
prior based BM	66.82	57.87	48.86	42.29	38.18	31.86	47.65

1.6% absolute [*7] to the ML, whereas the feature-space discriminative training improved the WER by 3% further. We believe that the feature space was adapted for a target speaker to improve the WER and that this effect reduced the influence of other noises.

6.2 Feature Transformation

The MFCC features were transformed using LDA and MLLT. Table 3 (middle) shows the WER for this case, whereas LDA by itself (i.e., without MLLT) achieves 54.37% (ML). This shows that features that are highly discriminable from other phonemes can be obtained by LDA. The performance gains of LDA and MLLT were 2.0 and 2.7%, respectively. It is effective to use a long context to reduce the influence of non-stationary noises. Furthermore, although noises increase the correlations between MFCC coefficients in each dimension, MLLT reduced the correlations. The denominator lattices for discriminative training were re-generated using the ML (MFCC + LDA+MLLT) model. Discriminative training improved the WER by 4.6%.

6.3 Adaptation

Table 3 (lower) shows the WER when additional SAT and fMLLR were used. Because the amount of training data is very limited, transformation into a canonical space, which leads to an increase in the effective amount of training data, has a strong impact on the estimation accuracy of the acoustic models. Additionally, fMLLR adaptation for a target speaker reduced the influence of noises and improved the WER by 5.0%. The denominator lattices for discriminative training were also re-generated using this adapted ML model. Discriminative training improved the WER by 5.5%.

6.4 Noise Suppression

In order to clarify the effectiveness of the prior-based proposed BM, **Table 4** shows the WERs of the proposed BM compared with those of the conventional BM [37] by using baseline GMM with MFCC features. As mentioned in Section 3, the conventional BM improved the performance significantly, whereas the proposed BM improved the WER in all SNRs by 7% to 9%. The best warping parameters for the proposed BM α was 0.25. Directional noises were effectively suppressed by our proposed method, but diffused noises such as music remained.

Table 5 shows the WER with feature adaptation and discriminative training. Combination of them with noise suppression was effective. The employed adaptation improved the WER by 9.0% and discriminative training improved it by 5.6%.

6.5 Deep Neural Network

Tables 6 and **7** provide the WERs of a DNN. Table 6 shows the result without noise suppression and Table 7 shows that with noise suppression. Using the same MFCC features, at the ML and CE baseline, the DNN result outperformed the GMM results by 8.9% (without noise suppression) and 6.3% (without noise suppression), respectively.

Table 7 (the second division) shows that the FBANK features outperformed the MFCC features, as previous studies have shown. The performance of MFCC + LDA+MLLT was worse than that of the FBANK features for a DNN-HMM system. When combined with GMM-based speaker adaptation techniques (SAT+fMLLR), DNN slightly outperformed f-bMMI even without discriminative training when Table 7 is compared with Ta-

[*7] In this paper, WER improvements are shown in absolute values.

Table 5 WER[%] of GMM-HMM for **si_dt_05** with noise suppression by prior-based BM. MFCC features (upper) and MFCC + LDA+MLLT + SAT+fMLLR (lower).

○MFCC + Δ + ΔΔ							
	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
f-bMMI	63.40	54.05	44.28	38.87	33.72	29.90	44.04
○MFCC + LDA+MLLT + SAT+fMLLR							
ML	59.94	47.93	39.83	33.01	28.00	23.47	38.70
f-bMMI	52.93	42.62	34.59	27.63	24.27	20.24	33.71
(+DLM)	53.16	42.93	34.36	27.26	23.72	19.47	33.48
(+MBR)	52.65	42.04	33.75	27.05	23.74	19.91	33.19
(+DLM+MBR)	52.54	42.09	33.72	27.02	23.66	19.66	33.11

Table 6 WER[%] of DNN-HMM for **si_dt_05** without noise suppression. MFCC features (upper) and MFCC + LDA+MLLT (lower).

○MFCC + Δ + ΔΔ							
	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
CE	67.47	57.55	48.78	43.43	36.10	31.76	47.52
○MFCC + LDA+MLLT							
CE	64.39	53.67	44.28	38.56	32.70	28.09	43.62

Table 7 WER[%] of DNN-HMM for **si_dt_05** with noise suppression by prior-based BM. MFCC features (first), FBANK features (second), MFCC + LDA+MLLT (third), MFCC + LDA+MLLT + SAT+fMLLR (fourth) and PLP + LDA+MLLT + SAT+fMLLR (fifth). Hypotheses of two systems (*1 and *2) were combined by ROVER (last).

