MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Dual system combination approach for various reverberant environments with dereverberation techniques

Tachioka, Y.; Narita, T.; Weninger, F.; Watanabe, S.

TR2014-032 May 2014

Abstract

The recently introduced REVERB challenge includes a reverberant speech recognition task. We focus on state-of-the-art ASR techniques such as discriminative training and various feature trans- formations including Gaussian mixture model, sub-space Gaussian mixture model, and deep neural networks, in addition to the pro- posed single channel dereverberation method with reverberation time estimation and multi-channel beamforming that enhances direct sound compared with the reflected sound. In addition, because the best performing system is different from environment to environment, we perform a system combination approach using different feature and different types of systems to handle these various environments in the challenge. Moreover, we use our discriminative training technique for system combination that improves system combination by making systems complementary. Experiments show the effectiveness of these approaches, reaching 6.76% and 18.60% word error rate on the REVERB simulated and real test sets, which are 68.8% and 61.5% relative improvements over the baseline.

IEEE REVERB Workshop

DUAL SYSTEM COMBINATION APPROACH FOR VARIOUS REVERBERANT ENVIRONMENTS WITH DEREVERBERATION TECHNIQUES

Yuuki Tachioka, Tomohiro Narita

Mitsubishi Electric Corporation
Information Technology R&D center
5-1-1, Ofuna, Kamakura, Kanagawa, Japan

Felix Weninger, Shinji Watanabe

Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, MA, USA

ABSTRACT

The recently introduced REVERB challenge includes a reverberant speech recognition task. We focus on state-of-the-art ASR techniques such as discriminative training and various feature transformations including Gaussian mixture model, sub-space Gaussian mixture model, and deep neural networks, in addition to the proposed single channel dereverberation method with reverberation time estimation and multi-channel beamforming that enhances direct sound compared with the reflected sound. In addition, because the best performing system is different from environment to environment, we perform a system combination approach using different feature and different types of systems to handle these various environments in the challenge. Moreover, we use our discriminative training technique for system combination that improves system combination by making systems complementary. Experiments show the effectiveness of these approaches, reaching 6.76% and 18.60% word error rate on the REVERB simulated and real test sets, which are 68.8% and 61.5% relative improvements over the baseline.

Index Terms— Reverberation, Dereverberation, Discriminative training, Feature transformation, System combination

1. INTRODUCTION

The REVERB challenge is a recently introduced task for reverberant speech processing [1]. This paper focuses on the speech recognition task, which provides a middle-size vocabulary continuous speech recognition task in order to evaluate the automatic speech recognition (ASR) performance under reverberant environments.

In this scenario, speech enhancement before ASR processing is important and affects the ASR performance. We have proposed a single-channel dereverberation method with estimation of reverberation time [2], which is the most important parameter for characterizing the extent of reverberation. In addition, in order to exploit the eight-channel data provided by the REVERB challenge, we use a beamforming approach [3] with direction of arrival estimation [4, 5].

Recently, ASR performance has been significantly improved owing to the discriminative training methods [6, 7] and various types of feature transformations [8, 9, 10, 11, 12, 13]. We have showed the effectiveness of discriminative training for noisy environments [14, 15]. In previous evaluation campaigns for noise-robust ASR such as the CHiME Challenge [16], it was necessary to handle many types of non-stationary additive noises, but the variety in the room acoustics (i.e., reverberation pattern) is very limited. However, it is well known that room reverberation degrades the ASR performance similarly to additive noise; thus, to address reverberation seems as important as to address noise. For matched conditions where training and evaluation conditions are close, discriminative training is effective in general, therefore, it is necessary to validate its effectiveness for different room types where training and evaluation conditions are different. The REVERB challenge [1] includes eight different reverberant environments: three rooms with near/far microphone settings for simulated data and one room with near/far microphone settings for real recorded data with stationary noise. Because of the variety in the evaluation environments and because of the mismatch between simulated training data and real test data, discriminative training may be ineffective. Thus, the first aim of this paper is to present the effectiveness of the state-of-the-art ASR techniques for varying reverberant and noisy environments.

