FEATURE ENHANCEMENT BASED ON GENERATIVE-DISCRIMINATIVE HYBRID APPROACH WITH GMMS AND DNNS FOR NOISE ROBUST SPEECH RECOGNITION

Masakiyo Fujimoto and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Japan
{fujimoto.masakiyo, nakatani.tomohiro}@lab.ntt.co.jp

ABSTRACT

This paper presents a technique that combines generative and discriminative approaches with Gaussian mixture models (GMMs) and deep neural networks (DNNs) for model-based feature enhancement. Typical model-based feature enhancement employs a generative model approach. The enhanced features are obtained by using the weighted sum of linear transformations given by each Gaussian component contained in GMMs and corresponding posterior probabilities. The computation of posterior probabilities is a crucial factor for this kind of feature enhancement, and can also be formulated as the class discrimination problem of observed noisy features. The prominent discriminability of DNNs is a well-known solution to this discrimination problem. Therefore, we propose the use of DNNs for computing posterior probabilities. The proposed method incorporates the benefit of the discriminative approach into the generative approach. For AURORA2 task evaluations, the proposed method provided noticeable improvements compared with results obtained using the conventional generative model approach.

Index Terms— feature enhancement, generative-discriminative hybrid approach, deep neural networks, unsupervised modeling

1. INTRODUCTION

Ensuring the noise robustness of automatic speech recognition (ASR) is becoming a more critical, because speech applications including voice searches are now used in various environments. Noise robust ASR techniques are generally classified into two types. The front-end processing of ASR attempts to remove the influence of noise from observed signals, and includes robust feature extraction [1, 2], feature space normalization [3, 4, 5], and feature enhancement (noise suppression) [6, 7, 8, 9]. Recently, deep neural network (DNN)-based approaches have attracted attention and they include a DNN bottle neck feature [10], a denoising autoencoder (DAE) [11, 12], a recurrent neural network-based DAE [13], and DNN-based ideal binary masking (IBM) [14, 15]. On the other hand, back-end processing techniques attempt to adapt an acoustic model to observed signals. For the traditional Gaussian mixture model (GMM)-hidden Markov model (HMM) systems, there are various techniques of model compensation [16, 17, 18, 19] and model adaptation [20, 21, 22]. For recent DNN-HMM systems, various training techniques have been proposed including noise adaptive training [23] and noise aware training [24]. Of these various techniques, we have focused our research on model-based feature enhancement.

For model-based feature enhancement, we have recently proposed unsupervised joint speaker adaptation and noise GMM estimation by using minimum mean squared error (MMSE) estimates of the clean speech and noise [9]. This method estimates the linear transformation from an observed noisy feature to a clean feature by using GMMs of clean speech and observed signals. Then, the GMM of the observed signals is obtained by composing GMMs of clean speech and noise. Unsupervised joint speaker adaptation and noise GMM estimation are also carried out by using a given observed signal. With the MMSE approach, the linear transformation is estimated from the weighted sum of each linear transformation basis. These sums are obtained by using the mean vector of each Gaussian component contained in GMMs of clean speech and observed signals. The weights are given as the posterior probability of the observed signal. This method is based on the generative model approach, and efficiently utilizes the property of GMMs.

The generative model approach is effective for the implementation of model-based feature enhancement, because it allows straightforward linear transformation estimation and unsupervised model parameter estimation. However, generative models, and especially mixture models including the GMM, have a class discrimination problem. The class referred to here means the latent variable of the GMM, which indicates the Gaussian components that generate the observed signal. To achieve accurate feature enhancement, we should select suitable Gaussian components that correspond to the observed signal. Consequently, the computation of the posterior probability, i.e., the estimation of the class discrimination probability, is a critical factor as regards the GMM-based approach.

In terms of the above problem, a discriminative model that focuses on the class discrimination of given data by utilizing supervised learning is usually superior to generative model-based discrimination. As the discriminative model, the prominent discriminability of DNNs is well known in recent research; therefore, the use of DNNs is a reasonable techniques for realizing class discrimination (computation of posterior probability). With this consideration, by combining a GMM-based approach and DNN-based discrimination, we investigate a generative-discriminative hybrid approach designed to incorporate the benefits of both the generative and the discriminative models. This hybrid approach applies the GMM-based generative model to linear transformation estimation and unsupervised model parameter estimation, and computes posterior probability with the DNN-based discriminative model. With this hybrid approach, we can realize accurate feature enhancement performance.

The proposed method was evaluated on the AURORA2 task [25]. The evaluation results reveal that the proposed method successfully improves the ASR accuracies of both tasks in results obtained with the conventional generative model approach.

