DISCRIMINATIVE METHODS FOR NOISE ROBUST SPEECH RECOGNITION: A CHIME CHALLENGE BENCHMARK

Yuuki Tachioka

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

Shinji Watanabe, Jonathan Le Roux, John R. Hershey

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

ABSTRACT

The recently introduced second CHiME challenge is a difficult two-microphone speech recognition task with non-stationary interference. Current approaches in the source-separation community have focused on the front-end problem of estimating the clean signal given the noisy signals. Here we pursue a different approach, focusing on state-of-the-art ASR techniques such as discriminative training and various feature transformations, in addition to simple noise suppression methods based on prior-based binary masking with estimated angle of arrival. In addition, we propose an augmented discriminative feature transformation that can introduce arbitrary features to a discriminative feature transform, an efficient combination method of Discriminative Language Modeling (DLM) and Minimum Bayes Risk (MBR) decoding in an ASR post-processing stage, and preliminarily investigate the effectiveness of deep neural networks for reverberated and noisy speech recognition. Using these techniques we present a benchmark on the middle-vocabulary subtask of CHiME challenge, showing their effectiveness for this task. Promising results were also obtained for the proposed augmented feature transformation and combination of DLM and MBR decoding. A part of the training code has been released as an advanced ASR baseline, using the Kaldi speech recognition toolkit.

***Index Terms*—** CHiME challenge, Noise robust ASR, Discriminative methods, Feature transformation, Prior-based binary masking

1. INTRODUCTION

The 2nd CHiME challenge is a recently introduced task for noise-robust speech processing [1]. The scenario involves recognizing speech from a single target speaker binaurally recorded in a domestic environment. Unlike the 1st CHiME challenge, the second edition contains a medium vocabulary task in which the speech is taken from the Wall Street Journal (WSJ0) 5k vocabulary read speech corpus, and convolved with binaural room impulse responses before mixing with binaural recordings of a noisy domestic environment. This task is much more difficult from a speech recognition point of view.

Whereas, in the 1st CHiME challenge, participants have focused more on source separation approaches, here we focus on state-of-the-art ASR techniques such as discriminative training and various feature transformations, using only simple noise suppression methods based on estimated time difference of arrival (TDOA) in the front-end. The goal is to understand how much can be gained from the discriminative training ASR approach, as well as to improve the baseline recognition systems used to test source-separation-based approaches, in order to allow researchers who may not be experts in ASR to better evaluate the benefit of these methods.

Recent advances in Automatic Speech Recognition (ASR) [2], have greatly improved the accuracy of speech recognition systems. Over the past ten years model training techniques have migrated from Maximum Likelihood (ML) estimation to discriminative training [3, 4]. In addition, various types of feature transformations have been proposed and showed effectiveness [5, 6, 7, 8, 9, 10]. Although it is well known that the state-of-the-art ASR techniques are very effective in clean speech conditions, we need further investigation of their effectiveness in challenging conditions such as environmental reverberation and noise. In this paper we focus on discriminative training and feature transformations for the 2nd CHiME challenge. This paper deals with several feature transformation approaches, which convert original features to new features based on linear transformations (Linear Discriminant Analysis (LDA) [5], Maximum Likelihood Linear Transformation (MLLT) [6, 7], Speaker Adaptive Training (SAT) [8]), and discriminative non-linear feature transformation [9]. LDA uses long context by context expansion (e.g., contiguous 9 frames) to exploit feature dynamics, which reduces the influence of non-stationary noises. MLLT finds a linear transformation of features to reduce state-conditional feature correlations. SAT and feature-space Maximum Likelihood Linear Regression (fMLLR) improve the recognition accuracy by adapting to unknown and changing noise conditions.

Discriminative non-linear feature transformations can provide yet further gains in performance, because the transformation is optimized to reduce the error rate in the context of the decoder (e.g., [11]). Some of the popular non-linear transforms provide an approximately piece-wise linear transform by the inclusion of "region-based" features based on Gaussian posterior probabilities. We propose to extend this basic approach by augmenting the set of region-based features to include additional non-linear features that may be relevant in noisy conditions. We call this method augmented discriminative feature transformation. As an alternative discriminative non-linear feature transformation, this paper also preliminarily investigates the effectiveness of Deep Neural Networks (DNN) [10].