In addition to discriminative training, this paper deals with several feature transformation approaches, which convert original features to new features based on linear transformations (Linear Discriminant Analysis (LDA) [8], Maximum Likelihood Linear Transformation (MLLT) [9, 10]), and discriminative non-linear feature transformation [12]. LDA uses long context by context expansion (e.g., contiguous nine frames) to exploit feature dynamics, which reduces the influence of reverberation. The effect of context size on the ASR performance will be validated. MLLT finds a linear transformation of features to reduce state-conditional feature correlations. To improve the recognition accuracy by adapting to unknown conditions, model adaptation is effective. In this paper, Speaker Adaptive Training (SAT) [11]) and basis feature-space Maximum Likelihood Linear Regression (basis fMLLR) are used.

This paper also deals with Deep Neural Networks (DNN) [13] that have recently attracted great attention. This modeling includes feature transformation and acoustic modeling, and can optimize both of them simultaneously [17]. We have shown the promising results for noisy environments [15] and validate its effectiveness for reverberant environments.

These studies above are mainly focused on the single ASR system. On the other hand, the use of multiple systems is another solution to improve the ASR performance [18, 19, 20]. In scenarios where the best performing system differs from environment to environment, combining their outputs can improve the performance. In this paper, as mentioned above, various systems are constructed and, in addition to this, we have proposed a discriminative training method in order to introduce complementarity of systems intentionally based on the lattice-based discriminative training framework [21, 22]. The results from various recognizers will be combined using recognizer output voting error reduction (ROVER) [18].

In summary, there are two objectives in this paper. First, we validate the effectiveness of feature transformation and discriminative training for reverberant environments including various types of acoustic modelings such as Gaussian Mixture Model (GMM), sub-

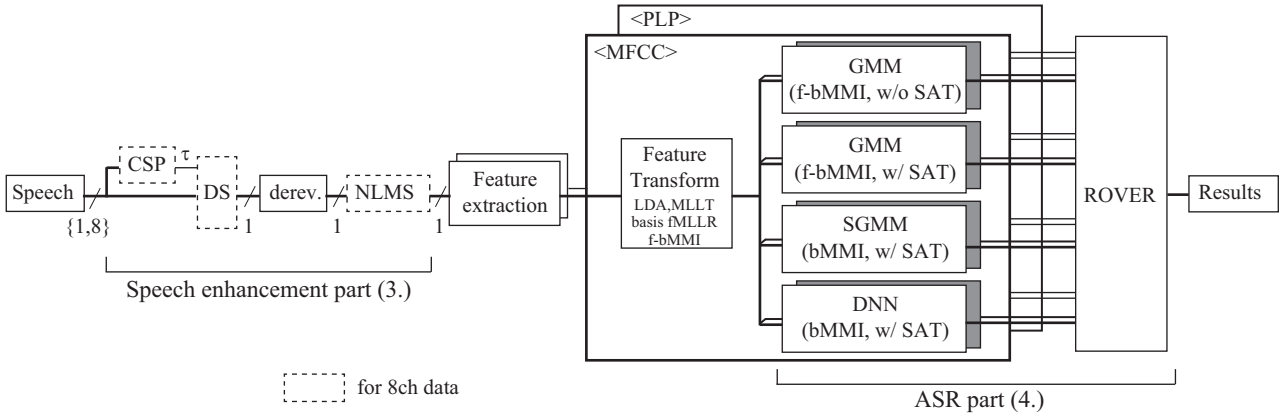

Fig. 1. Schematic diagram of the proposed system. (CSP: cross spectrum phase analysis, DS: delay-and-sum beamformer, derev.: proposed dereverberation method, NLMS: normalized least-mean-squares algorithm, gray blocks are complementary systems for each system type)

space Gaussian mixture model (SGMM), and DNN after speech enhancement and discriminative feature transformation. Second, to address the variety of reverberant environments, system combination approach is introduced. These systems are constructed using Kaldi toolkit [23].

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, described in Section 3. It consists of 1) a multi-channel delay-and-sum beamformer 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 eliminates late reverberation, and 3) the normalized least-mean squares (NLMS) algorithm that eliminates short-term distortions. 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 4.2). 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 expected 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 (boosted maximum mutual information (boosted MMI), presented in Sections 4.1 for GMM and SGMM and 4.3 for DNN). Three types of systems (GMM, SGMM, and DNN) are constructed. In addition to the speaker adaptive training (SAT) model, we also constructed a GMM model without SAT, because we verified on the development set that its outputs are different from those of a GMM with SAT. Moreover, we propose a dual system combination approach, which uses a pair of systems discriminatively trained by our proposed method (Section 4.4). Various system outputs are combined using ROVER.

