Tachioka *et al. EURASIP Journal on Advances in Signal Processing* (2015) 2015:52
DOI 10.1186/s13634-015-0241-y

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Effectiveness of dereverberation, feature transformation, discriminative training methods, and system combination approach for various reverberant environments

Yuuki Tachioka[1*], Tomohiro Narita[1] and Shinji Watanabe[2]

Abstract

The recently released REverberant Voice Enhancement and Recognition Benchmark (REVERB) challenge includes a reverberant automatic speech recognition (ASR) task. This paper describes our proposed system based on multi-channel speech enhancement preprocessing and state-of-the-art ASR techniques. For preprocessing, we propose a single-channel dereverberation method with reverberation time estimation, which is combined with multichannel beamforming that enhances direct sound compared with the reflected sound. In addition, this paper also focuses on state-of-the-art ASR techniques such as discriminative training of acoustic models including the Gaussian mixture model, subspace Gaussian mixture model, and deep neural networks, as well as various feature transformation techniques. Although, for the REVERB challenge, it is necessary to handle various acoustic environments, a single ASR system tends to be overly tuned for a specific environment, which degrades the performance in the mismatch environments. To overcome this mismatch problem with a single ASR system, we use a system combination approach using multiple ASR systems with different features and different model types because a combination of various systems that have different error patterns is beneficial. In particular, we use our discriminative training technique for system combination that achieves better generalization by making systems complementary with the modified discriminative criteria. Experiments show the effectiveness of these approaches, reaching 6.76 and 18.60 % word error rates on the REVERB simulated and real test sets. These are 68.8 and 61.5 % relative improvements over the baseline.

Keywords: Reverberant speech recognition; Dereverberation; Discriminative training; Feature transformation; System combination; REVERB challenge

1 Introduction

Automatic speech recognition (ASR) using distant microphones can overcome application restrictions of places and devices and widen the usage of speech interfaces. For example, users can control distant home appliances by voice without touching the devices. However, in such a scenario, it is necessary to address reverberation, which is composed of reflected sounds from walls, ceilings, or furniture, in addition to the direct sound from a sound source. Reverberation as well as noise degrades the intelligibility of speech for humans, and it also significantly degrades ASR performance.

The REverberant Voice Enhancement and Recognition Benchmark (REVERB) challenge is an Audio and Acoustic Signal Processing (AASP) challenge sponsored by the IEEE Signal Processing Society in 2013, and has recently been released for studying reverberant speech enhancement and recognition techniques [1]. This paper focuses on the speech recognition task, which is a medium-sized vocabulary continuous speech recognition task, in order to evaluate the ASR performance in reverberant environments.

In such a scenario, speech enhancement before ASR is important and impacts ASR performance. We have proposed a single-channel dereverberation method [2]. This

*Correspondence: Tachioka.Yuki@eb.MitsubishiElectric.co.jp
[1] Information Technology R&D Center, Mitsubishi Electric, 5-1-1, Ofuna, Kamakura, Japan
Full list of author information is available at the end of the article

method first estimates a reverberation time, which is one of the most important parameters for characterizing the extent of reverberation, and attempts to eliminate the reverberant components based on the estimated reverberation time. In addition, in order to exploit the eight-channel data provided by the REVERB challenge, we use a beamforming (BF) approach [3] with a direction-of-arrival estimation [4, 5].

In addition to the speech enhancement process, we focus on the state-of-the-art ASR techniques. Recently, ASR performance has been significantly improved owing to various types of discriminative training [6, 7] and feature transformations [8–13]. In the previous Computational Hearing in Multisource Environments (CHiME) challenge [14], we showed the effectiveness of discriminative training and feature transformations in noisy environments [15, 16], and this time, also our proposed system employs these techniques. However, the CHiME challenge and other existing evaluation campaigns for noise-robust ASR [14, 17] mainly focus on the variety of non-stationary additive noises, and the variety of room shapes or room types in these campaigns is very limited. On the other hand, the REVERB challenge [1] includes eight different reverberant environments: four rooms, which are composed of three simulated rooms and one real recorded room, multiplied by two types of source-to-microphone distances. In this scenario, due to the variety in the evaluation environments and the mismatch between simulated training data and real test data, discriminative training would cause over-training problems, although discriminative training is very powerful for matched conditions where training and evaluation conditions are close, in general. Therefore, it is important to confirm that speech recognition systems with discriminative training and feature transformations perform robustly in various reverberant environments.

This paper deals with two feature transformation approaches: linear transformation and non-linear discriminative feature transformation. The former approach converts original feature vectors to new feature vectors based on linear transformation matrices. This paper deals with linear discriminant analysis (LDA) [8] and maximum likelihood linear transformation (MLLT) [9, 10] to estimate the transformation matrices. LDA uses long context input features, which are obtained by concatenating multiple features in contiguous frames, as original feature vectors to exploit feature dynamics. Therefore, LDA can reduce the influence of reverberation because the long context input features can handle the distorted speech features across several frames due to the influence of longer reverberation than the window size of the short-time Fourier transform (STFT) [18, 19]. This property is particularly effective for reverberant speech recognition, and this paper investigates the effectiveness of LDA on ASR performance in detail with varying context sizes. In addition, MLLT finds a linear transformation of features to reduce state-conditional feature correlations. For the latter approach, we use non-linear discriminative feature transformation [12], which directly reduces ASR errors by estimating non-linear feature transformation matrix with discriminative criteria.

The above feature transformation techniques estimate transformation matrices in the training stage. However, to improve recognition accuracy for unknown conditions in the evaluation stage, the adaptation strategy of estimating feature transformation matrices for evaluation data is also effective. This paper deals with basis feature-space maximum likelihood linear regression (basis fMLLR) [20], which can estimate transformation matrices robustly even in the cases of short utterances. In addition, in the training stage, speaker adaptive training (SAT) [11] is also used. It trains acoustic models in a canonical speaker space based on the MLLR framework in order to obtain better feature transformation in the adaptation stage.

After the feature transformations, Gaussian mixture model (GMM)-based acoustic models are obtained by using discriminative training techniques [6, 7] and also this paper deals with deep neural networks (DNN) [13] that have recently attracted great attention, and we have shown promising results in noisy environments [16]. Note that the lower layers of a DNN play the role of discriminative feature transformation [21], and our DNN system skips discriminative feature transformation, which is already included in a DNN.

The studies above mainly focus on a single ASR system. On the other hand, the use of multiple systems is another solution to improve the robustness of ASR performance [22–24]. For our proposed method, which exploits discriminative training methods, the best performing system is different from environment to environment due to the variety of evaluation data or mismatch between training and evaluation data. The system combination methods relax the degradation of speech recognition performance coming from these varieties or mismatches, e.g., [25, 26] proposed to use a complementary system for system combination. This paper constructs various systems that have different properties, and in particular, our proposed discriminative training method introduces complementary systems intentionally within a lattice-based discriminative training framework [27, 28]. The results from various recognizers will be combined using recognizer output voting error reduction (ROVER) [22].

In summary, there are three objectives in this paper: First, the effectiveness of dereverberation and microphone-array speech enhancement techniques is validated. Second, the effectiveness of feature transformation and discriminative training for reverberant environments is validated. The objectives here are various

types of acoustic modeling such as the GMM, subspace Gaussian mixture model (SGMM) [29], and DNN and their discriminative training. Third, to address the variety of reverberant environments, a system combination approach is introduced and its effectiveness is validated.

There are two main differences between this paper and the REVERB challenge workshop paper [30]: First, we add detailed descriptions about validated techniques and the experimental setup. For example, we detail the speech enhancement, feature transformation, and speaker adaptation parts. Second, we compare our proposed method with other participants' systems that were submitted to the workshop, which clarifies the effectiveness of our proposed method.

2 System overview

Figure 1 shows a schematic diagram of the proposed system, which consists of three components. The first component is based on a speech enhancement step, which is described in Section 3. This paper focuses on single- and eight-channel data. The speech enhancement part consists of (1) a multichannel delay-and-sum BF with direction-of-arrival estimation that enhances the direct sound compared with the reflected sound, (2) a single-channel dereverberation technique with reverberation time estimation that attempts to eliminate late reverberation, and (3) a normalized least-mean-squares (NLMS) adaptive filter algorithm that attempts to eliminate short-term distortions such as microphone difference or speech distortions caused by speech enhancement methods.

The second component is based on a feature transformation step, including several feature-level transformations (LDA, MLLT, and basis fMLLR) and discriminative feature transformation (Section 4.1). This step uses two types of features [Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP)]. By using two different types of features, it is believed that complementary hypotheses can be obtained for system combination.

The third component is based on the ASR decoding step that uses a discriminatively trained acoustic model with margin control. Three types of systems (GMM, SGMM, and DNN) are constructed. Boosted maximum mutual information (bMMI) is used for GMM and SGMM in Sections 4.2 and for DNN in Section 4.4.