In addition to testing the above methods in isolation, we consider some minimal signal processing in the front end to take advantage of the binaural nature of the recordings. The method forms a masking function using the discrepancy between the instantaneous inter-microphone phase difference and the expected phase difference for the target speaker location. In our ASR post processing step, we deal with an N-best re-ranking technique based on Discriminative Language Modeling (DLM) [12, 13, 14], and Minimum Bayes Risk (MBR) decoding [15, 16, 17][1]. We propose an efficient combination method of DLM and MBR decoding, which further improves ASR

[1] Note that [17] performs DLM with the MBR criterion, while we combine DLM and MBR *decoding* in a cascade form.

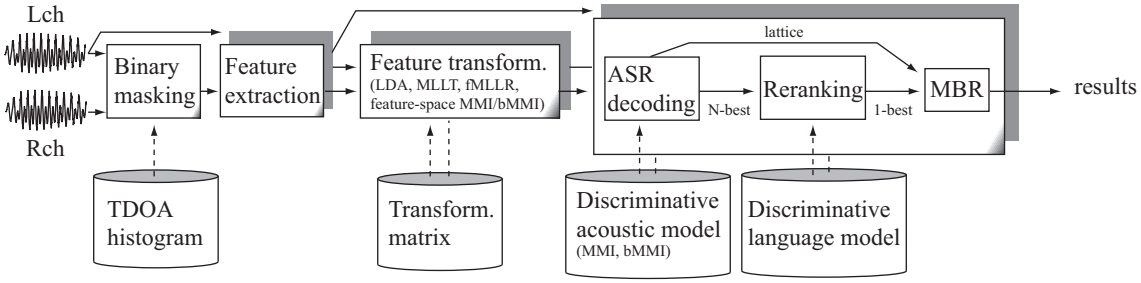

Fig. 1. Schematic diagram of the proposed system.

performance.

In summary, the goal of this paper is to evaluate the effectiveness of the discriminative training and feature transformation for reverberated and noisy speech for 2nd CHiME challenge Track 2. In addition we aim to build a CHiME challenge benchmark using public tools that can be shared with the community. We use a Kaldi toolkit [18] to provide an advanced ASR back-end and compare with the default HTK [19] based ML baseline associated with the CHiME challenge. In addition we also experiment with augmented discriminative feature transformations, combination of DLM and MBR decoding, and angle-of-arrival-based processing, all of which show some promising improvements to recognition performance.

2. SYSTEM OVERVIEW

Fig. 1 shows a schematic diagram of the proposed system, which consists of three components. First is a noise suppression step, described in Section 3). It consists in a prior-based binary masking, which suppresses directional interferences. Second is a feature transformation step, including feature-level transformation (LDA, MLLT, fMLLR) and discriminative feature transformation (feature-space techniques, presented in Section 4.2)[20]. Third is a decoding step. ASR decoding uses a discriminative acoustic model with margin (MMI and boosted MMI, presented in Section 4.1). Results are re-ranked using discriminative language model in Section 4.3 and minimum Bayes risk decoding (MBR) is performed based on lattice using re-ranked 1-best as a reference in Section 4.4.

3. PRIOR-BASED BINARY MASKING

In the CHiME challenge, two-channel recordings are provided and the target speaker is assumed to be in a frontal position with respect to the microphones. As binary masking based on the TDOA has been shown [21] to be more effective than beamforming for speech recognition with a small number of microphones, we investigate here its usage in our system. In the frontal position setting, without reverberation, the TDOA for signals coming from the target speaker should be equal to zero. Hence, time-frequency bins for which the inter-microphone phase difference is not close to zero are less likely to contain energy of the target speaker. However, with reverberation, the phase differences for a source from a frontal position may not be zero. Fig. 2 shows the phase difference histograms for 250 Hz and 1 kHz in "reverberated" speech. For 250 Hz, the histogram is almost symmetrical and variance is small but for 1 kHz, the mean is drifted and variance is large. The extent to which the phase difference is affected by noises and reverberation is significantly different for each frequency. Thus, a simple binary mask only using physical information will not work, and indeed, preliminary experiments