3. SPEECH ENHANCEMENT PART

3.1. Delay-and-sum beamformer after the direction of arrival estimation using CSP method

To enhance the direct sound from the source, a frequency domain delay-and-sum beamformer is applied [3]. The enhanced spectrum $\tilde{\mathbf{y}}_t$ is obtained as a sum of the observed short-time Fourier transform (STFT) spectrum of the m-th microphone $\mathbf{x}_t(m)$ as

$$\tilde{\mathbf{y}}_t = \sum_m \mathbf{x}_t(m) \odot \exp(-\jmath\boldsymbol{\omega}\tau_{1,m}), \tag{1}$$

where t is the index of the current frame, $\odot$ is an element-wise multiplication, and $\boldsymbol{\omega}$ is a set of angular frequencies. The arrival time delay of the m-th microphone from the first microphone $\tau_{1,m}$ is related to the direction of arrival and is estimated by the cross-spectrum phase (CSP) analysis, which uses a cross-power spectrum between two microphones [4] as

$$\tau_{1,m} = \arg\max \mathcal{S}^{-1}\left[\frac{\mathbf{x}_t(1) \odot \mathbf{x}_t(m)^*}{|\mathbf{x}_t(1)||\mathbf{x}_t(m)|}\right], \tag{2}$$

where $\mathcal{S}$ is the STFT operation and * denotes the complex conjugate. To improve the performance of the original CSP method, we used a peak-hold process [24] and noise component suppression, which sets the cross power spectrum to zero when the estimated SNR is below 0 dB [5]. Using three or more microphones reduces noise influence by synchronously adding pair-wise CSP coefficients [25].

3.2. Single-channel dereverberation with estimation of reverberation time

As a single-channel dereverberation method, we employ the algorithm proposed in [2]. When the reverberation time T_r is much longer than the frame size, an observed power spectrum $|\mathbf{x}|^2$ is modeled as a weighted sum of the source's power spectrum $|\hat{\mathbf{y}}|^2$ to be estimated with a stationary noise power spectrum $|\mathbf{n}|^2$ as

$$|\mathbf{x}_t|^2 = \sum_{\mu=0}^{t} w_\mu |\hat{\mathbf{y}}_{t-\mu}|^2 + |\mathbf{n}|^2, \tag{3}$$

where μ and w are the delay frame and the weight coefficient, respectively. The source's power spectrum $|\hat{\mathbf{y}}|^2$ is related to $|\mathbf{x}|^2$ as

$$|\hat{\mathbf{y}}_{t-\mu}|^2 = \eta(T_r)|\mathbf{x}_{t-\mu}|^2 - |\mathbf{n}|^2, \tag{4}$$

where η is the ratio of direct sound components to the sum of the direct and reflected sound components, which is a decreasing function

of T_r because for longer T_r, the energy of the reflected sound components increases. Assuming that w_0 is unity, Eq. (5) can be derived from above relations:

$$|\hat{\mathbf{y}}_t|^2 = |\mathbf{x}_t|^2 - \sum_{\mu=1}^{t} w_\mu \left[\eta(T_r)|\mathbf{x}_{t-\mu}|^2 - |\mathbf{n}|^2\right] - |\mathbf{n}|^2. \quad (5)$$

Reverberation is divided into two stages: early reverberation and late reverberation. The threshold between them, after arrival of the direct sound, is denoted by D (in frames). Early reverberation is complex but can be ignored because the ASR performance is mainly degraded by the late reverberation. The proposed method focuses on the late reverberation where the sound-energy density decays exponentially with time according to Polack's statistical model [26]. Hence, w is determined as

$$w_\mu = \begin{cases} 0 & (1 \le \mu \le D), \\ \frac{\alpha_s}{\eta(T_r)} e^{-2\Delta\varphi\mu} & (D < \mu), \end{cases} \quad (6)$$

where φ is the frame shift and α_s is the subtraction parameter to be set. The upper condition and lower condition are corresponding to early and late reverberation, respectively. Assuming η is constant, Eq. (5) is a process similar to spectral subtraction [27]. If the subtracted power spectrum $|\hat{\mathbf{y}}|^2$ is less than $\beta|\mathbf{x}|^2$, it is substituted by $\beta|\mathbf{x}|^2$, where β is a flooring parameter. We define the floored ratio r as the ratio of the number of floored time-frequency bins to the total number of bins.