2. RELATED WORK

Recently, model-based approaches have been widely used as powerful tools for noise robust ASR. As a representative technique of model-based noise suppression, a vector Taylor series (VTS)-based approach [8] and its various extensions have been proposed [17, 18, 19]. The VTS-based approach compensates the model of an observed signal with models of clean speech and noise with Taylor series-based linear approximation. Then, model parameters are updated by using the EM-algorithm and the given observed signals. A typical VTS-based approach employs a single Gaussian distribution for the noise model. Since non-stationary noise has a multi-modal distribution and a temporal structure, a single Gaussian distribution is unsuitable for the model of the non-stationary noise. Therefore, a model with a complex structure, e.g., a GMM or an HMM, is needed to ensure robustness against non-stationary noise.

A stereo-based approach, which utilizes corresponding clean and noisy speech data, has also been proposed. This approach trains a mapping function from a noisy feature to a clean feature by using stereo data. Representative stereo-based approaches are stereo-based piecewise linear compensation for environments (SPLICE) [26], DAE [11, 12, 13], and IBM [14, 15]. SPLICE trains a linear mapping function from a noisy feature to a clean feature with a joint probability density function (PDF). DAE directly estimates a clean feature from a noisy feature by using neural networks. IBM trains various binary time-frequency masks, and selects a suitable mask for a given observed signal. These stereo-based approaches provide considerable improvement in ASR accuracy with sufficient training of the mapping function, whereas they are not always robust in unknown noise environments. As an extension of SPLICE, a discriminative criterion-based SPLICE has been proposed [27]. This method also utilizes the DNN output for the weighted sum of linear transformations. However, since linear transformations are estimated by using stereo data and the DNN output, the environmental dependency increases greatly.

To cope with problems of presented by the above studies, our proposed method introduces an unsupervised scheme and a discriminative criterion into feature enhancement and model parameter estimation.

3. UNSUPERVISED JOINT SPEAKER ADAPTATION AND NOISE GMM ESTIMATION

This section briefly reviews our previous work, namely model-based feature enhancement based on unsupervised joint speaker adaptation and noise GMM estimation [9].

3.1. Definition of GMMs

In our method, the speaker independent (SI) clean speech model is given by a GMM with K Gaussian components in the D-dimensional log mel-filter bank (LMFB) domain, and has model parameters that consist of the mixture weight $w_{S,k}$, the mean vector $\boldsymbol{\mu}_{S,k} \triangleq \{\mu_{S,k,d}\}_{d=0}^{D-1}$, and the diagonal variance matrix $\boldsymbol{\Sigma}_{S,k} \triangleq \mathrm{diag}\,\{\sigma_{S,k,d}\}_{d=0}^{D-1}$. k and d denote the indices of the Gaussian component and the element of a vector or a diagonal component of a matrix. Then, we apply the global bias-based adaptation [28] to the mean vector of the SI clean speech GMM, i.e., $\tilde{\boldsymbol{\mu}}_{S,k} = \boldsymbol{\mu}_{S,k} + \boldsymbol{b}$, where $\tilde{\boldsymbol{\mu}}_{S,k} \triangleq \{\tilde{\mu}_{S,k,d}\}_{d=0}^{D-1}$ and $\boldsymbol{b} \triangleq \{b_d\}_{d=0}^{D-1}$ denote the adapted mean vector and the bias vector, respectively.

On the other hand, the noise model is also given by a GMM with L Gaussian components in the LMFB domain, and has model parameters that consist of the mixture weight $w_{N,l}$, the mean vector $\boldsymbol{\mu}_{N,l} \triangleq \{\mu_{N,l,d}\}_{d=0}^{D-1}$, and the diagonal variance matrix $\boldsymbol{\Sigma}_{N,l} \triangleq \mathrm{diag}\,\{\sigma_{N,l,d}\}_{d=0}^{D-1}$, where l denotes the Gaussian index.

3.2. Mismatch function and model compensation

With the LMFB vectors of the clean speech $\boldsymbol{S}_t \triangleq \{S_{t,d}\}_{d=0}^{D-1}$ and the noise $\boldsymbol{N}_t \triangleq \{N_{t,d}\}_{d=0}^{D-1}$ at the t-th frame, the LMFB vector of the observed signal $\boldsymbol{O}_t \triangleq \{O_{t,d}\}_{d=0}^{D-1}$ is derived by the following mismatch function.

$$O_{t,d} = S_{t,d} + \log\left(1 + \exp\left(N_{t,d} - S_{t,d}\right)\right) \equiv h\left(S_{t,d}, N_{t,d}\right) \quad (1)$$