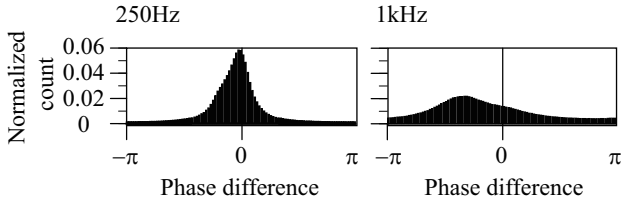

Fig. 2. Histogram of phase differences for two frequency bins.

showed worse word error rate (WER) than the baseline. As in [22], a statistical model is needed. To consider the offset and variance of the phase difference from the anechoic value, a prior-based binary masking is proposed. The phase difference $\theta_{\omega,t}$ at frequency bin ω and time frame t is calculated for each time-frequency bin as

$$X^L_{\omega,t}/X^R_{\omega,t} = A_{\omega,t}\mathrm{e}^{\jmath\theta_{\omega,t}}, \tag{1}$$

where $\jmath$ is the imaginary unit, $A_{\omega,t}$ is a positive real number, and X^L and X^R are the short-time Fourier spectrum for the left and right channels, respectively. In classical binary masking, a mask W is designed by using the following thresholding:

$$W_{\omega,t} = \begin{cases} \epsilon & \text{if } |\theta_{\omega,t}| > \theta_c, \\ 1 & \text{if } |\theta_{\omega,t}| \leq \theta_c, \end{cases} \tag{2}$$

where ϵ is a very small constant and θ_c is a threshold determined in advance. In our prior-based binary masking, the mask W' is determined using a frequency-dependent prior q_ω, here obtained from a phase difference histogram, as

$$W'_{\omega,t} = \begin{cases} \epsilon & \text{if } q_\omega(\theta_{\omega,t})/\bar{q}_\omega < q_c, \\ (q_\omega(\theta_{\omega,t})/\bar{q}_\omega)^\alpha & \text{if } q_\omega(\theta_{\omega,t})/\bar{q}_\omega \geq q_c, \end{cases} \tag{3}$$

where $\bar{q}_\omega = \max_\theta q_\omega(\theta)$, q_c is a threshold probability, and α is a warping coefficient.

4. BACK-END PROCESSING BASED ON DISCRIMINATIVE METHODS

4.1. Discriminative training

Discriminative training is a supervised training algorithm that minimizes the Bayes risk of posteriors for correct labeling and recognition results. This paper uses boosted MMI (bMMI) [23], where a boosting factor $b \geq 0$ is used to introduce a weight depending on phoneme accuracies. The objective function is given as

$$\mathcal{F}_{\mathrm{bMMI}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda\left(\{\mathbf{x}_t\}_r|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda\left(\{\mathbf{x}_t\}_r|\mathcal{H}_s\right)^\kappa p_L(s)\mathrm{e}^{-bA(s,s_r)}}, \tag{4}$$

where R is the number of training utterances and $\{\mathbf{x}_t\}_r$ is the r^{th} utterance's feature sequence. The acoustic model parameters λ are optimized by the extended Baum-Welch. $\mathcal{H}_{s_r}$ and $\mathcal{H}_s$ are the HMM sequences of a correct label s_r and a recognition result s, respectively; p_λ is the acoustic model likelihood, κ is the acoustic scale, and p_L is the language model likelihood; $A(s, s_r)$ is the phoneme accuracy of s for a reference s_r. In this paper, we compare the performances of MMI (corresponding to $b = 0$) and bMMI to that of ML.

4.2. Discriminative feature transforms with augmented features

In addition to discriminative training, feature transformation based on the discriminative training criterion can be used. This method is referred to as feature-space discriminative training [9]. It estimates a matrix $\mathbf{M}$ that projects rich high-dimensional features down to low-dimensional transformed features, as follows:

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{M}\mathbf{h}_t, \tag{5}$$

where $\mathbf{x}_t$ are the original I-dimensional features, $\mathbf{h}_t$ are J-dimensional features which depend on $\mathbf{x}$, with $J \gg I$, and $\mathbf{y}_t$ are the transformed features, and $\mathbf{M}$ is an $I \times J$ matrix. In this study, we validate the effectiveness of feature-space MMI (f-MMI) and its extension, feature-space boosted MMI (f-bMMI).