Two observations are exploited to estimate T_r from floored ratios r. First, when some arbitrary reverberation times (T_a) are assumed, r increases monotonically with T_a. This is modeled as linear with the inclination Δ_r. Second, r increases with T_r at the same T_a. Since the actual $\eta(T_r)$ decreases with T_r, the power spectrum after dereverberation assuming constant η is more likely to be floored for a longer T_r because the second term of Eqn. (5) is larger than that of actual one in the condition with longer T_r. Therefore, T_r has a positive correlation with Δ_r and this can be modeled as $T_r = a\Delta_r - b$ with two constants a and b. The estimation process of T_r is summarized as follows: Calculate r and the inclination Δ_r by least-squares regression for some values of T_a, and estimate T_r.

4. ASR PART

4.1. 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 boosted MMI (bMMI) [28], where a boosting factor $b \ge 0$ is used to introduce a weight depending on phoneme accuracies. The objective function is given as

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

where $\mathbf{x}^r$ is the r-th utterance's feature sequence. The acoustic model parameters λ are optimized by the extended Baum-Welch algorithm. $\mathcal{H}_{s_r}$ 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, and p_L is the language model likelihood; $A(s, s_r)$ is the phoneme accuracy of s for s_r. In this paper, we compare the performances of bMMI training of GMM and SGMM to those of Maximum Likelihood (ML) training.

4.2. Discriminative feature transforms

The extension of discriminative training to feature transformation is referred to as feature-space discriminative training [12]. It estimates an $I \times J$ matrix $\mathbf{M}$ that projects rich high-dimensional features down to low-dimensional transformed features, as follows:

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{M}\mathbf{h}_t, \quad (8)$$

where $\mathbf{x}_t$ are the original I-dimensional features at frame t, $\mathbf{y}_t$ are the transformed I-dimensional features, and $\mathbf{h}_t$ are J-dimensional auxiliary features with $J \gg I$. Usually, Gaussian posteriors of universal background model is used for $\mathbf{h}_t$. The matrices $\mathbf{M}$ are optimized by maximizing the objective function $\mathcal{F}_{\text{f-bMMI}}(\mathbf{M})$, which can be obtained by simply replacing $\mathbf{x}^r$ by the r-th utterance's transformed feature sequence $\mathbf{y}^r$ in Eq. (7):

$$\mathcal{F}_{\text{f-bMMI}}(\mathbf{M}) = \sum_r \log \frac{p_\lambda\left(\mathbf{y}^r|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda\left(\mathbf{y}^r|\mathcal{H}_s\right)^\kappa p_L(s)\mathrm{e}^{-bA(s,s_r)}}. \quad (9)$$

In this study, we validate the effectiveness of feature-space boosted MMI (f-bMMI).

4.3. Discriminative training of DNN

In a DNN-HMM hybrid system, sequential discriminative training according to the (b)MMI criterion (7) has been proposed [29] in addition to 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

$$p_\theta\left(\mathbf{x}^r|j\right) = \frac{p_\theta\left(j|\mathbf{x}^r\right)}{p_0\left(j\right)}, \quad (10)$$

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

$$p_\theta(j|\mathbf{x}^r) = \frac{\exp a_j(\mathbf{x}^r)}{\sum_{j'} \exp a_{j'}(\mathbf{x}^r)}, \quad (11)$$

where a_j is the activation of the j-th unit in the output layer. These activations are trained discriminatively according to the bMMI criterion. The bMMI objective function is the same as Eq. (7), simply replacing λ by θ: $\mathcal{F}_{\mathrm{bMMI}}(\theta)$.