In addition, Section 4.5 describes our proposed system combination approach that combines discriminatively trained complementary systems. In addition to the three types of SAT model, a GMM model without SAT is also constructed; our proposed method constructed

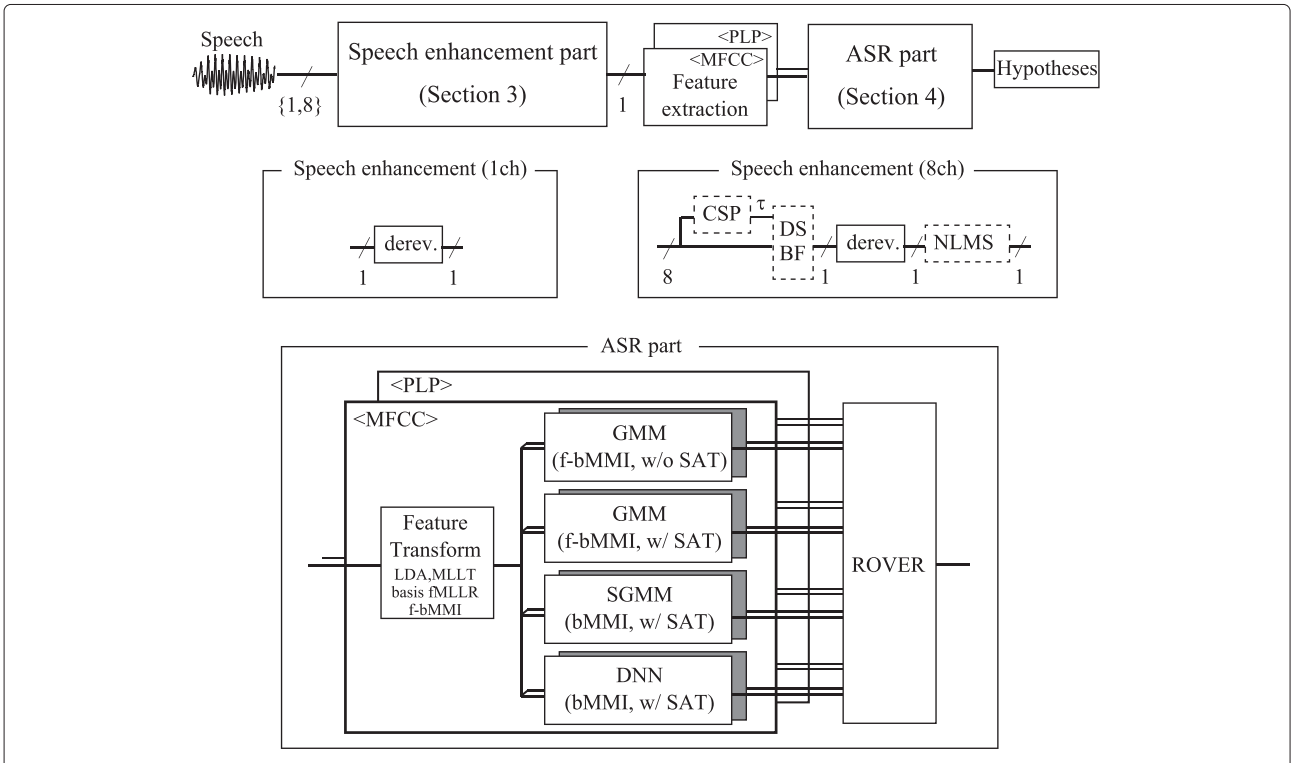

Fig. 1 Schematic diagram of the proposed system. (*CSP* cross-spectrum phase analysis, *DS-BF* delay-and-sum beamformer, *derev.* proposed dereverberation method, and *NLMS* normalized least-mean-squares adaptive filter algorithm.). *Gray blocks* are complementary systems for each system type

complementary systems for each system. The output results of 16 systems are combined using ROVER, and the final hypotheses are obtained.

3 Speech enhancement

This section deals with speech enhancement methods: delay-and-sum BF with cross-spectrum phase (CSP) analysis in Section 3.1, a proposed dereverberation method in Section 3.2, and an NLMS algorithm that attempts to eliminate short-term distortion in Section 3.3. We describe them step by step. The delay-and-sum BF using the CSP method and NLMS adaptive filter algorithm is used for an 8-channel (ch) system; the dereverberation method is used for both the 1-ch and 8-ch systems.

3.1 Delay-and-sum BF after direction-of-arrival estimation using CSP method

To enhance the direct sound from the source, a frequency-domain delay-and-sum BF is applied [3]. The time-domain sth sample $z_m(s)$ observed by the mth microphone is transformed into the STFT spectrum. The spectrum $x_{t,m}(n)$ at the tth frame and nth frequency bins obtained as

$$x_{t,m}(n) = \sum_{s=0}^{N_F-1} [\phi(s) z_m(\varphi \cdot t + s)] \exp\left(-2\pi \jmath \frac{s}{N_F} n\right), \tag{1}$$

where φ is a frame shift, and ϕ is a window function with the window length N_F. A vector form of the spectrum $\boldsymbol{x}_{t,m}$ denotes $[x_{t,m}(0), \ldots, x_{t,m}(N_F-1)]^\top \in \mathbb{C}^{N_F}$, where $\top$ denotes a transpose of vectors or matrices. The enhanced spectrum $\tilde{x}_t(n)$ is obtained by summing the spectrum $x_{t,m}(n)$ with a compensation of a time delay as

$$\tilde{x}_t(n) = \sum_m x_{t,m}(n) \cdot \exp\left(-2\pi \jmath \frac{n}{N_F} \tau_{t,m}\right). \tag{2}$$

The arrival time delay $\tau_{t,m}$ of the mth microphone from the first microphone is related to the direction of arrival at the tth frame (here $\tau_{t,1} = 0$). This time delay is estimated by CSP analysis [4]. First, an inverse STFT transform a cross-power spectrum between first and mth microphones into the time domain as

$$\psi_{t,m}(s) = \frac{1}{N_F} \sum_{n=0}^{N_F-1} \left[\phi(n) \frac{x_{t,1}(s) \cdot x^*_{t,m}(s)}{|x_{t,1}(s)||x_{t,m}(s)|}\right] \exp\left(2\pi \jmath \frac{n}{N_F} s\right), \tag{3}$$

where "*" denotes a complex conjugate. The highest correlated point is the maximum point of elements among $\{\psi_{t,m}(0), \ldots, \psi_{t,m}(N_F-1)\}$. Thus, the time delay $\tau_{t,m}$ is calculated as

$$\tau_{t,m} = \max_{s \in [0,\ldots,N_F-1]} \psi_{t,m}(s) \times \frac{1}{f_{\text{samp}}}, \tag{4}$$

where f_{samp} is a sampling frequency. To improve the performance of the original CSP method, we used a peak-hold process [31] and noise component suppression, which sets the cross-power spectrum to zero when the estimated signal-to-noise ratio (SNR) is below 0 dB [5]. Synchronous addition of multiple microphone pair-wise CSP coefficients reduces the noise influence [32].

3.2 Single-channel dereverberation with estimation of reverberation time

For a single-channel dereverberation method, we employ an algorithm proposed in [2]. The proposed algorithm is briefly described below, and detailed discussions are found in [2]. Since the proposed method is independently processed across microphones, we omit the microphone index m. When reverberation time T_r is much longer than the frame size, an observed power spectrum $\boldsymbol{X}_t = \left[|\boldsymbol{x}_t(0)|^2, \ldots, |\boldsymbol{x}_t(N_F-1)|^2\right]^\top$ is modeled as a weighted sum of the source's power spectrum $\hat{\boldsymbol{X}}_t \in \mathbb{R}^{N_F}$. The source's power spectrum is estimated as follows in the existence of stationary noise $\boldsymbol{N} \in \mathbb{R}^{N_F}$ when the spectrum between frequency bins is independent:

$$\boldsymbol{X}_t = \sum_{\mu=0}^{t} w_\mu \hat{\boldsymbol{X}}_{t-\mu} + \boldsymbol{N}, \tag{5}$$

where μ and w are the delay frame and the weight coefficient, respectively. The source's power spectrum $\hat{\boldsymbol{X}}_t$ is related to $\boldsymbol{X}_t$ as

$$\hat{\boldsymbol{X}}_{t-\mu} = \eta(T_r) \boldsymbol{X}_{t-\mu} - \boldsymbol{N}, \tag{6}$$

where η is the ratio of a direct sound component to the sum of the direct and reflected sound components, which is a decreasing function of T_r because longer T_r increases the energy of the reflected sound components. Here, we assume that the reverberation time T_r and η are independent of frequency bins, for simplicity.

Assuming that w_0 is unity to normalize reverberation decay for the direct sound, Eq. (7) can be derived from the above relations:

$$\hat{\boldsymbol{X}}_t = \boldsymbol{X}_t - \sum_{\mu=1}^{t} w_\mu \left[\eta(T_r) \boldsymbol{X}_{t-\mu} - \boldsymbol{N}\right] - \boldsymbol{N}. \tag{7}$$

Reverberation is divided into two stages: early reverberation and late reverberation. The threshold between them is denoted by D (frames) after the arrival of a direct sound.