In addition, as it is often effective to use different types of features for noisy speech recognition, such as in the tandem approach, we propose a method that obtains new transformed features $\mathbf{y}'_t$ by adding features $\mathbf{h}'_t$ to $\mathbf{h}_t$ as

$$\mathbf{y}'_t = \mathbf{x}_t + \begin{bmatrix}\mathbf{M}\ \mathbf{M}'\end{bmatrix} \begin{bmatrix}\mathbf{h}_t \\ \mathbf{h}'_t\end{bmatrix}. \tag{6}$$

The concatenated matrices $\mathbf{M}$ and $\mathbf{M}'$ are optimized by maximizing the following objective function:

$$\mathcal{F}_{\text{af-MMI}}\left(\begin{bmatrix}\mathbf{M}\ \mathbf{M}'\end{bmatrix}\right) = \sum_{r=1}^{R} \log \frac{p_\lambda\left(\{\mathbf{y}'_t\}_r|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda\left(\{\mathbf{y}'_t\}_r|\mathcal{H}_s\right)^\kappa p_L(s)}. \tag{7}$$

By augmenting discriminative feature transforms in this way we can consider a wider class of transforms depending upon the chosen auxiliary features. One could consider alternative region-based features, features deriving from source-separation considerations such as signal-to-noise ratios, masking values, or any other set of alternative features.

4.3. Discriminative language modeling

Discriminative Language Modeling (DLM) learns patterns of errors in the hypotheses output by a speech recognizer, and adjusts scores of hypotheses to reduce the errors. The score can be simply obtained by the inner product of a feature vector $\phi(\mathcal{H}_s)$ (e.g., N-gram counts) extracted from a hypothesis (recognition result) $\mathcal{H}_s$ and a weight vector $\mathbf{w}$. The weight vectors for each utterance are estimated based on an on-line manner by using the following perceptron learning rule, $\mathbf{w} = \mathbf{w} + (\phi(\mathcal{H}_{s_r}) - \phi(\mathcal{H}_s))$. Then, an aggregate weight vector is obtained by averaging the weight vectors for all utterances (i.e., averaged perceptron). In our paper, the approach is realized by re-ranking the hypotheses, and $\mathcal{H}_{s_r}$ is selected from the lowest WER (oracle) hypothesis in an N-best list computed from the corresponding reference.

4.4. Minimum Bayes risk decoding for discriminative language modeling

Minimum Bayes Risk (MBR) decoding finds the hypothesis that approximately minimizes the Bayes risk with respect to the word error rate based on the lattice representation. To efficiently combine DLM based on the N-best re-scoring framework and MBR decoding based on the lattice framework, we use the algorithm of [16]. This algorithm forms a consensus by choosing a word sequence that has the minimum expected edit distance to each sequence in the lattice:

$$\mathcal{H}_{\hat{s}} = \arg\min_{s'} \sum_{s \in \mathcal{L}} p_\lambda\left(\{\mathbf{y}'_t\}|\mathcal{H}_s\right)^\kappa p_L(s) L(\mathcal{H}_s, \mathcal{H}_{s'}), \tag{8}$$

where $L(\mathcal{H}_s, \mathcal{H}_{s'})$ is the edit distance between a hypothesis in the lattice $\mathcal{H}_s$ and that of the argument of the minimization, $\mathcal{H}_{s'}$.

The edit distance is approximately computed based on the probability $\gamma(q, u)$ of which symbol (including the epsilon symbol) u is aligned at the position q in the word sequence $\mathcal{H}_{\hat{s}}$. The approximate objective is iteratively updated, conventionally starting at the current 1-best hypothesis from the lattice, and forming alignments with the lattice sequences. Our approach improves upon the initialization point by starting with 1-best result in an N-best list re-scored by DLM, rather than the conventional 1-best result. $\gamma(q, u)$ is approximately computed by using the original (non-rescaled) arc scores of the DLM 1-best result[2]. Thus, we efficiently combine DLM-based N-best re-scoring and minimum Bayes risk decoding.