○MFCC + Δ + ΔΔ							
	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
CE	62.44	51.59	42.93	35.27	30.11	25.76	41.35
○FBANK + Δ + ΔΔ							
CE	55.60	44.52	36.16	30.62	26.02	22.32	35.87
bMMI	51.70	39.43	31.75	26.83	23.35	19.86	32.15
○MFCC + LDA+MLLT							
CE	57.21	45.85	36.21	30.61	26.36	23.31	36.59
○MFCC + LDA+MLLT + SAT+fMLLR							
CE	52.78	42.50	34.08	27.05	24.13	20.12	33.44
bMMI	47.34	36.33	28.96	23.40	20.03	17.05	28.85
(+DLM)	47.37	36.48	28.94	23.09	20.02	16.93	28.80
(+MBR)	46.79	35.68	28.44	22.88	19.91	16.64	28.39
*1 (+DLM+MBR)	46.67	35.55	28.38	22.84	19.83	16.65	28.32
○PLP + LDA+MLLT + SAT+fMLLR							
*2 (+DLM+MBR)	47.38	35.29	27.89	22.70	19.38	15.92	28.09
○ROVER							
*1+*2	**45.12**	**34.34**	**26.73**	**21.71**	**19.09**	**15.39**	**27.06**

Table 8 WER[%] of GMM-HMM for **si_et_05** without noise suppression.

○MFCC + Δ + ΔΔ							
	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
-GMM-HMM							
ML	69.79	62.71	55.86	46.89	42.07	37.49	52.47
-DNN-HMM							
CE	62.79	53.19	46.46	38.26	32.30	30.34	43.89

○MFCC + LDA+MLLT + SAT+fMLLR							
-GMM-HMM							
ML	60.83	52.14	43.51	34.28	29.22	23.82	40.63
f-bMMI	**54.70**	**45.11**	**35.98**	**28.64**	**24.38**	**21.39**	**35.04**

ble 5 [*8]. With discriminative training (bMMI) for DNN, the DNN outperformed f-bMMI of GMM by 4.9%. This shows the effectiveness of DNN for noise-robust ASR. The performance gains by discriminative training of acoustic models were around 5% for both GMM and DNN.

6.6 Discriminative Language Modeling and Minimum Bayes Risk Decoding

Tables 5 (lower) and 7 (the fourth division) show that DLM improved the average WER by 0.2% and 0.05%, respectively, especially for the 9 dB case of GMM, which resulted in a 0.8% improvement. DLM was not always effective because, while error tendencies were dependent on a particular SNR, training was performed on the whole multi-condition training set, which included all SNRs. This led to a mismatch between training and recognition, thereby degrading performance. DLM was less effective for DNN than GMM.

Tables 5 (lower) and 7 (the fourth division) show that MBR improved the WER by 0.5% for both GMM and DNN. The performance of MBR was stable with respect to SNR. The combination of DLM and MBR as mentioned in Section 4.5 improved the WER further by 0.1% for both cases because DLM refined the initial 1-best result and adapts to error tendencies inherent to the decoder. Thus, MBR was effective for both GMM and DNN.

6.7 System Combination

Table 7 (the fifth division) shows the WER using PLP features for the best case of DNN. This (PLP) result was equivalent to the condition of 1) of the fourth division. PLP was slightly better than MFCC but preliminary experiments show that simple concatenation of MFCC and PLP features for DNN degraded the performance. Table 7 (the last division) shows that ROVER, which combined the 1-best hypotheses of MFCC and PLP, improved the WER by 1% and this was effective in all SNR cases.

6.8 Evaluation Set

Table 8 shows the WERs on the evaluation set using the models tuned on the development set. Tendencies were the same to those of the development set. DNN was still effective for the evaluation set. Using both discriminative training and feature transformation (f-bMMI) achieved a 33.2% error reduction relative to the baseline (ML). Thus, we show the effectiveness of both discriminative training and feature transformation for reverberated and noisy speech.

[*8] This type of adaptation cannot be directly applied for the FBANK feature due to their high dimensionality and correlation across feature dimensions [35].

Table 9 WER[%] of GMM- and DNN-HMM for **si_et_05** with noise suppression.