4.4. A generalized framework for constructing complementary system for dual system combination

We describe a discriminative method that constructs complementary systems for appropriate system combination [21, 22]. Complementary systems are constructed by discriminatively training a model starting from an initial model. The proposed discriminative training method for complementary systems is extended from a discriminative training principle. Assuming Q base systems have already been constructed, the discriminative training objective function $\mathcal{F}$ is generalized to the following proposed objective function $\mathcal{F}^c$, which subtracts from the original objective function involving the correct labels s_r, the objective functions involving the 1-best hypotheses (lattice) $s_{q,1}$ of the q-th base systems:

$$\mathcal{F}^c(\varphi) = (1 + \alpha_c)\mathcal{F}(\varphi) - \frac{\alpha_c}{Q}\sum_{q=1}^{Q}\mathcal{F}(\varphi), \quad (12)$$

where φ is the set of model parameters of a complementary system to be optimized (that is λ, $\mathbf{M}$, and θ) and α_c is a scaling factor. 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. (12) promotes good performance according to the discriminative training criterion, whereas the second term makes the target system generate hypotheses that have a different tendency from the original base models. This procedure is commonly used to obtain the objective functions of Sections 4.1, 4.2, and 4.3.

5. EXPERIMENTS

5.1. Task description

We validated the effectiveness of our proposed approaches for reverberated speech on the REVERB challenge [1] data, which features a medium-vocabulary task in reverberant environments, whose utterances are taken from the Wall Street Journal database (WSJ0). 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 rooms (Room 1, 2, and 3), and "REALDATA" created by recording real-world speech at a distance of 0.5 m (near) or 2 m (far) from the microphones in one room (Room 1) with stationary noise such as air conditioner noise. Eight microphones were arranged on the circle whose radius is 0.1 m. The training data set (**tr**) contains 7,861 utterances by 92 speakers, the evaluation data set (**eva**) contains 2,176 utterances by 28 speakers for SIMDATA and 372 utterances by 10 speakers for REALDATA, and the development set (**dev**) contains 1,484 utterances by 20 speakers for SIMDATA and 179 utterances by five speakers for REALDATA. Acoustic models were trained using **tr** and 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 tri-gram language model is used. All experiments in this paper were "utterance-based batch processing", i.e., allowing multiple decoding passes per utterance (such as for calculating the fMLLR matrix), but decoding each test utterance separately, without taking into account information from other test utterances, or speaker identities.

5.2. Speech enhancement

The REVERB challenge provides one, two, and eight channel data. We used one and eight channel data. For single channel data, the proposed dereverberation technique with reverberation time estimation was used. Parameters for dereverberation technique were set as $D = 9$, $\alpha = 5$, $\beta = 0.05$, $a = 0.005$, and $b = 0.6$. For eight channel data, before dereverberation, delay-and-sum beamforming with direction of arrival estimation was performed. After dereverberation, the NLMS algorithm with 200 taps was used to eliminate short-term distortions. Totally ${}_8C_2 (= 28)$ pairs of microphones were used for beamforming and direction of arrival estimation.

5.3. Feature extraction and transformation and acoustic model adaptation

We describe the settings of acoustic feature and feature transformation, which are detailed in [14, 15]. The baseline acoustic features are MFCC and PLP (0-12 order MFCCs/PLPs + Δ + $\Delta\Delta$). After concatenating static MFCCs in nine contiguous frames, a total of 117-dimensional features are compressed into 40 dimensions by an LDA with classes corresponding to tri-phone HMM states (2,500 states). In addition to this, to decrease correlations between features, MLLT is used.

For acoustic model adaptation, instead of ordinary fMLLR transformation, which did not improve the performance because environments are various even for the same speaker and using speaker IDs at test time is prohibited, we used basis fMLLR [30], which can perform adaptation in short utterances. Moreover, to address the large variations among speakers, SAT [11] is typically used: in SAT, acoustic model training is conducted after having transformed the training speech into a canonical space so as to reduce the variances across speakers. Note that for the training set, speaker IDs are assumed to be known. In this study, we validate the effectiveness of feature transformation (LDA, MLLT) and speaker adaptation (basis fMLLR and SAT).

5.4. Discriminative methods

In discriminative feature transformation (Section 4.2), 400 Gaussians were used and offset features were calculated for each of the 40 MFCC dimensions with context expansion (9 frames). The number of dimensions of the feature vector $\mathbf{h}_t$ was $400 \times 40 \times 9$. Features with the top 2 GMM posteriors were selected and all other features were ignored. The Boosting factor of bMMI and f-bMMI was 0.1. For constructing complementary systems, the additional boosting factor was 0.3 and α_c is 0.75.