5. EXPERIMENTAL SETUP

5.1. Task description

We validated the effectiveness of our proposed approach for reverberated and noisy speech on Track 2 of the 2$^{\text{nd}}$ CHiME challenge [1]. Track 2 is a medium-vocabulary task in reverberant and noisy environment, whose utterances are taken from the Wall Street Journal database (WSJ0). The training data set (si_tr_s) contains 7138 utterances by 83 speakers (si84), the evaluation data set (si_et_05) contains 330 utterances by 12 speakers (Nov'92), and the development set (si_dt_05) contains 409 utterances by 10 speakers. Acoustic models were trained using si_tr_s and some of the parameters (e.g., language model weights) were tuned based on the WERs of si_dt_05. The language model size was 5 k (basic). This database simulates a realistic environment. There are two types of data: "reverberated," created by convolving clean speech with binaural room impulse responses corresponding to a frontal position at a distance of 2 m from the microphones in a family living room, and "isolated," created by adding real-world noises recorded in the same room to "reverberated" and selecting the noise excerpts to obtain signal-to-noise ratio (SNR) ranges of -6, -3, 0, 3, 6, and 9 dB without rescaling. Noises are non-stationary such as other speakers' utterances, home noises, or music.

5.2. Feature extraction and transformation

We describe the settings of acoustic feature and feature transformation. The baseline acoustic features are MFCC and PLP (1-13 order MFCCs (PLPs) + Δ + $\Delta\Delta$). It is well known that LDA transforms

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

the features of a class to make them as discriminable as possible to those of the other classes. After concatenating 1-13 order static MFCCs in nine contiguous frames, a total of 117-dimensional features are compressed into 40 dimensions by an LDA whose class is a tri-phone HMM state (2500 states). Because the acoustic features are high dimensional, it is impossible to use full-covariance models (which consider correlations between dimensions), and, instead, diagonal-covariance models are widely used. This simplification decreases the model's performance. Several methods for transforming a feature space so as to decrease correlations between features have thus been proposed, among which MLLT is a widely used example. Moreover, large variations among speakers degrade the acoustic models. To address this problem, SAT is typically used: in SAT, training is conducted after having transformed the training speech into a canonical space so as to reduce the variances across speakers. In this study, we validated the effectiveness of LDA, MLLT, and SAT.

5.3. 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 dimension of the feature vector $\mathbf{h}_t$ was $400 \times 40 \times 9$. Features with the top 2 posteriors were selected and all other features were ignored. For the DNN, we used a CPU version of neural network training implemented in Kaldi with 3 hidden layers and 500,000 parameters. The initial learning rate was 0.01 and was decreased to 0.001 at the end of training.

5.4. Experimental procedure

We summarize the experimental procedure based on the above setup as follows: First, a clean acoustic model was trained. The number of mono-phones was 40, including silence ("sil"). In the tri-phone model, the number of states was 2500 and the total number of Gaussians was 15000. Second, using their alignments and tri-phone tree structures, reverberated acoustic models were trained using the "reverberated" dataset. Third, noisy acoustic models were trained multi-conditionally using the "isolated" dataset without any pre-processing such as blind source separation. Finally, using this ML model, we validated the effectiveness of the discriminative training and feature transformation for the "isolated" dataset. The parameters used in our experiments were set as those in the WSJ tutorial attached to the Kaldi toolkit.

6. RESULTS AND DISCUSSION

6.1. Maximum likelihood baseline

We retrained the initial tri-phone model (trained on clean data) using reverberated and noisy data. Reverberation and noises cause errors in the alignment and reconstruction of tree structures. We consider whether alignment (A) and tree structures (T) are reconstructed (y), i.e., retrained on noisy data, or not (n), i.e., the same to those of clean model. There are three conditions: (A=n,T=n), (A=y,T=n), and (A=y,T=y). From now on, we evaluate the WER on the development set (si_dt_05). For the "reverberated" case, the WERs of the tri-phone models (ML) are 12.69% (A=n,T=n), 12.05% (A=y,T=n), 12.35% (A=y,T=y). Using an alignment by the (A=y,T=n) model (the model that achieves the best performance), we retrained models on the "isolated" dataset. The averages of these ML models are 56.29% (A=n,T=n), 56.37% (A=y,T=n), and 56.98% (A=y,T=y). The performance of the (A=y,T=y) model is inferior to that of the other models. The performances of (A=n,T=n) and (A=y,T=n) are almost equivalent; we use the (A=y,T=n) model as a baseline model. Discriminative training and feature transformation were carried out using this model as the initial model.