Generally, late reverberation mainly degrades speech recognition performance and early reverberation can be ignored. Therefore, the proposed method only focuses on late reverberation. Early reverberation is complex because it greatly depends on room shapes and distributions of room materials, whereas late reverberation is statistical and the sound-energy density decays exponentially with time under the assumption of a diffuse sound field. These are modeled according to Polack's statistical model [33], and w_μ is determined as

$$w_\mu = \begin{cases} 0 & (1 \le \mu \le D) \\ \frac{\alpha_s}{\eta(T_r)} e^{-2\frac{3\log 10}{T_r}\varphi\mu} & (D < \mu) \end{cases}, \tag{8}$$

which corresponds to a reverberation decay in Fig. 2. Here, α_s is a subtraction parameter to be set. The upper condition and lower condition correspond of Eq. (8) to the early and late reverberations, respectively. Assuming η is constant, Eq. (7) is a process similar to spectral subtraction [34]. If the subtracted power spectrum $\hat{\boldsymbol{X}}_t$ is less than $\beta \boldsymbol{X}_t$, it is substituted with $\beta \boldsymbol{X}_t$. This process is called a flooring, and β is a flooring parameter. We define the floored ratio ρ as a ratio of the number of floored time-frequency bins to the total number of bins.

The proposed method estimates a reverberation time T_r from a flooring ration ρ. Two observations are exploited for this estimation. First, when some arbitrary reverberation times (T_a) are assumed, ρ increases monotonically with T_a because a longer T_a increases the extent of subtraction. This is modeled as a linear relation with the inclination Δ_ρ. Second, ρ increases with T_r at the same T_a. Since actual $\eta(T_r)$ decreases with T_r, the power spectrum after dereverberation assuming a constant η is more likely to be floored for a longer T_r because the second term of Eq. (7) is larger than that of the actual one in the condition with a longer T_r. Therefore, T_r has a positive correlation with Δ_ρ. This can be modeled as

$$T_r = a\Delta_\rho - b, \tag{9}$$

with two predetermined constants a and b.

The estimation process of T_r is summarized as follows: Calculate ρ and the inclination Δ_ρ by a least-squares regression for some values of arbitrary assumed reverberation times T_a, and estimate an actual reverberation time T_r by Eq. (9).

3.3 NLMS adaptive filter algorithm

The goal of the NLMS adaptive filter algorithm is to eliminate short-term distortions from an observed distorted signal sequence $z_s = [\, z(s-N_L+1), \ldots, z(s)]^\top \in \mathbb{R}^{N_L}$ based on a desired signal d_s [35] by using a linear filter with the tap length N_L. Filters $w'_s \in \mathbb{R}^{N_L}$ that realize these requirements are recursively trained in a manner where errors between filtered signals and desired signals are minimized as

$$\min_{w'_s} |d_s - z_s^\top w'_s|^2. \tag{10}$$

An LMS algorithm uses instantaneous values for the estimation of a gradient, and an NLMS algorithm normalizes the step size parameter by the signal power. Thus, the update formula of an NLMS algorithm is obtained as

$$w'_s = w'_{s-1} + \frac{\varrho}{\epsilon + |z_s|^2} z_s [\, d_s - z_s^\top w'_{s-1}], \tag{11}$$

where ϱ is a step size, and ϵ is a very small constant that avoids the instability of the update formula. The initial value of filter w'_0 is 0. In this case, z_s is a reverberant speech, and d_s is a clean speech without reverberation. A filter w' is obtained from the entire training data set. For evaluation, desired signals d_s cannot be obtained; thus, the filter cannot be changed. The tap length of NLMS is short because the goal of this filter is to eliminate a short-term distortion, whereas the proposed dereverberation algorithm (3.2) attempts to eliminate late reverberation.

4 Speech recognition

4.1 Feature transformation and speaker adaptation

Static features concatenated during the left L frames, current frame, and the right R frames are compressed into low-dimensional (I'-dimensional) features by using LDA.

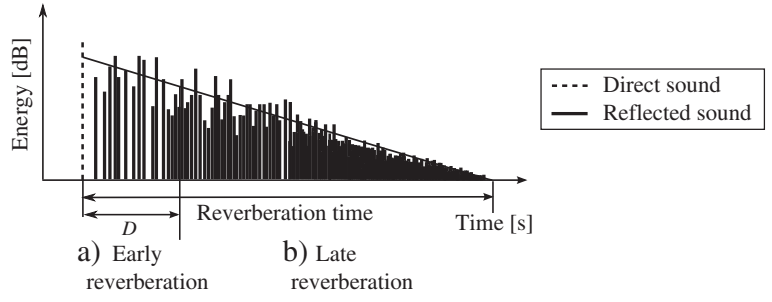

Fig. 2 *a* Early and *b* late reverberation. Early reverberation has complex and sparse reflections. Late reverberation has dense reflections and an exponentially decayed shape

The class of LDA is the state of the triphone HMM. In addition to this, to reduce the correlation between feature dimensions, MLLT is used. Combined feature transformation is realized as

$$\boldsymbol{y}'_t = \boldsymbol{A}^M \left[\boldsymbol{A}^L [\boldsymbol{y}_{t-L}^\top, \ldots, \boldsymbol{y}_t^\top, \ldots, \boldsymbol{y}_{t+R}^\top]^\top \right], \tag{12}$$

where $\boldsymbol{y}_t$ is an original I-dimensional feature at the tth frame, and $\boldsymbol{y}'_t$ is an I'-dimensional transformed feature; $\boldsymbol{A}^L \in \mathbb{R}^{I' \times (I \times (L+R+1))}$ is a transform matrix of LDA, and $\boldsymbol{A}^M \in \mathbb{R}^{I' \times I'}$ is a transform matrix of MLLT.

For adaptation, instead of a normal fMLLR transformation, the basis fMLLR [20] is used. It can robustly estimate transform matrices and bias terms even for short utterances. This method realizes the transformation of original features y'_t into adapted features y''_t by using pre-trained bases of transform matrices and bias terms and estimating their weights as

$$\boldsymbol{y}''_t = \sum_\nu \pi_\nu \left[\boldsymbol{A}^f_\nu \boldsymbol{y}'_t + \boldsymbol{b}^f_\nu \right], \tag{13}$$

where $\boldsymbol{A}^f_\nu \in \mathbb{R}^{I' \times I'}$ and $\boldsymbol{b}^f_\nu \in \mathbb{R}^{I'}$ are the νth pre-trained basis of an fMLLR transform matrix and bias term, respectively, which are estimated from entire training data. For evaluation, only their weights π_ν are estimated.

Moreover, to address the wide variety between speakers, SAT as an acoustic model adaptation [11] is frequently used. In SAT training, acoustic models are trained on speaker-adapted training data, which are transformed into canonical speaker space by using speaker adaptation techniques, in this case, fMLLR. This can reduce the influence of a speaker variation. This paper validates the effectiveness of feature transformations (LDA and MLLT) and adaptation techniques (basis fMLLR and SAT).

4.2 MMI discriminative training of acoustic model

MMI discriminative training is a supervised training algorithm that maximizes the mutual information between correct labels and recognition hypotheses. This paper focuses on bMMI [36], where a boosting factor $b \geq 0$ is used to introduce a weight depending on phoneme accuracies. The objective function is given as

$$\mathcal{F}_b(\lambda) = \sum_r \log \frac{p_\lambda \left(\boldsymbol{y}^r | \mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda \left(\boldsymbol{y}^r | \mathcal{H}_s\right)^\kappa p_L(s) \mathrm{e}^{-bA(s,s_r)}}, \tag{14}$$

where $\boldsymbol{y}^r = \left[\boldsymbol{y}_0{}^\top, \ldots, \boldsymbol{y}_{T(r)-1}{}^\top\right]^\top$ is the rth utterance's feature sequence and $T(r)$ is the total frame number of the rth utterance. The acoustic model parameters λ are optimized by the extended Baum-Welch algorithm. λ is a mean, variance, and mixture weight of GMM. $\mathcal{H}_{s_r}$ and $\mathcal{H}_s$ are the HMM sequences of the correct label s_r and a hypothesis s, respectively; p_λ is the acoustic model likelihood; κ is the acoustic scale; p_L is the language model likelihood; and $A(s, s_r)$ is the phoneme accuracy of s for s_r. This paper compares the performances of bMMI training of GMM and SGMM to those of maximum likelihood (ML) training.