5.5. Experimental procedure

The experimental procedure based on the above setup can be summarized as follows: First, a clean acoustic model was trained. The number of mono-phones was 44, including silence ("sil"). In the triphone model, the number of states was 2,500 and the total number of Gaussians was 15,000. Second, using their alignments and tri-phone tree structures, reverberated acoustic models were trained using the reverberated dataset. Finally, using this ML model, we validated the effectiveness of the discriminative training and feature transformation. For the DNN, we used a CPU version of neural network training implemented in Kaldi [23] with 2 hidden layers and 2,000,000 parameters. The initial learning rate of cross-entropy training was 0.02 and was decreased to 0.004 at the end of training. The training targets for the DNN were determined by forced alignment 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 were modified.

5.6. ASR acoustic models

We prepared three types of ASR acoustic model systems for the challenge: GMM, SGMM [31], and DNN. To improve the performance of respective systems, discriminative methods were used. Feature-space boosted MMI was employed for GMM systems, and boosted MMI for SGMMs and DNNs [29]. Two systems were constructed for each of these systems. Moreover, these systems were trained both for MFCC and PLP features; thus, totally sixteen systems were prepared. Minimum Bayes risk decoding [32], which slightly improved the performance, was commonly used after decoding.

5.7. Black box optimization

Bayesian optimization using Gaussian processes [33] was applied to various speech recognition problems including neural network [34] and HMM topology optimization [35]. In this paper, we also apply this technique to the selection of combined systems and the parameter optimization for ROVER. The objective function of the optimization was WER.

Table 1. WER [%] by room and microphone distance on the REVERB Challenge (**dev**) using single channel data. (MFCC)

			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

Table 2. WER [%] on the REVERB Challenge (**dev**) using eight channel data. (MFCC)

			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

6. RESULTS AND DISCUSSION

6.1. Baseline and speech enhancement techniques

Tables 1 and 2 show the WERs on the development set (**dev**) for various rooms and distances between microphones and speakers. Table 1 is based on single channel one and Table 2 is based on 8 channel one. "Kaldi baseline" in Table 1 shows the WER using acoustic model trained on reverberant speech without speech enhancement. "derev." is a proposed dereverberation method with reverberation time estimation. Although, for some cases in room 1 whose reverberation time is fairly short, the performances were degraded. However, for other cases and in average, the performances were improved. For eight channel input (Table 2), because the direction of arrival estimation was stable, beamforming with "derev." improved the performance significantly. "NLMS" improved the WER by 2.04% on REALDATA, but degraded the WER by 0.63% for SIMDATA. However, because this decrease in performance was fairly low, we used NLMS below.

The results above used MFCC features. Experimental results using PLP features are shown in Table 3. In average, the ASR performance using PLP features was slightly lower than that using MFCC features; however, their error tendencies were fairly different, which was a good property for system combination.

6.2. Feature transformation and discriminative training

LDA and MLLT feature transformations significantly improved the performance. Table 4 shows the effect of LDA context size on the performance. Performance on SIMDATA could not be increased by

Table 3. WER [%] on the REVERB Challenge (**dev**). (PLP)

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

context sizes longer than 4. On REALDATA, performance could be improved in several cases by adding more right context, but generally not by left context. In the end, because these results are rather mixed, we kept the context size at the default setting, $L = R = 4$.

Tables 1 and 2 show that discriminative training was effective for reverberant environments. The performances of f-bMMI training were higher than those of bMMI training in all cases. Because the performances of our complementary systems are only slightly lower than those of the base systems, they appear to be very well suited to system combination.

Table 5 shows the effect of the iteration numbers of bMMI and f-bMMI on the development set performance. For f-bMMI, in one iteration, f-bMMI for the matrix $\mathbf{M}$ was coupled with bMMI for the acoustic models. Results shows that best performance is achieved at four iterations.

6.3. Subspace GMM and deep neural network

Tables 1 and 2 show that the performance of SGMM acoustic models for SIMDATA was higher than that of GMMs, however, for REALDATA the performance was lower than that of GMMs. Because

Table 4. Effect of LDA context size (left (L) and right (R)) on the REVERB Challenge (**dev**) using eight channel data.