Table 1. WER[%] for isolated speech (**si_dt_05**) without noise suppression. Tri-phone model, discriminative training with MFCC features (upper), MFCC+LDA+MLLT (middle), MFCC+LDA+MLLT+SAT (lower).

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
ML	74.20	66.57	58.24	51.84	46.73	40.64	56.37
MMI	73.40	65.60	56.88	51.17	45.40	41.20	55.61
bMMI	72.78	64.71	55.69	50.83	44.00	40.27	54.71
f-MMI	69.94	62.50	54.51	48.74	42.73	38.34	52.79
f-bMMI	68.64	61.56	53.11	47.65	41.73	36.98	51.61
ML	70.95	62.62	53.98	47.37	40.27	34.84	51.67
MMI	68.55	61.12	53.41	46.32	39.52	34.30	50.54
bMMI	68.74	60.98	51.95	45.86	38.16	32.85	49.76
f-MMI	66.19	58.24	49.23	43.58	36.89	31.35	47.58
f-bMMI	66.65	57.46	48.25	42.99	35.71	31.07	47.02
ML	68.36	58.30	48.80	40.73	35.09	28.54	46.64
MMI	65.13	55.27	45.89	39.64	33.12	27.29	44.39
bMMI	64.60	55.10	45.82	39.05	32.72	26.86	44.03
f-MMI	63.09	52.62	42.44	36.29	31.01	25.52	41.83
f-bMMI	**62.43**	**52.23**	**42.17**	**35.31**	**29.84**	**24.72**	**41.12**

6.2. Discriminative training and feature transformation

First, with regard to the MFCC features, the improvement of the WER by discriminative training from the ML baseline is shown in Table 1 (upper). The mixture of speech and noise increases the likelihood of detecting erroneous phonemes and leads to incorrect recognition. These errors could be modified by discriminative training. The boosted model improves the WER by 1% relative to the non-boosted one, whereas the feature space technique improves the WER by 3%. We believe that the feature space is adapted for a target speaker to improve the WER and that this effect reduces the influence of other noises. In these tables, the boosting factor is set to 0.1. The preliminary experiments show that the performance does not heavily depend on the boosting factors and that the optimized values of the boosting factor are approximately 0.1-0.2. Denominator lattices for discriminative training are generated using ML model.

Second, the MFCC features are transformed using LDA and MLLT. Table 1 (middle) shows the WER, whereas LDA by itself achieves 54.37% (ML). This shows that features that are highly discriminable from other phonemes can be obtained by LDA. These significant improvements are partly a result of the characteristics of the CHiME database. As mentioned in the Introduction, LDA and MLLT improve the model performance in ordinary noise conditions. Additionally, the CHiME database's noises include many utterances by other people. These types of noises are best suited to be handled by LDA, because if two or more phonemes exist in the same frame when sources are mixed, the model can possibly discriminates between these phonemes separately, as if it were a source separation problem. It is also effective to use context to reduce the influence of non-stationary noises. Furthermore, although noises increase the correlations between MFCC coefficients in each dimension, MLLT reduces the correlations and improves the WER. Denominator lattices for discriminative training are re-generated using

Table 2. WER[%] for isolated speech (**si_dt_05**) without noise suppression. Tri-phone model, discriminative feature transformation with PLP (P) features.

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
ML(M)	74.20	66.57	58.24	51.84	46.73	40.64	56.37
ML(P)	74.57	67.50	59.76	53.02	47.00	42.23	57.35
f-MMI	69.94	62.50	54.51	48.74	**42.73**	38.34	52.79
(+P)	**69.52**	**62.31**	**54.48**	**48.59**	42.94	**37.90**	**52.62**

Table 3. WER[%] for isolated speech (**si_dt_05**) with noise suppression by prior-based binary masking. Tri-phone model, discriminative training with MFCC features (upper), MFCC+LDA+MLLT (middle), MFCC+LDA+MLLT+SAT (lower).