4.3 Discriminative feature transforms

The extension of a discriminative training to a feature transformation is referred to as a feature-space discriminative training [12]. It estimates a matrix $\boldsymbol{M} \in \mathbb{R}^{I' \times J}$ that projects rich, high-dimensional features $\boldsymbol{h}_t \in \mathbb{R}^J$ ($J \gg I'$) down to low-dimensional transformed features, as follows:

$$\boldsymbol{\nu}_t = \boldsymbol{y}''_t + \boldsymbol{M}\boldsymbol{h}_t. \tag{15}$$

Usually, Gaussian posteriors of an N_g-mix universal background model (UBM) are used for $\boldsymbol{h}_t$ [37]. The objective function can be obtained simply by replacing $\boldsymbol{y}^r$ with the rth utterance's transformed feature sequence $\boldsymbol{\nu}^r = \left[\boldsymbol{\nu}_0{}^\top, \ldots, \boldsymbol{\nu}_{T(r)-1}{}^\top\right]^\top$ in Eq. (14) as

$$\mathcal{F}_b(\boldsymbol{M}) = \sum_r \log \frac{p_\lambda \left(\boldsymbol{\nu}^r | \mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda \left(\boldsymbol{\nu}^r | \mathcal{H}_s\right)^\kappa p_L(s) \mathrm{e}^{-bA(s,s_r)}}. \tag{16}$$

The matrices $\boldsymbol{M}$ are optimized by maximizing the objective function $\mathcal{F}_b(\boldsymbol{M})$. In this study, we validate the effectiveness of a feature-space bMMI (f-bMMI).

4.4 Discriminative training of DNN

In a DNN-HMM hybrid system, sequential discriminative training according to the (b)MMI criterion (14) has been proposed [38] in addition to a usual cross-entropy (CE) training. The DNN provides posterior probabilities for the HMM state j. The acoustic likelihood p_θ is replaced by a pseudo likelihood as

$$p_\theta \left(\boldsymbol{y}^r | j\right) = \frac{p_\theta \left(j | \boldsymbol{y}^r\right)}{p_0 \left(j\right)}, \tag{17}$$

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

$$p_\theta(j|\boldsymbol{y}^r) = \frac{\exp a_j(\boldsymbol{y}^r)}{\sum_{j'} \exp a_{j'}(\boldsymbol{y}^r)}, \tag{18}$$

where a_j is the activation of the jth unit in the output layer. θ is a parameter in weight matrices and bias terms

of DNN. These activations are trained discriminatively according to the bMMI criterion. The bMMI objective function is the same as Eq. (14), simply by replacing λ with θ: $\mathcal{F}_b(\theta)$.

4.5 Constructing complementary system suitable for system combination

We describe a discriminative method that constructs complementary systems for appropriate system combination [27, 28]. Complementary systems are constructed by discriminatively training a model, which begins with 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 and fixed, the discriminative training objective function $\mathcal{F}^c$ for building a complementary system is

$$\mathcal{F}^c(\mathcal{M}) = (1+\alpha_c)\mathcal{F}_b(\mathcal{M}) - \frac{\alpha_c}{Q}\sum_{q=1}^{Q}\mathcal{F}_{b_1}(\mathcal{M}), \qquad (19)$$

where $\mathcal{F}_{b_1}$ is a $\mathcal{F}_b$ just replaced by b with b_1. Derived formula was

$$\begin{aligned}\mathcal{F}^c(\mathcal{M}) = \mathcal{F}_b(\mathcal{M}) \\ + \alpha_c \sum_r \Bigg[\frac{1}{Q}\sum_{q=1}^{Q} \log p_{\mathcal{M}}\left(\boldsymbol{y}^r|\mathcal{H}_{s_q}\right)^{\kappa} p_L(s_q) \mathrm{e}^{-b_1 A(s_q,s_r)} \\ - \log \sum_s p_{\mathcal{M}}\left(\boldsymbol{y}^r|\mathcal{H}_s\right)^{\kappa} p_L(s) \mathrm{e}^{-bA(s,s_r)} \Bigg],\end{aligned} \qquad (20)$$

where $\mathcal{M}$ is the set of model parameters of a complementary system to be optimized; that is, λ, $\boldsymbol{M}$, and θ. α_c is a scaling factor. The model parameter $\boldsymbol{M}$ is shared among the original $\mathcal{F}$ and the Q base models' $\mathcal{F}$ to be optimized. This subtracts an objective function related to the one-best hypothesis of the qth base system, s_q, from an objective function related to the correct label s_r. The discriminative criterion $\mathcal{F}$ is selected as bMMI or f-bMMI. If α_c equals zero, this objective function matches the original $\mathcal{F}$. The first term in Eq. (19) promotes a good performance according to the discriminative training criterion, whereas the second term makes the target system generate hypotheses that have different tendencies from the original Q base models. This procedure is commonly used to obtain the objective functions of Sections 4.2, 4.3, and 4.4.

5 Experimental setup

5.1 REVERB challenge speech recognition task

We validated the effectiveness of our proposed approaches for a reverberated speech recognition task on the REVERB challenge [1] data. The task is a medium-vocabulary ASR in reverberant environments, whose utterances are taken from the *Wall Street Journal* (*WSJ*) database (WSJCAMO [39]). This database includes two types of data: SimData created by convolving clean speech with six types of room impulse responses at a distance of 0.5 m (near) or 2 m (far) from the microphones in three offices (Rooms 1, 2, and 3) whose reverberation times are 0.25, 0.5, and 0.75 s, respectively, with relatively stationary noise at 20 dB SNR; and RealData created by recording real-world speech at a distance of 1 m or less (near) or 2.5 m or less (far) from the microphones in one room (Room 1) with stationary noise such as air conditioner noise. Eight microphones were arranged on the circle with a radius of 0.1 m. The number of speakers and utterances of the training set (*tr*), evaluation set (*eva*), and development set (*dev*) is shown in Table 1.

Acoustic models were trained using *tr*. Some of the parameters, e.g., language model weights, were tuned based on the WERs of *dev*. The vocabulary size is 5 k, and a trigram language model is used. The REVERB challenge speech recognition task is categorized in terms of processing techniques, training data of the acoustic model, recognizer type, and number of channels used, as shown in Table 2. All experiments in this paper were "utterance-based batch processing,"[1] "acoustic model trained on the challenge provided multicondition (MC) training data," "own recognizer," and "single- or eight-channel data". These systems were constructed by using the Kaldi toolkit [40].

5.2 Speech enhancement

The REVERB challenge provides single-, two-, and eight-channel data. We used single- and eight-channel data. For single- and eight-channel data, the proposed dereverberation technique was used with parameters: $D = 9$, $\alpha = 5$, $\beta = 0.05$, $a = 0.005$, and $b = 0.6$. For eight-channel data, before dereverberation, delay-and-sum BF with a direction of arrival estimation by CSP analysis was performed, which used a total of ${}_8C_2 (= 28)$ pairs of microphones. After dereverberation, NLMS adaptive filters with $N_L = 200$ taps were applied.

Table 1 Number of speakers and utterances of training (*tr*), development (*dev*), and evaluation (*eva*) set for the REVERB challenge

Set		Number of speakers	Number of utterances
tr	–	92	7861
dev	SimData	20	1484
	RealData	5	179
eva	SimData	28	2 176
	RealData	10	372

Table 2 Category of the REVERB challenge speech recognition task

	Type
Processing scheme	Full batch, *utterance-based*, real-time
Training data of acoustic model	Own dataset, *multi-condition*, clean
Recognizer type	*own recognizer*, Challenge baseline recognizer
Number of channels used	*1*, *2*, *8*

Italicized data denotes the category to which this paper belongs

5.3 Feature extraction and transformation and acoustic model adaptation

We describe the settings of acoustic features and feature transformations, which are detailed in [15, 16]. The baseline acoustic features were 0–12 order MFCCs and PLPs with first and second dynamic features. After concatenating static MFCCs/PLPs during $L + R + 1$ frames without using delta feature, a total of $(13 \times (L+R+1))$-dimensional features were compressed into 40 dimensions by the LDA.

For adaptation, when speaker IDs were known for the training set, bases $\boldsymbol{A}_{\nu}^{f}$ and $\boldsymbol{b}_{\nu}^{f}$ were estimated. For the development and evaluation set, speaker IDs are assumed to be unknown, and weight vector π_{ν} was estimated.

5.4 Discriminative methods

In discriminative feature transformation (Section 4.3), a UBM with $N_g = 400$-mix Gaussians was used. The offset features were calculated for each composed of 40-dimensional features, including MFCC/PLP features with dynamic features (39 dimensions in total) and the posterior probability of it, with context expansion (contiguous nine frames). The number of dimensions of feature vector $\boldsymbol{h}_t$ was 400[Gauss] $\times$ 40[dim/(Gauss $\cdot$ frame)] $\times$ 9[frame]. Features with the top two GMM posteriors were selected and all other features were ignored.

The boosting factor b of bMMI and f-bMMI was 0.1. To construct complementary systems, the additional boosting factor b_1 in the second term of Eq. (19) was 0.3 and α_c was 0.75. For f-bMMI, in one iteration, f-bMMI for the matrix $\boldsymbol{M}$ was coupled with bMMI for the acoustic model parameters λ.