○MFCC+ Δ + ΔΔ							
	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
-GMM-HMM							
ML	60.58	52.87	45.60	37.70	33.38	29.24	43.23

○MFCC + LDA+MLLT + SAT+fMLLR							
-GMM-HMM							
ML	50.91	41.64	33.89	26.30	21.61	18.85	32.20
f-bMMI	44.54	35.91	29.24	22.31	17.77	15.88	27.61
(+DLM)	44.27	35.48	28.75	21.61	17.34	15.37	27.14
(+MBR)	44.51	35.42	28.81	21.46	17.41	14.98	27.10
(+DLM+MBR)	44.12	35.46	28.12	21.20	17.43	14.83	26.86
-DNN-HMM							
bMMI	37.98	28.26	21.86	17.71	12.61	11.75	21.70
(+DLM)	38.00	27.82	21.80	16.64	12.22	11.62	21.35
(+MBR)	37.14	27.35	21.41	16.94	12.55	11.54	21.16
*1 (+DLM+MBR)	37.16	27.44	21.24	16.66	12.40	11.49	21.07

○PLP + LDA+MLLT + SAT+fMLLR							
-DNN-HMM							
*2 (+DLM+MBR)	38.22	27.93	22.57	16.91	13.49	12.14	21.88

○ROVER							
*1+*2	**36.43**	**26.02**	**20.96**	**15.84**	**11.99**	**11.17**	**20.40**

Table 9 shows the WERs after noise suppression. Using a GMM-HMM system with both discriminative training and feature transformation (f-bMMI) achieved a 37.9% error reduction relative to the baseline (ML). These results were submitted to the CHiME challenge workshop [41]. Moreover, for this case, DNN with bMMI and system combination of two systems (ROVER) achieved a 52.6% error reduction, which means that errors were reduced by more than half.

7. Conclusions

We developed a state-of-the-art recognition system for the second CHiME challenge track 2, which is a medium-size automatic speech recognition task under noisy environments, and validated the effectiveness of both feature transformation and discriminative methods. For realistic reverberated and noisy environments of this task, we proposed a prior-based binary masking and show its effectiveness. Combination of minimum Bayes risk decoding and discriminative language modeling improved the word error rate by considering error tendencies, which are inherent to the decoder. Deep neural networks are also effective; they outperformed the feature-space boosted maximum mutual information technique, which had been the state-of-the-art acoustic modeling technique for conventional Gaussian mixture model based systems. This superior performance was achieved even without discriminative training; with the combination of sequential discriminative training and system combination, the best performance was achieved. Experiments show that these techniques are effective for non-stationary interference and reverberation.

Future work will be an extension of our approaches to various tasks. For handling distant speech, reverberation effect is also important [22]. In this scenario, because the speaker moves freely,

our prior-based binary masking approach needs modifications to include multiple priors according to the speaker direction.

References

[1] Anastasakos, T., McDonough, J., Schwartz, R. and Makhoul, J.: A Compact Model for Speaker-adaptive Training, *Proc. ICSLP*, pp.1137–1140 (1996).

[2] Bahl, L., Brown, P., de Souza, P. and Mercer, R.: Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition, *Proc. ICASSP*, Vol.11, pp.49–52 (1986).

[3] Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C., Morgan, N. and O'Shaughnessy, D.: Research Developments and Directions in Speech Recognition and Understanding Part 1, *IEEE Signal Processing Magazine*, Vol.26, pp.75–80 (2009).

[4] Barker, J., Vincent, E., Ma, N., Christensen, C. and Green, P.: The PASCAL CHiME Speech Separation and Recognition Challenge, *Computer Speech and Language*, Vol.27, No.3, pp.621–633 (2013).

[5] Byrne, W.: Minimum Bayes Risk Estimation and Decoding in Large Vocabulary Continuous Speech Recognition, *IEICE Trans. Inf. Syst.*, Vol.E89-D, pp.900–907 (2006).

[6] Collins, M.: Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms, *Proc. ACL Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp.1–8 (2002).

[7] Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., Kubo, Y., Souden, M., Hahm, S.-J. and Nakamura, A.: Speech Recognition in Living Rooms: Integrated Speech Enhancement and Recognition System Based on Spatial, Spectral and Temporal Modeling of Sounds, *Computer Speech and Language*, Vol.27, pp.851–873 (2013).