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
ML	66.82	57.87	48.86	42.29	38.18	31.86	47.65
(+MBR)	66.16	57.09	48.15	41.47	37.16	31.23	46.88
bMMI	65.73	56.98	46.95	41.57	36.27	31.02	46.42
f-bMMI	63.40	54.05	44.28	38.87	33.72	29.90	44.04
ML	64.64	54.24	46.35	37.91	32.75	28.96	44.14
bMMI	63.39	52.54	44.56	35.60	30.98	28.10	42.53
f-bMMI	60.92	50.41	41.76	33.59	29.56	25.90	40.36
DNN	57.21	45.85	36.21	30.61	26.36	23.31	36.59
ML	59.94	47.93	39.83	33.01	28.00	23.47	38.70
bMMI	56.90	45.79	37.60	30.31	26.15	21.74	36.42
f-bMMI	52.93	42.62	34.59	27.63	24.27	20.24	33.71
(+DLM)	53.16	42.93	34.36	27.26	23.72	**19.47**	33.48
(+MBR)	52.65	**42.04**	33.75	27.05	23.74	19.91	33.19
(both)	**52.54**	42.09	**33.72**	**27.02**	**23.66**	19.66	**33.11**

ML (MFCC+LDA+MLLT) model.

Third, we added SAT and fMLLR to the model described in the second step. Table 1 (lower) shows the WER. As the amount of training data is very limited, transformation into a canonical space, which leads to an increase in the effective amount of training data, has a strong impact on the estimation accuracy of the acoustic models. Additionally, fMLLR adaptation for a target speaker reduces the influence of noises. Denominator lattices for discriminative training are re-generated using ML model.

6.3. Augmented discriminative feature transformation

Table 2 shows the WER of ML and f-MMI whose auxiliary features $\mathbf{h}'_t$ in Eq. (6) are PLP (13 dimensions each), respectively. In the ML model, we observe that the performance of PLP is worse than that of MFCC by about 1% absolute. However, adding PLP to the discriminative feature transformation improves the WER. Thus, it is effective to obtain new features that contain information that cannot be obtained using the features $\mathbf{h}_t$.

6.4. Noise suppression

Table 3 shows the WER with noise suppression by prior-based binary masking. Binary masking improves the WER in all cases by 7% to 9% absolute. We tried several α and $\alpha = 0.25$ achieved the best WER. Directional noise is suppressed to some extent but diffused noises such as music still remain.

Table 4. WER[%] for isolated speech (**si_et_05**) without noise suppression. The baseline is ML (MFCC), whereas on top of MFCC+LDA+MLLT+SAT, "Best 1" is ML and "Best 2" is feature-space boosted MMI.

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
Baseline	69.79	62.71	55.86	46.89	42.07	37.49	52.47
Best 1	60.83	52.14	43.51	34.28	29.22	23.82	40.63
Best 2	**54.70**	**45.11**	**35.98**	**28.64**	**24.38**	**21.39**	**35.04**

6.5. Deep neural network

Table 3 also provides the WER of a DNN based on ML baseline with MFCC+LDA+MLLT (middle) after the noise suppression step. The DNN result outperformed bMMI and f-bMMI results by 2.8% at the most and was comparable to the SAT system (lower). This shows the potential effectiveness of DNN for reverberated noisy speech recognition. Although DNN is not embedded to our total system currently, the integration of DNN and our system is likely to further improve ASR performance.

6.6. Discriminative language modeling and minimum Bayes risk decoding

Weights $\mathbf{w}$ of a discriminative language model are learned on the training data set using 100-best recognition candidates, where the weight w_0 associated with the original score is set to 20. Using these weights, results are re-ranked, with w_0 set to 13. Weights are obtained by averaged perceptron at three iterations. Features are counts of uni-grams, bi-grams, and tri-grams. DLM improves WER by 0.23% on average, especially for 9dB with a 0.77% improvement. DLM is not always effective because, while error tendencies are dependent on SNR, training is performed on the whole training set, which includes all SNRs. This leads to a mismatch between training and recognition, damaging performance.

MBR improves the WER by 0.77% for ML (MFCC) and 0.52% for f-bMMI (MFCC+LDA+MLLT+SAT). The performance of MBR is stable with respect to SNR. Combination of DLM and MBR as mentioned in Section 4.4 improves the WER because DLM refines the initial 1-best result and adapts to error tendencies inherent to the decoder.