5.5 Building acoustic models

First, clean acoustic models were trained. The number of monophones was 45, including silence ("sil"). Triphone model has 2500 states and 15,000 Gaussian distributions. Second, using the alignments and triphone tree structures of the clean model, reverberated acoustic models were trained on the MC dataset according to the ML

Table 3 WER [%] in terms of rooms and microphone distances on the REVERB challenge *dev* set using single-channel data and MFCC features

			SimData							RealData		
			Room 1		Room 2		Room 3		Avg	Room 1		Avg
	Feature	Type	Near	Far	Near	Far	Near	Far		Near	Far	
Kaldi baseline	MFCC	ML	10.96	12.56	15.70	34.21	19.61	39.24	22.05	48.53	47.37	47.95
derev.			12.41	14.68	14.03	27.16	16.39	33.85	19.75	47.04	44.57	45.81
GMM	+LDA+MLLT	ML	9.46	11.01	11.51	22.04	13.08	28.09	15.87	39.99	40.67	40.33
	+basis fMLLR		7.77	10.00	9.76	19.28	11.05	24.90	13.79	33.00	35.54	34.27
		bMMI	7.13	9.61	9.12	16.19	10.46	21.98	12.42	30.69	35.20	32.95
		f-bMMI	6.27	8.73	8.28	14.89	9.37	19.54	11.18	*28.32*	*31.31*	*29.82*
		f-bMMI$_c$	7.06	9.05	8.58	14.96	10.16	20.43	11.71	29.01	31.72	30.37
	+SAT	ML	8.87	11.21	9.71	19.89	10.95	24.04	14.11	36.06	36.23	36.15
		bMMI	6.56	8.51	7.76	16.24	9.03	19.88	11.33	34.19	37.53	35.86
		f-bMMI	5.88	7.60	7.25	14.59	8.09	17.51	10.15	31.63	34.72	33.18
		f-bMMI$_c$	6.07	7.82	7.22	14.89	8.43	17.51	10.32	32.38	35.27	33.83
SGMM		ML	6.47	9.07	8.18	17.11	9.55	20.40	11.80	33.13	34.93	34.03
		bMMI	5.53	7.23	7.00	14.44	7.76	17.48	9.91	31.50	33.36	32.43
		bMMI$_c$	5.68	7.28	7.02	14.44	7.94	17.68	10.01	30.94	33.08	32.01
DNN		CE	6.71	8.85	8.70	15.58	9.15	19.07	11.34	30.88	35.82	33.35
		bMMI	5.29	7.06	6.95	13.09	7.57	15.53	9.25	28.45	32.67	30.56
		bMMI$_c$	*5.14*	*6.74*	*6.51*	*12.37*	*7.27*	*15.50*	*8.92*	*28.32*	33.49	30.91

The proposed dereverberation method was used. Three types of acoustic models (GMM, SGMM, and DNN) were constructed with feature transformation (LDA + MLLT), adaptation (basis fMLLR and SAT), and discriminative training (bMMI and f-bMMI). Subscript letter "c" represents the proposed "complementary" system. Italicized data were the best systems in each condition

criterion. Finally, from this ML model, we performed the discriminative training and feature transformations.

For DNNs, we used Povey's implementation of neural network training in Kaldi [40]. DNN has two hidden layers was two and each hidden layer has 642 nodes. The total number of parameters was 2 M. The initial learning rate of CE training was 0.02, and this decreased to 0.004 at the end of training. The training targets for the DNN were determined by the forced alignments on reverberant speech using a GMM model with SAT. The parameters used in our experiments were set as those in the *WSJ* tutorial (s6) attached to the Kaldi toolkit, although some settings such as the number of model parameters or some minor parameters were modified.

5.6 System combination

We prepared three types of ASR acoustic model systems for the challenge: GMM, SGMM, and DNN. To improve the performance of the respective systems, for GMM, f-bMMI was used; whereas for SGMM and DNN, bMMI was used. On the development set, because output tendencies of GMM with and without SAT model were different, both systems were used for a system combination. For each system, complementary systems were constructed by the proposed method as shown in 4.5. These systems were trained both for MFCC and PLP features; thus, a total of 16 systems were prepared. After decoding for generated lattices, minimum Bayes risk decoding [41], which slightly improved the performance, was commonly used.

5.7 Black-box optimization

Bayesian optimization using Gaussian processes [42] was applied to various speech recognition problems including neural network [43] and HMM topology optimization [44]. In this paper, we also applied this technique to the selection of combined systems and the parameter optimization for ROVER. The objective function of the optimization was WER of the development set.

6 Results and discussion

6.1 Baseline and speech enhancement techniques

Tables 3 and 4 show the WERs of the development set (*dev*) for three simulated rooms and one real room with two types of source-to-microphone distances (near/far). Table 3 is based on a single-channel one and Table 4 is based on an eight-channel one. The "Kaldi baseline" in Table 3 is an acoustic model trained on the MC data without speech enhancement. "derev." is the proposed dereverberation method with a reverberation time estimation. Although, for some cases in room 1, the reverberation time is fairly short and the proposed method degraded

Table 4 WER [%] on the REVERB challenge *dev* set using eight-channel data and MFCC features

			SimData							RealData		
			Room 1		Room 2		Room 3		Avg	Room 1		Avg
	Feature	Type	Near	Far	Near	Far	Near	Far		Near	Far	
CSP+BF+derev.	MFCC	ML	10.79	12.19	11.02	16.71	11.47	20.43	13.77	40.36	42.83	41.60
+NLMS			11.11	12.27	11.81	17.40	12.34	21.46	14.40	38.37	40.74	39.56
GMM	+LDA+MLLT	ML	8.38	10.30	9.91	14.94	10.19	17.28	11.83	34.06	37.18	35.62
	+basis fMLLR		7.74	9.22	8.80	13.33	9.05	15.28	10.57	27.39	30.14	28.77
		bMMI	6.64	8.21	7.25	11.39	7.10	11.50	8.68	24.89	27.96	26.43
		f-bMMI	6.19	7.40	7.39	10.13	6.58	10.24	7.99	*22.58*	*26.25*	*24.42*
		f-bMMI$_c$	6.39	7.33	7.44	9.86	6.70	10.44	8.03	22.71	27.41	25.06
	+SAT	ML	7.25	9.32	8.70	12.79	8.33	13.80	10.03	28.88	32.88	30.88
		bMMI	5.24	7.10	6.56	9.93	5.98	10.98	7.63	26.58	30.83	28.71
		f-bMMI	5.01	6.76	*5.96*	*9.07*	5.84	*9.40*	7.01	24.27	29.60	26.94
		f-bMMI$_c$	5.16	6.93	6.11	9.49	5.96	9.67	7.22	24.27	29.73	27.00
SGMM		ML	5.65	7.62	7.47	10.97	7.00	11.45	8.36	25.27	30.35	27.81
		bMMI	*4.57*	*6.05*	6.19	9.27	6.01	9.89	*7.00*	24.70	30.01	27.36
		bMMI$_c$	4.72	6.10	6.09	9.56	6.18	10.01	7.11	24.39	30.01	27.20
DNN		CE	6.49	7.45	7.84	11.44	7.25	11.97	8.74	25.27	29.32	27.30
		bMMI	5.56	6.27	6.24	9.29	5.71	10.44	7.25	23.27	28.84	26.06
		bMMI$_c$	5.26	*6.05*	6.21	9.10	*5.61*	10.06	7.05	22.65	28.50	25.58

In addition to the proposed dereverberation method, BF with direction of arrival estimation by CSP analysis and NLMS adaptive filters were used. Subscript letter "c" represents the proposed "complementary" system. Italicized data were the best systems in each condition

Table 5 Average WER [%] on the REVERB challenge *dev* set using PLP features

			1 ch		8 ch	
	Feature		SimData	RealData	SimData	RealData
Kaldi baseline	PLP	ML	22.96	48.90		
derev.			19.84	44.15		
CSP+BF+derev.					13.98	42.21
+NLMS					14.97	41.15
GMM	+LDA+MLLT	ML	15.63	40.36	12.13	35.11
	+basis fMLLR		13.70	34.21	10.73	29.21
		bMMI	12.78	33.43	8.94	26.84
		f-bMMI	11.91	30.67	8.10	*25.72*
		f-bMMI$_c$	12.20	31.67	8.26	26.30
	+SAT	ML	13.55	36.25	10.17	30.85
		bMMI	11.05	35.63	8.06	28.45
		f-bMMI	10.14	33.29	7.32	26.78
		f-bMMI$_c$	12.20	31.67	7.61	27.59
SGMM		ML	11.90	32.95	8.43	26.99
		bMMI	10.25	33.10	7.13	26.67
		bMMI$_c$	10.30	33.14	7.19	27.21
DNN		CE	11.30	31.87	8.75	27.33
		bMMI	9.44	30.19	7.25	26.06
		bMMI$_c$	*9.40*	*30.13*	*6.74*	26.37

Subscript letter "c" represents the proposed "complementary" system. Italicized data were the best systems in each condition

performance, for other cases and on average, performance was improved by approximately 2 %. Weninger et al. [45] showed that our proposed dereverberation technique is effective even with a state-of-the-art denoising auto-encoder. For the eight-channel data shown in Table 4, BF with "derev." significantly improved performance by approximately 6.3–8.3 % on average, because the direction of arrival estimation was stable and reliable. "NLMS" improved the WER by 2.0 % for the RealData, but degraded the WER by 0.6 % for the SimData. However, because these decreases in performance have less impact than the improvements, we used NLMS below.