	SIMDATA				REALDATA			
L \ 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

Table 5. Effect of iteration number of discriminative training (bMMI and f-bMMI with SAT) on the REVERB Challenge (**dev**) using eight channel data.

bMMI

	MFCC				PLP			
iter	1	2	3	4	1	2	3	4
SIM	8.70	8.41	8.18	**7.63**	9.02	8.64	8.47	**8.06**
REAL	29.21	28.34	**28.16**	28.71	29.74	29.26	28.91	**28.45**

f-bMMI

	MFCC				PLP			
iter	1	2	3	4	1	2	3	4
SIM	8.07	7.56	7.30	**7.01**	8.47	7.93	7.57	**7.32**
REAL	27.70	27.29	27.16	**26.94**	29.36	27.86	27.15	**26.78**

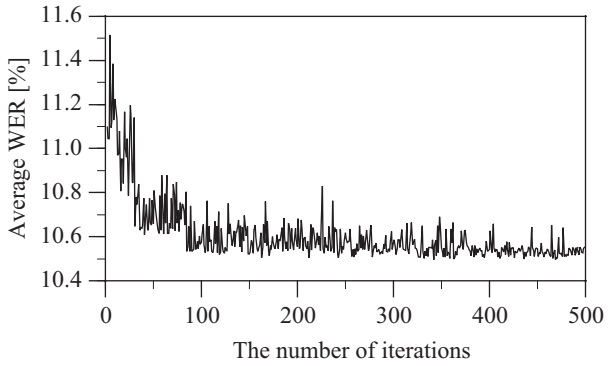

Fig. 2. Average WER [%] of black box optimization of the system selection and parameter setting for ROVER in terms of the number of iterations.

REALDATA were noisier than SIMDATA, the estimation of speaker vector can be unstable.

DNN acoustic models achieved the best performance for SIMDATA. Although the best system for REALDATA was GMM without SAT, DNN was the second best. On average over SIMDATA and REALDATA, DNNs achieved the best performance. Discriminative training for DNN systems turned out to be as effective as for other systems.

6.4. System combination

We tried five types of system combinations as shown in Table 6. The ID 1) system was a combination of SAT-GMMs (f-bMMI) using both MFCC and PLP features. The performance for REALDATA was improved by 4.15% (1ch) and 1.15% (8ch) over the f-bMMI with SAT (MFCC) single system. With the GMM system without SAT, using f-bMMI (ID 2)), the WER was improved by 1.49% and 0.6% (1ch) and 0.19% and 1.36% (8ch) for SIMDATA and REALDATA, respec-

Table 6. WER [%] on the REVERB Challenge (**dev**) with system combination using both MFCC and PLP features. 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 system along with their complementary systems. ROVER 6) uses black box optimization at the stage of the system selection and parameter optimization for ROVER.

		Number of systems				SIMDATA							REALDATA		
						Room 1		Room 2		Room 3		Avg	Room 1		Avg
	ID	GMM	SAT-GMM	SGMM	DNN	near	far	near	far	near	far		near	far	
1ch	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**
8ch	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*

Table 7. WER [%] on the REVERB Challenge (**eva**). All systems except ROVER are single systems. MFCC feature was used for single system and MFCC and PLP features were used for ROVER 5).

		SIMDATA							REALDATA		
		Room 1		Room 2		Room 3		Avg	Room 1		Avg
		near	far	near	far	near	far		near	far	
1ch	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
	f-bMMI	7.27	8.17	8.82	14.11	10.54	18.76	11.28	28.65	29.54	29.10
	SAT+f-bMMI	6.44	7.22	7.57	13.97	9.52	18.44	10.53	28.87	29.78	29.33
	SGMM+bMMI	5.81	6.54	7.22	13.84	8.70	18.17	10.05	27.75	28.36	28.06
	DNN+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**
8ch	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
	f-bMMI	6.57	6.93	6.80	9.93	7.47	12.76	8.41	20.22	23.19	21.71
	SAT+f-bMMI	6.17	6.64	6.51	10.13	7.40	13.15	8.33	20.63	23.67	22.15
	SGMM+bMMI	5.86	6.44	6.29	9.23	6.96	12.83	7.94	20.66	23.50	22.08
	DNN+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

tively. Including the complementary systems (ID 3)), the WER was slightly improved. For the best case, WER was improved by 0.40%, while for the worst case, WER was decreased by 0.07%. This shows the effectiveness of our proposed method. Adding in SGMMs (ID 4)), which was effective for SIMDATA, the performance for SIMDATA was further improved by 0.28% (1ch) and 0.38% (8ch). Taking into account DNNs (ID 5)), the performance was again improved and this system achieved the best performance in average.

In all cases except Room 1/far(8ch) condition, the performances were better than those of the best system.