6.7. Evaluation set

Table 4 shows the WERs on the evaluation set using the models tuned using the development set. The baseline is ML (MFCC), whereas on top of MFCC+LDA+MLLT+SAT, "Best 1" is ML and "Best 2" is f-bMMI. Using both discriminative training and feature transformation ("Best 2") achieves 33.22% error reduction relative to the baseline. Thus, we show the effectiveness of both discriminative training and feature transformation for reverberated and noisy speech. The WERs after noise suppression are shown in Table 5, which represents a 37.9% error reduction.

The HTK baseline results for si_dt_05 and si_et_05 using our front end are shown in Table 6 as a reference. "Denoised" are the results obtained with HMMs retrained on denoised data. "Noisy" are the results obtained with the original HMMs trained on the noisy data. Performance is lower than that of Kaldi, but the settings are different and only limited tuning was performed for HTK.

Table 5. WER[%] for isolated speech (**si_et_05**) with noise suppression. The baseline is ML (MFCC), whereas on top of MFCC+LDA+MLLT+SAT, "Best 1" is ML and "Best 2" is feature-space boosted MMI.

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
Baseline	60.58	52.87	45.60	37.70	33.38	29.24	43.23
Best 1	50.91	41.64	33.89	26.30	21.61	18.85	32.20
Best 2	44.54	35.91	29.24	22.31	17.77	15.88	27.61
(+DLM)	44.27	35.48	28.75	21.61	**17.34**	15.37	27.14
(+MBR)	44.51	**35.42**	28.81	21.46	17.41	14.98	27.10
(both)	**44.12**	35.46	**28.12**	**21.20**	17.43	**14.83**	**26.86**

Table 6. WER[%] for isolated speech with noise suppression by prior-based binary masking (tri-phone model) using HTK with MFCC features.

si_dt_05

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
denoised	72.18	66.16	57.95	53.99	48.36	43.58	57.04
noisy	74.67	68.08	61.12	56.61	51.33	47.65	59.91

si_et_05

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
denoised	68.56	61.97	56.34	48.76	43.51	40.58	53.29
noisy	72.00	65.27	59.05	52.34	48.57	44.14	56.90

7. CONCLUSIONS

We developed a state-of-the-art recognition system following a simple prior-based binary masking for realistic reverberated and noisy environments and validated the effectiveness of both feature transformation and discriminative methods. Combination of MBR and DLM improves the WER by considering error tendencies, which are inherent to the decoder. Experiments show that these techniques are effective for non-stationary interference and reverberation.

8. REFERENCES

[1] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The 2nd 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. ICASSP*, 2013.

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

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

[4] 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 Trans. Audio, Speech, Language Process.*, vol. 15, pp. 203–223, Jan. 2007.

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

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

[7] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 272–281, Jul. 1999.

[8] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

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

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Process. Mag.*, vol. 28, pp. 82-97, Nov. 2012.

[11] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the AMI and AMIDA projects," in *Proc. ASRU*, 2007, pp. 238–247.

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

[13] T. Oba, T. Hori, A. Nakamura, and A. Ito, "Round-robin duel discriminative language models," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, pp. 1244–1255, May 2012.

[14] E. Dikici, M. Semarci, M. Saraçlar, and E. Alpaydin, "Classification and ranking approaches to discriminative language modeling for ASR," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, pp. 291–300, Feb. 2013.

[15] V. Goel and W.J. Byrne, "Minimum Bayes-risk automatic speech recognition," *Computer Speech & Language*, vol. 14, pp. 115–135, Apr. 2000.

[16] H. Xu, D. Povey, L. Mangu, and J. Zhu, "An Improved Consensus-like method for minimum Bayes risk decoding and lattice Combination," in *Proc. ICASSP*, 2010, pp. 4938-4941.

[17] H. Kuo, L. Mangu, E. Arisoy, and G. Saon, "Minimum Bayes risk discriminative language models for Arabic speech recognition," in *Proc. of ASRU*, 2011, pp. 208-213.

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

[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.4.1)," *http://htk.eng.cam.ac.uk*, March 2009.

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

[21] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, pp. 516–527, Mar. 2011.

[22] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S. Hahm, and A. Nakamura, "Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation," in *Proc. CHiME*, 2011, pp. 12-17.

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