These results above used MFCC features. Experimental results using PLP features are shown in Table 5. On average, the ASR performances using PLP features were approximately 0.2–1 % lower than those using MFCC features; however, their error tendencies were fairly different, which was a good property for system combination.

6.2 LDA and MLLT feature transformation and adaptation

LDA and MLLT feature transformations significantly improved performance by approximately 2.6–5.5 %. Table 6 shows the effect of an LDA context size on performance. The performance of the SimData could not be improved by context sizes longer than 4. For the RealData, performance could be improved in several cases by adding more right context, but generally not by adding left context. In reverberant environments, because reverberant components of current frames give an influence on the features in the right context, the right context can be useful for improving speech recognition performance. In the end, we kept the context size at the default setting, $L = R = 4$.

Tables 3 and 4 show that the adaptation technique, basis fMLLR, improved performance by approximately 1.3–6.9 %. The effect of SAT is unstable between environments.

Table 6 Average WER[%] investigating the effect of LDA context sizes [left (*L*) and right (*R*)] on the REVERB challenge *dev* set using eight-channel data

	SimData				RealData			
$L \setminus R$	4	5	6	7	4	5	6	7
4	*11.83*	12.20	12.10	12.57	35.62	34.31	34.10	36.22
5	12.14	12.32	12.46	12.72	34.71	35.34	34.44	*33.31*
6	12.57	12.33	12.56	12.87	35.49	35.29	34.19	35.11
7	12.83	12.94	13.43	13.49	35.13	35.90	35.67	36.00

Italicized data were the best systems in each condition

Table 7 Average WER [%] investigating the effect of iteration numbers of bMMI and f-bMMI discriminative training with SAT on the REVERB challenge *dev* set using eight-channel data

	MFCC				PLP			
	Number of iterations							
	1	2	3	4	1	2	3	4
bMMI								
SimData	8.70	8.41	8.18	*7.63*	9.02	8.64	8.47	*8.06*
RealData	29.21	28.34	*28.16*	28.71	29.74	29.26	28.91	*28.45*
f-bMMI								
SimData	8.07	7.56	7.30	*7.01*	8.47	7.93	7.57	*7.32*
RealData	27.70	27.29	27.16	*26.94*	29.36	27.86	27.15	*26.78*

Italicized data were the best systems in each condition

6.3 Discriminative training of acoustic model and discriminative feature transformation

Tables 3 and 4 show that the discriminative training was effective for reverberant environments. The performances of f-bMMI training were higher than those of bMMI training in all cases by approximately 0.6–1.7 %. The WERs of our complementary systems were only slightly lower (0.2–0.7 %) than those of the base systems, and they have different tendencies from base systems; thus, they appear to be well suited to system combination.

Table 7 shows the effect of the iteration numbers of bMMI and f-bMMI on the development set performance. The results show that the best performance was achieved at four iterations.

6.4 SGMM and DNN

Tables 3 and 4 show the performance of SGMM acoustic models. For the SimData, the performance of SGMMs was higher than that of GMMs. However, for the RealData, the performance was lower than that of GMMs. Because the RealData were noisier than the SimData, the estimation of speaker vector can be unstable.

DNN acoustic models achieved the best performance for the SimData. Although the best system for the RealData was GMM without SAT, DNN was the second best. On average over the SimData and RealData, DNNs achieved the best performance. Although DNN was trained discriminatively even by CE training according to the frame-level discriminative criterion, sequence discriminative training, bMMI, for DNN systems turned out to be as effective as for other systems.

6.5 System combination

We tested five types of system combinations, as shown in Table 8. The number 2 stands for one MFCC system and one PLP system. The number 4 stands for two MFCC and two PLP systems composed of a base system and the proposed complementary system. These systems' outputs are combined by using ROVER. The ID 1) system was a combination of SAT-GMMs (f-bMMI) using both MFCC and PLP features. The performance for the RealData improved by 1.2–4.2 % over the f-bMMI with a SAT (MFCC) single system. For the GMM system without SAT, using f-bMMI [ID 2)], the WER improved by 0.2–1.5 % for the SimData and 0.6–1.4 % for the RealData. Including the complementary systems

Table 8 WER [%] on the REVERB challenge *dev* set, with system combination using both MFCC and PLP features

						SimData							RealData		
		Number of systems				Room 1		Room 2		Room 3		Avg	Room 1		Avg
	ID	GMM	SAT-GMM	SGMM	DNN	Near	Far	Near	Far	Near	Far		Near	Far	
1 ch	1)		2			6.00	8.19	7.52	14.37	8.78	18.35	10.54	27.70	30.35	29.03
	2)	2	2			5.31	6.37	6.58	12.62	7.42	16.00	9.05	27.26	29.60	28.43
	3)	4	4			5.33	6.39	6.63	12.67	7.49	15.60	9.02	27.01	29.67	28.34
	4)	4	4	4		5.01	6.34	6.33	12.45	6.87	15.43	8.74	26.64	29.80	28.22
	5)	4	4	4	4	4.67	5.88	6.31	*11.93*	6.63	*14.89*	8.39	26.58	*28.91*	*27.75*
	6)	2	2	2	2	*4.52*	*5.68*	*6.29*	12.00	*6.50*	15.06	*8.34*	*26.45*	29.80	28.13
8 ch	1)		2			4.72	5.83	5.96	8.92	5.37	8.75	6.59	23.27	28.30	25.79
	2)	2	2			4.72	6.02	5.72	8.26	5.14	8.56	6.40	22.27	26.59	24.43
	3)	4	4			4.72	5.83	5.77	8.21	5.19	8.38	6.35	22.52	26.52	24.52
	4)	4	4	4		*4.08*	5.16	5.62	*7.79*	*4.80*	8.38	5.97	22.40	27.00	24.70
	5)	4	4	4	4	4.18	*5.11*	*5.50*	*7.74*	4.85	*8.23*	*5.94*	21.90	26.52	24.21
	6)	3	1	4	2	4.18	5.51	*5.50*	*7.74*	4.97	8.43	6.06	*21.58*	*26.32*	*23.95*

For GMM systems, f-bMMI is used, while for SGMM and DNN systems, bMMI is used. The number 2 stands for MFCC and PLP systems, and the number 4 stands for MFCC and PLP systems along with their complementary systems. ROVER 6) uses black-box optimization at the stage of system selection and parameter optimization for ROVER. Italicized data were the best systems in each condition

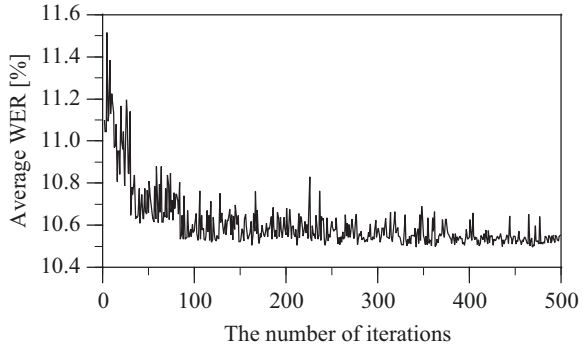

Fig. 3 WER [%] averaged over SIMDATA and REALDATA through black-box optimization of the system selection and parameter setting for ROVER in terms of the number of iterations

[ID 3)], the WER improved slightly. For the best case, WER improved by 0.4 %, while for the worst case, WER decreased by 0.1 %. This shows the effectiveness of our proposed method. Adding in SGMMs [ID 4)], which was effective for the SIMDATA, the performance for the SIMDATA further improved by 0.3–0.4 %. Taking into account DNNs [ID 5)], the performance was again improved; this system, which combined 16 systems in total, achieved the best average performance on the development set. For the reference, the results of eight system combination without using our proposed combination are added to the last line of 1 ch case [ID 6)]. The WER on REALDATA was worse than those of the proposed 16 system combination, which shows that the complementary training generalizes the ASR results for unseen data conditions more.

In all cases except for the room 1/far(8-ch) condition,[2] the performances were better than those of the best system. This shows that the system combination approach is effective for the case where reverberant environments are various.

6.6 Black-box optimization

For eight-channel data, black-box optimization was performed. Figure 3 shows the average WER in terms of the iteration number. WER almost decreased monotonically and, after 100 iterations, it converged. Among these iterations, the results that achieved the best WER on average, are shown in the last column of Table 8. The performance improved mainly for the REALDATA.