6.5. Black box optimization

For eight channel data, black box optimization was performed. The results were shown at the last column of Table 6. The performance was improved mainly for REALDATA. Figure 2 shows the average WER by the iteration number. WER almost decreased monotonically and, after 100 iterations, it converged.

6.6. Evaluation set

Table 7 shows the results for the evaluation set (**eva**). The DNN with discriminative training achieved the best performance for SIMDATA and REALDATA among single systems. This shows the robustness of DNN in unseen conditions. Moreover, system combination (ROVER 5)) improved the WER by 1.26% and 2.13% (1ch) and 1.03% and 2.22% (8ch) for SIMDATA and REALDATA, respectively. Among system combination systems, the performance of ROVER 5) was better than that of ROVER 6), which could be due to overtuning on the development set.

7. CONCLUSIONS

We validated the effectiveness of feature transformations and discriminative training for reverberated speech. Experiments show that these state-of-the-art techniques are effective across various types of reverberation as well as for noisy environments. Moreover, the system combination approach was used in order to improve the robust-

ness in various environments. We constructed multiple systems, because the best performing system was different from environment to environment. System combination improved the performance in almost all the cases even from the best single system for each environment. Our proposed method to specifically provide desired complementary systems for system combination improved the performance further.

8. 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, and B. Raj, "The REVERB Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. of WASPAA*, 2013.

[2] Y. Tachioka, T. Hanazawa, and T. Iwasaki, "Dereverberation method with reverberation time estimation using floored ratio of spectral subtraction," *Acoustical Science and Technology*, vol. 34, no. 3, pp. 212–215, 2013.

[3] D.H. Johnson and D.E. Dudgeon, *Array Signal Processing*, Prentice-Hall, 1993.

[4] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 8 1976.

[5] Y. Tachioka, T. Narita, and T. Iwasaki, "Direction of arrival estimation by cross-power spectrum phase analysis using prior distributions and voice activity detection information," *Acoustical Science and Technology*, vol. 33, pp. 68–71, 1 2012.

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

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

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

[9] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. of ICASSP*, 1998, pp. 661–664.

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

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

[12] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. of ICASSP*, 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, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 28, pp. 82–97, 11 2012.

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

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

[16] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: an overview of challenge systems and outcomes," in *Proc. of ASRU*, 2013.

[17] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. of ICASSP*, 2012, pp. 4273–4276.

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

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

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

[21] Y. Tachioka and S. Watanabe, "Discriminative training of acoustic models for system combination," in *Proc. of INTERSPEECH*, 2013.

[22] Y. Tachioka, S. Watanabe, J. Le Roux, and J.R. Hershey, "A generalized framework of discriminative training for system combination," in *Proc. of ASRU*, 2013.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, 2011.

[24] T. Suzuki and Y. Kaneda, "Sound source direction estimation based on subband peak-hold processing," *The journal of the Acoustical Society of Japan*, vol. 65, no. 10, pp. 513–522, 10 2009.

[25] T. Nishiura, T. Yamada, T. Nakamura, and K. Shikano, "Localization of multiple sound sources based on a CSP analysis with a microphone array," in *Proc. of ICASSP*, 2000, vol. 2, pp. 1053–1056.

[26] E.A.P. Habets, "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation*, P.A. Naylor and N.D. Gaubitch, Eds. Springer, 2010.

[27] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 4 1979.

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

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

[30] D. Povey and K. Yao, "A basis representation of constrained MLLR transforms for robust adaptation," *Computer Speech and Language*, vol. 26, pp. 35–51, 2012.

[31] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R.C. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model –A structured model for speech recognition," *Computer Speech and Language*, vol. 25, no. 2, pp. 404–439, 4 2011.

[32] H. Xu, D. Povey, L. Mangu, and J. Zhu, "An improved consensus-like method for minimum Bayes risk decoding and lattice combination," in *Proc. of ICASSP*, 2010, pp. 4938–4941.

[33] J. Snoek, H. Larochelle, and R.P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Neural Information Processing Systems*, 2012.

[34] G.E. Dahl, T.N. Sainath, and G.E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. of ICASSP*, 2013, pp. 8609–8613.

[35] S. Watanabe and J. Le Roux, "Black box optimization for automatic speech recognition," in *Proc. of ICASSP*, 2014, submitted.