[8] Deoras, A., Filimonov, D., Harper, M. and Jelinek, F.: Model Combination for Speech Recognition Using Empirical Bayes Risk Minimization, *Proc. Spoken Language Technology Workshop* (*SLT*), pp.235–240, IEEE (2010).

[9] Dikici, E., Semarci, M., Saraąlar, M. and Alpaydin, E.: Classification and Ranking Approaches to Discriminative Language Modeling for ASR, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.21, pp.291–300 (2013).

[10] Evermann, G. and Woodland, P.: Posterior Probability Decoding, Confidence Estimation and System Combination, *Proc. NIST Speech Transcription Workshop* (2000).

[11] Fiscus, J.: A Post-processing System to Yield Reduced Error Word Rates: Recognizer Output Voting Error Reduction (ROVER), *Proc. ASRU*, pp.347–354 (1997).

[12] Gales, M.: Maximum Likelihood Linear Transformations for HMM-based Speech Recognition, *Computer Speech and Language*, Vol.12, pp.75–98 (1998).

[13] Gales, M.: Semi-tied Covariance Matrices for Hidden Markov Models, *IEEE Trans. Speech and Audio Processing*, Vol.7, pp.272–281 (1999).

[14] Goel, V. and Byrne, W.: Minimum Bayes-risk Automatic Speech Recognition, *Computer Speech and Language*, Vol.14, pp.115–135 (2000).

[15] Gopinath, R.: Maximum Likelihood Modeling with Gaussian Distributions for Classification, *Proc. ICASSP*, pp.661–664 (1998).

[16] Haeb-Umbach, R. and Ney, H.: Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition, *Proc. ICASSP*, pp.13–16 (1992).

[17] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition, *IEEE Signal Processing Magazine*, Vol.28, pp.82–97 (2012).

[18] Hoffmeister, B., Klein, T., Schlüter, R. and Ney, H.: Frame Based System Combination and a Comparison with Weighted ROVER and CNC, *Proc. ICSLP*, pp.537–540 (2006).

[19] Hsiao, R.: *Generalized Discriminative Training for Speech Recognition*, PhD thesis for Carnegie Mellon University (2012).

[20] Johnson, D. and Dudgeon, D.: *Array Signal Processing*, Prentice-Hall, New Jersey (1993).

[21] Kingsbury, B., Sainath, T. and Soltau, H.: Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization, *Proc. INTERSPEECH*, pp.485–488 (2012).

[22] Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W., Maas, R., Gannot, S. and Raj, B.: The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech, *Proc. WASPAA*, New Paltz, NY, USA, pp.1–4 (2013).

[23] Knapp, C. and Carter, G.: The Generalized Correlation Method for Estimation of Time Delay, *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.24, No.4, pp.320–327 (1976).

[24] Kuo, H., Mangu, L., Arisoy, E. and Saon, G.: Minimum Bayes Risk Discriminative Language Models for Arabic Speech Recognition, *Proc. ASRU*, pp.208–213 (2011).

[25] Leggetter, C. and Woodland, P.: Flexible Speaker Adaptation for Large Vocabulary Speech Recognition, *Proc. EUROSPEECH*, pp.1155–1158 (1995).

[26] McDermott, E., Hazen, T., Le Roux, J., Nakamura, A. and Katagiri, S.: Discriminative Training for Large-Vocabulary Speech Recognition Using Minimum Classification Error, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.15, pp.203–223 (2007).

[27] Oba, T., Hori, T., Nakamura, A. and Ito, A.: Round-robin Duel Discriminative Language Models, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.20, pp.1244–1255 (2012).

[28] Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H. and Zweig, G.: fMPE: Discriminatively Trained Features for Speech Recognition, *Proc. ICASSP*, pp.961–964 (2005).

[29] Povey, D. and Woodland, P.: Minimum Phone Error and I-smoothing for Improved Discriminative Training, *Proc. ICASSP*, Vol.I, pp.105–108 (2002).

[30] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Petr, M., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G. and Veselý, K.: The Kaldi Speech Recognition Toolkit, *Proc. ASRU*, pp.1–4 (2011).

[31] Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G. and Visweswariah, K.: Boosted MMI for Model and Feature-space Discriminative Training, *Proc. ICASSP*, pp.4057–4060 (2008).

[32] Ren, B., Wang, L., Lu, L., Ueda, Y. and Kai, A.: Combination of bottleneck feature extraction and dereverberation for distant-talking speech recognition, *Multimedia Tools and Applications*, Vol.75, No.9, pp.5093–5108 (2016).