6.7 Evaluation set

Table 9 shows the results for the evaluation set (*eva*). Legend of the table is the same to the development set. The optimal system combination is determined based on the WER on the development set. The discriminative training of acoustic model (bMMI) and feature-space discriminative training (f-bMMI) significantly improved the performance. SGMM was better than GMM because model adaptation was well performed. DNN outperformed GMM and SGMM. The DNN with discriminative

Table 9 WER [%] on the REVERB challenge *eva* set

		SIMDATA							REALDATA		
		Room 1		Room 2		Room 3		Avg	Room 1		Avg
		Near	Far	Near	Far	Near	Far		Near	Far	
1 ch	Kaldi baseline	13.23	14.13	15.54	29.69	20.06	37.44	21.68	50.62	45.98	48.30
	derev.	12.50	13.43	14.61	24.71	17.09	32.62	19.16	44.75	43.32	44.04
	GMM (f-bMMI)	7.27	8.17	8.82	14.11	10.54	18.76	11.28	28.65	29.54	29.10
	GMM (SAT,f-bMMI)	6.44	7.22	7.57	13.97	9.52	18.44	10.53	28.87	29.78	29.33
	SGMM (SAT, bMMI)	5.81	6.54	7.22	13.84	8.70	18.17	10.05	27.75	28.36	28.06
	DNN (SAT, bMMI)	5.90	6.84	7.35	12.57	9.40	16.55	*9.77*	25.97	25.69	*25.83*
	ROVER 5)	*5.30*	*5.61*	*6.30*	*11.16*	*7.76*	*14.95*	*8.51*	*23.79*	*23.60*	*23.70*
8 ch	CSP+BF+derev.	10.94	11.69	10.98	16.33	12.79	21.39	14.02	34.33	36.93	35.63
	+NLMS	10.94	12.32	11.38	17.59	13.46	22.96	14.78	35.32	35.28	35.30
	GMM (f-bMMI)	6.57	6.93	6.80	9.93	7.47	12.76	8.41	20.22	23.19	21.71
	GMM (SAT, f-bMMI)	6.17	6.64	6.51	10.13	7.40	13.15	8.33	20.63	23.67	22.15
	SGMM (SAT, bMMI)	5.86	6.44	6.29	9.23	6.96	12.83	7.94	20.66	23.50	22.08
	DNN (SAT, bMMI)	5.64	6.18	6.16	9.29	7.08	12.40	*7.79*	19.35	22.28	*20.82*
	ROVER 5)	*4.96*	5.62	*5.58*	8.18	*5.73*	*10.47*	*6.76*	*16.90*	*20.29*	*18.60*
	ROVER 6)	5.00	*5.56*	*5.38*	*8.15*	*5.73*	10.70	*6.75*	17.47	20.36	18.93

All systems except ROVER are single systems. MFCC feature was used for single system, and MFCC and PLP features were used for ROVER 5). Italicized data were the best systems in each condition

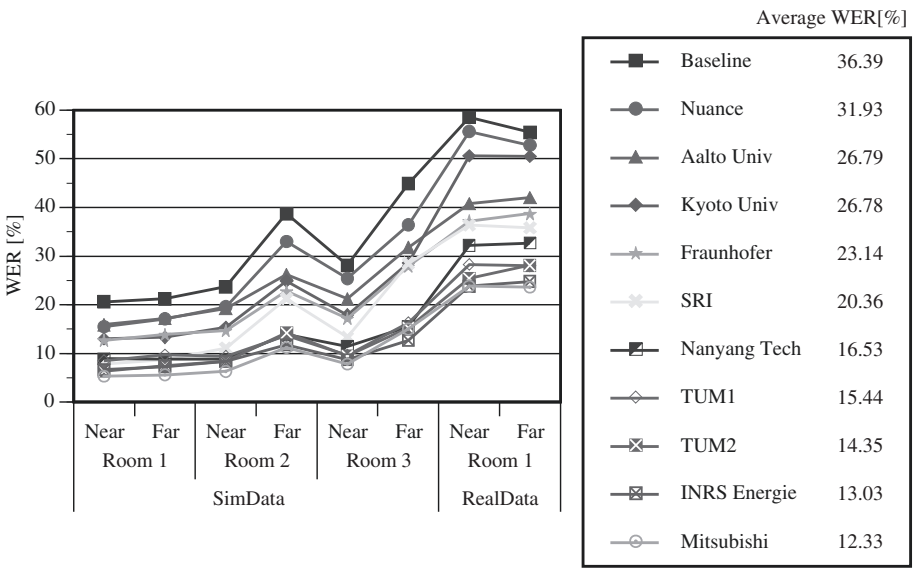

Fig. 4 Comparison of WER [%] on REVERB challenge *eva* set among REVERB challenge participants for 1-ch data. (Multicondition/clean, own recognizer/challenge baseline recognizer.)

training achieved the best performance for the SIMDATA and REALDATA among single systems. This shows the robustness of DNN in unseen conditions. Moreover, system combination [ROVER 5)] improved the WER by 1.0–1.3 % for the SIMDATA and 2.1–2.2 % for the REALDATA. Among system combination systems, the performance of ROVER 5) was better than that of ROVER 6), which used black-box optimization and could be overly tuned on the development set.

6.8 Comparison to other participants' results in the REVERB challenge workshop

The results in the previous section were submitted to the REVERB challenge workshop. Figure 4 shows the WERs for the single-channel data of other participants who belong to the same category, which corresponds to all cases except "own dataset" in the training data of the acoustic models in Table 2. Figure 5 shows those for the eight-channel data. For speech enhancement purposes, a long–short-term memory recurrent neural network (LSTM-RNN) was effective [46] ("TUM2" in the figure). Many participants used DNN-based acoustic modeling (e.g., [47] "Nanyang Tech" in the figure). Speaker adaptation of DNN based on the *i*-vector technique in addition to robust features, also performed well [48] ("INRS Energie" in the figure). We achieved the best performances in both single- and eight-channel cases.[3]

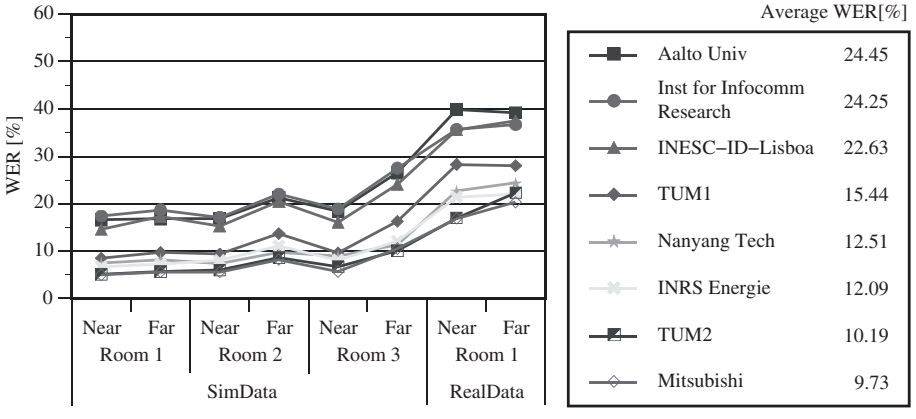

Fig. 5 Comparison of WER [%] among REVERB challenge participants for 8-ch data

7 Conclusions

We evaluated the medium-sized vocabulary continuous speech recognition task of the REVERB challenge in order to validate the effectiveness of single-channel dereverberation and multi-channel beamforming techniques and discriminative training of acoustic model and feature transformation in reverberant environments. For speech enhancement, experiments show the effectiveness of dereverberation of the late reverberation components, and beamforming using multiple microphones that enhances direct sounds compared to the reflected sounds.

For speech recognition, we validated the effectiveness of feature transformations and discriminative training. Experiments show that these techniques are effective across various types of reverberation as well as in noisy environments. To improve robustness in eight types of environments, the system combination approach was used. Systems from 2 to 16 were constructed to address the problem where the best performing system was different from environment to environment. System combination improved performance; in almost all cases, the combined system outperformed the best performing single system. Our proposed method to specifically provide desired complementary systems for system combination further improved performance. The best results were submitted to the REVERB challenge workshop, and our results were the best among the challenge participants in the same category, which clarifies the effectiveness of our proposed approach.

Endnotes

[1] This allows for multiple decoding passes per utterance, such as for calculating the fMLLR matrix, but decodes each test utterance separately, without taking into account information from other test utterances, or speaker identities.

[2] In this case, GMM(f-bMMI) exhibited the best performance (26.25 % WER).

[3] Among all the participants, [49] was the best. This is a state-of-the-art system composed of a liner-prediction based dereverberation technique, DNN based acoustic modeling, and rescoring using RNN language model. The main difference from our system was the use of the "own dataset" that can compensate for the mismatches between training data and evaluation data (especially for the REALDATA) and improve the performance.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
YT developed a single-channel dereverberation method and discriminative training for complementary system, carried out whole experiments, and drafted the manuscript. TN developed multi-channel speech enhancement methods. SW developed a discriminative training for complementary system and black-box optimization method. All authors read and approved the final manuscript.

Acknowledgements
We appreciate that Mr. Felix Weninger, who belongs to the TU München and MERL, constructed the baseline ASR system.

Author details
[1] Information Technology R&D Center, Mitsubishi Electric, 5-1-1, Ofuna, Kamakura, Japan. [2] Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, US.

Received: 6 February 2015 Accepted: 15 June 2015
Published online: 30 June 2015

References

1. K Kinoshita, M Delcroix, T Yoshioka, T Nakatani, E Habets, R Haeb-Umbach, V Leutnant, A Sehr, W Kellermann, R Maas, S Gannot, B Raj, in *Proceedings of WASPAA*. The REVERB Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech (IEEE, 2013)
2. Y Tachioka, T Hanazawa, T Iwasaki, Dereverberation method with reverberation time estimation using floored ratio of spectral subtraction. Acoust. Sci. Technol. **34**(3), 212–215 (2013)
3. D Johnson, D Dudgeon, *Array Signal Processing*. (Prentice-Hall, New Jersey, 1993)
4. C Knapp, G Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acous. Speech, and Signal Process. **24**, 320–327 (1976)
5. Y Tachioka, T Narita, T Iwasaki, Direction of arrival estimation by cross-power spectrum phase analysis using prior distributions and voice activity detection information. Acoust. Sci. Technol. **33**, 68–71 (2012)
6. D Povey, P Woodland, in *Proceedings of ICASSP*. Minimum phone error and I-smoothing for improved discriminative training, vol. I (IEEE, 2002), pp. 105–108
7. E McDermott, T Hazen, J Le Roux, A Nakamura, S Katagiri, Discriminative training for large-vocabulary speech recognition using minimum classification error. IEEE Trans. Audio Speech Lang. Process. **15**, 203–223 (2007)
8. R Haeb-Umbach, H Ney, in *Proceedings of ICASSP*. Linear discriminant analysis for improved large vocabulary continuous speech recognition (IEEE, 1992), pp. 13–16
9. R Gopinath, in *Proceedings of ICASSP*. Maximum likelihood modeling with Gaussian distributions for classification (IEEE, 1998), pp. 661–664
10. M Gales, Semi-tied covariance matrices for hidden Markov models. IEEE Trans. Speech Audio Process. **7**, 272–281 (1999)
11. T Anastasakos, J McDonough, R Schwartz, J Makhoul, in *Proceedings of ICSLP*. A compact model for speaker-adaptive training (ISCA, 1996), pp. 1137–1140
12. D Povey, B Kingsbury, L Mangu, G Saon, H Soltau, G Zweig, in *Proceedings of ICASSP*. fMPE: Discriminatively trained features for speech recognition (IEEE, 2005), pp. 961–964
13. G Hinton, L Deng, D Yu, G Dahl, A Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, T Sainath, B Kingsbury, Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. **28**, 82–97 (2012)
14. E Vincent, J Barker, S Watanabe, Le Roux, J, F Nesta, M Matassoni, in *Proceedings of ICASSP*. The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines (IEEE, 2013), pp. 126–130
15. Y Tachioka, S Watanabe, J Hershey, in *Proceedings of ICASSP*. Effectiveness of discriminative training and feature transformation for reverberated and noisy speech (IEEE, 2013), pp. 6935–6939
16. Y Tachioka, S Watanabe, J Le Roux, J Hershey, in *Proceedings of the 2nd CHiME Workshop on Machine Listening in Multisource Environments*. Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark, (2013), pp. 19–24
17. H Christensen, J Barker, N Ma, P Green, in *Proceedings of INTERSPEECH*. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments (ISCA, 2010), pp. 1918–1921
18. G Saon, S Dharanipragada, D Povey, in *Proceedings of ICASSP*. Feature space Gaussianization, vol. I (IEEE, 2004), pp. 329–332

19. K Palomäki, H Kallasjoki, in *Proceedings of REVERB Workshop*. Reverberation robust speech recognition by matching distributions of spectrally and temporally decorrelated features, (2014)
20. D Povey, K Yao, A basis representation of constrained MLLR transforms for robust adaptation. Comput. Speech and Language. **26**, 35–51 (2012)
21. A Mohamed, G Hinton, G Penn, in *Proceedings of ICASSP*. Understanding how deep belief networks perform acoustic modelling (IEEE, 2012), pp. 4273–4276
22. J Fiscus, in *Proceedings of ASRU*. A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER) (IEEE, 1997), pp. 347–354
23. G Evermann, P Woodland, in *Proceedings of NIST Speech Transcription Workshop*. Posterior probability decoding, confidence estimation and system combination, (2000)
24. B Hoffmeister, T Klein, R Schlüter, H Ney, in *Proceedings of ICSLP*. Frame based system combination and a comparison with weighted ROVER and CNC (ISCA, 2006), pp. 537–540
25. F Diehl, P Woodland, in *Proceedings of INTERSPEECH*. Complementary phone error training (ISCA, 2012)
26. K Audhkhasi, A Zavou, P Georgiou, S Narayanan, Theoretical analysis of diversity in an ensemble of automatic speech recognition systems. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(3), 711–726 (2014)
27. Y Tachioka, S Watanabe, in *Proceedings of INTERSPEECH*. Discriminative training of acoustic models for system combination (ISCA, 2013), pp. 2355–2359
28. Y Tachioka, S Watanabe, J Le Roux, J Hershey, in *Proceedings of ASRU*. A generalized framework of discriminative training for system combination (IEEE, 2013), pp. 43–48
29. D Povey, L Burget, M Agarwal, P Akyazi, F Kai, A Ghoshal, O Glembek, N Goel, M Karafiát, A Rastrow, R Rose, P Schwarz, S Thomas, The subspace Gaussian mixture model –a structured model for speech recognition. Comput. Speech Lang. **25**(2), 404–439 (2011)
30. Y Tachioka, T Narita, S Watanabe, F Weninger, in *Proceedings of REVERB Challenge*. Dual system combination approach for various reverberant environments, (2014), pp. 1–8
31. T Suzuki, Y Kaneda, Sound source direction estimation based on subband peak-hold processing. J. Acoust. Soc. Japan. **65**(10), 513–522 (2009)
32. T Nishiura, T Yamada, T Nakamura, K Shikano, in *Proceedings of ICASSP*. Localization of multiple sound sources based on a CSP analysis with a microphone array, vol. 2 (IEEE, 2000), pp. 1053–1056
33. E Habets, in *Speech Dereverberation*, ed. by P Naylor, N Gaubitch. Speech dereverberation using statistical reverberation models (Springer London, 2010)
34. S Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acous. Speech Signal Process. **27**(2), 113–120 (1979)
35. AH Sayed, *Adaptive Filters*. (John Wiley & Sons, New Jersey, 2008)
36. D Povey, D Kanevsky, B Kingsbury, B Ramabhadran, G Saon, K Visweswariah, in *Proceedings of ICASSP*. Boosted MMI for model and feature-space discriminative training (IEEE, 2008), pp. 4057–4060
37. D Povey, in *Proceedings of INTERSPEECH*. Improvements to fMPE for discriminative training of features (ISCA, 2005), pp. 2977–2980
38. Veselý, A Ghoshal, L Burget, D Povey, in *Proceedings of INTERSPEECH*. Sequence-discriminative training of deep neural networks, (2013)
39. T Robinson, J Fransen, D Pye, J Foote, S Renals, in *Proceedings of ICASSP*. WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition (IEEE, 1995), pp. 81–84
40. D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, M Petr, Y Qian, P Schwarz, J Silovský, G Stemmer, K Veselý, in *Proceedings of ASRU*. The Kaldi speech recognition toolkit (IEEE, 2011), pp. 1–4
41. H Xu, D Povey, L Mangu, J Zhu, in *Proceedings of ICASSP*. An improved consensus-like method for minimum Bayes risk decoding and lattice combination (IEEE, 2010), pp. 4938–4941
42. J Snoek, H Larochelle, R Adams, in *Proceedings of Neural Information Processing Systems*. Practical bayesian optimization of machine learning algorithms, (2012)
43. G Dahl, T Sainath, G Hinton, in *Proceedings of ICASSP*. Improving deep neural networks for LVCSR using rectified linear units and dropout (IEEE, 2013), pp. 8609–8613
44. S Watanabe, J Le Roux, in *Proceedings of ICASSP*. Black box optimization for automatic speech recognition (IEEE, 2014), pp. 3280–3284
45. F Weninger, S Watanabe, Y Tachioka, B Schuller, in *Proceedings of ICASSP*. Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition (IEEE, 2014), pp. 4656–4660
46. F Weninger, S Watanabe, J Le Roux, J Hershey, Y Tachioka, JT Geiger, BW Schuller, G Rigoll, in *Proceedings of REVERB Challenge*. The MERL/MELCO/TUM system using deep recurrent neural network speech enhancement, (2014), pp. 1–8
47. X Xiao, Z Shengkui, DHH Nguyen, Z Xionghu, D Jones, E-S Chng, H Li, in *Proceedings of REVERB Challenge*. The NTU-ADSC systems for reverberation challenge 2014, (2014), pp. 1–8
48. MJ Alam, V Gupta, P Kenny, P Dumouchel, in *Proceedings of REVERB Challenge*. Use of multiple front-ends and i-vector-based speaker adaptation for robust speech recognition, (2014), pp. 1–8
49. M Delcroix, T Yoshioka, A Ogawa, Y Kubo, M Fujimoto, I Nobutaka, K Kinoshita, M Espi, T Hori, T Nakatani, A Nakamura, in *Proceedings of REVERB Challenge*. Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge, (2014), pp. 1–8