[33] Renals, S., Hain, T. and Bourlard, H.: Recognition and Understanding of Meetings the AMI and AMIDA Projects, *Proc. ASRU*, pp.238–247 (2007).

[34] Roark, B., Saraçlar, M., Collins, M. and Johnson, M.: Discriminative Language Modeling with Conditional Random Fields and the Perceptron Algorithm, *Proc. ACL*, pp.47–54 (2004).

[35] Sainath, T.N., Mohamed, A.-R., Kingsbury, B. and Ramabhadran, B.: Deep Convolutional Neural Networks for LVCSR, *Proc. ICASSP*, pp.8614–8618 (2013).

[36] Saon, G. and Chien, J.-T.: Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances, *IEEE Signal Processing Magazine*, Vol.29, No.6, pp.18–33 (2012).

[37] Sawada, H., Araki, S. and Makino, S.: Underdetermined Convolutive Blind Source Separation via Frequency Bin-wise Clustering and Permutation Alignment, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.19, pp.516–527 (2011).

[38] Schlüter, R., Macherey, W., Müller, B. and Ney, H.: Comparison of Discriminative Training Criteria and Optimization Methods for Speech Recognition, *Speech Communication*, Vol.34, pp.287–310 (2001).

[39] Shinoda, K. and Lee, C.: A Structural Bayes Approach to Speaker Adaptation, *IEEE Trans. Speech and Audio Processing*, Vol.9, pp.276–287 (2001).

[40] Tachioka, Y., Narita, T. and Iwasaki, T.: Direction of Arrival Estimation by Cross-power Spectrum Phase Analysis Using Prior Distributions and Voice Activity Detection Information, *Acoustical Science & Technology*, Vol.33, No.1, pp.68–71 (2012).

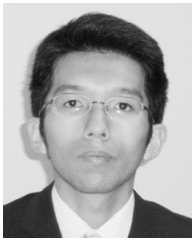
[41] Tachioka, Y., Watanabe, S., Le Roux, J. and Hershey, J.: Discriminative Methods for Noise Robust Speech Recognition: A CHiME Challenge Benchmark, *Proc. 2nd CHiME Workshop on Machine Listening in Multisource Environments*, pp.19–24 (2013).

[42] Tachioka, Y., Watanabe, S., Le Roux, J. and Hershey, J.: A Generalized Framework of Discriminative Training for System Combination, *Proc. ASRU*, pp.43–48, IEEE (2013).

[43] Veselý, K., Ghoshal, A., Burget, L. and Povey, D.: Sequence-discriminative Training of Deep Neural Networks, *Proc. INTERSPEECH*, pp.2345–2349 (2013).

[44] Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F. and Matassoni, M.: The Second 'CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines, *Proc. ICASSP*, pp.126–130 (2013).

[45] Xu, H., Povey, D., Mangu, L. and Zhu, J.: An Improved Consensus-like Method for Minimum Bayes Risk Decoding and Lattice Combination, *Proc. ICASSP*, pp.4938–4941 (2010).



Yuuki Tachioka is a researcher at the Information Technology R&D center, Mitsubishi Electric Corporation. His research interest is speech processing. He was graduated from the Faculty of Engineering of the University of Tokyo in 2006 and received M. Environmental studies from the same university in 2008.

Shinji Watanabe was a research scientist at NTT Communication Science Laboratories in Japan for 10 years, working on Bayesian learning for speech recognition, speaker adaptation, and language modeling, before joining MERL in 2012. His research interests include speech recognition, spoken language processing, and machine learning.

Jonathan Le Roux completed his B.Sc. and M.Sc. in Mathematics at the École Normale Superieure in Paris, France. Before joining MERL in 2011, he spent several years in Beijing and Tokyo. In Tokyo he worked as a postdoctoral researcher at NTT's Communication Science Laboratories. His research interests are in signal processing and machine learning applied to speech and audio.

John Hershey spent 5 years at IBM's T.J. Watson Research Center in New York, leading a Noise Robust Speech Recognition team, before joining MERL in 2010. He also spent a year as a visiting researcher in the speech group at Microsoft Research, after obtaining his Ph.D. from UCSD. He is currently working on machine learning for signal separation, speech recognition, language processing, and adaptive user interfaces.