Based on this mismatch function, the GMM parameters of the observed signal, which consist of the mixture weight $w_{O,k,l}$, the mean vector $\boldsymbol{\mu}_{O,k,l} \triangleq \{\mu_{O,k,l,d}\}_{d=0}^{D-1}$, and the diagonal variance matrix $\boldsymbol{\Sigma}_{O,k,l} \triangleq \mathrm{diag}\,\{\sigma_{O,k,l,d}\}_{d=0}^{D-1}$ are compensated as follows:

$$w_{O,k,l} = w_{S,k} \cdot w_{N,l} \quad (2)$$

$$\mu_{O,k,l,d} = h\left(\tilde{\mu}_{S,k,d}, \mu_{N,l,d}\right) \quad (3)$$

$$\sigma_{O,k,l,d} \simeq H_{k,l,d}^2 \cdot \sigma_{S,k,d} + \left(1 - H_{k,l,d}\right)^2 \cdot \sigma_{N,l,d} \ , \quad (4)$$

with the Jacobian $H_{k,l,d} = \partial h\left(\tilde{\mu}_{S,k,d}, \mu_{N,l,d}\right) / \partial\tilde{\mu}_{S,k,d}$.

3.3. Parameter estimation with MMSE estimates

The initial parameters of speaker adaptation and the noise GMM are given as

$$\boldsymbol{b} = \boldsymbol{0} \quad (5)$$

$$w_{N,l} = \frac{1}{L} \ , \ \boldsymbol{\mu}_{N,l} \sim \mathcal{N}\left(\cdot \left| \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{\Sigma}}_N \right.\right) \ , \ \boldsymbol{\Sigma}_{N,l} = \hat{\boldsymbol{\Sigma}}_N \ , \quad (6)$$

where $\boldsymbol{0}$ denotes the zero vector. $\boldsymbol{\mu}_{N,l}$ is initialized by the multivariate Gaussian random value $\mathcal{N}\left(\cdot \left| \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{\Sigma}}_N \right.\right)$ with $\hat{\boldsymbol{\mu}}_N = \frac{1}{U}\sum_{t=0}^{U-1} \boldsymbol{O}_t$ and $\hat{\boldsymbol{\Sigma}}_N = \mathrm{diag}\left\{\frac{1}{U}\sum_{t=0}^{U-1} \boldsymbol{O}_t \boldsymbol{O}_t^{\mathrm{T}} - \hat{\boldsymbol{\mu}}_N \hat{\boldsymbol{\mu}}_N^{\mathrm{T}}\right\}$, where $\mathcal{N}(\cdot|\cdot)$ denotes the PDF of the Gaussian distribution.

Each parameter is estimated by using the EM algorithm [9] with the MMSE estimates of the speech $\tilde{\boldsymbol{S}}_t$ and the noise $\tilde{\boldsymbol{N}}_t$ derived as:

$$\tilde{\boldsymbol{S}}_t = \boldsymbol{O}_t + \sum_{k,l} P_{O,t,k,l} \cdot \left(\tilde{\boldsymbol{\mu}}_{S,k} - \boldsymbol{\mu}_{O,k,l}\right) \quad (7)$$

$$\tilde{\boldsymbol{N}}_t = \boldsymbol{O}_t + \sum_{k,l} P_{O,t,k,l} \cdot \left(\boldsymbol{\mu}_{N,l} - \boldsymbol{\mu}_{O,k,l}\right) \ , \quad (8)$$

with the posterior probability

$$P_{O,t,k,l} = \frac{w_{O,k,l} \cdot \mathcal{N}\left(\boldsymbol{O}_t \left| \boldsymbol{\mu}_{O,k,l}, \boldsymbol{\Sigma}_{O,k,l} \right.\right)}{\sum_{k,l} w_{O,k,l} \cdot \mathcal{N}\left(\boldsymbol{O}_t \left| \boldsymbol{\mu}_{O,k,l}, \boldsymbol{\Sigma}_{O,k,l} \right.\right)} \ . \quad (9)$$

Then, the posterior probability w.r.t. the clean speech GMM $P_{S,t,k}$ and the posterior probability w.r.t. the noise GMM $P_{N,t,l}$ are given by marginalizing $P_{O,t,k,l}$ as follows:

$$P_{S,t,k} = \sum_{l} P_{O,t,k,l} \quad (10)$$

$$P_{N,t,l} = \sum_{k} P_{O,t,k,l} \ . \quad (11)$$

With $\tilde{\boldsymbol{S}}_t$, $\tilde{\boldsymbol{N}}_t$, $P_{S,t,k}$ and $P_{N,t,l}$, the target parameters are estimated as follows:

$$\boldsymbol{b} = \left(\sum_{t,k} P_{S,t,k} \cdot \boldsymbol{\Sigma}_{S,k}^{-1}\right)^{-1} \sum_{t,k} P_{S,t,k} \cdot \boldsymbol{\Sigma}_{S,k}^{-1}\left(\tilde{\boldsymbol{S}}_t - \boldsymbol{\mu}_{S,k}\right) \quad (12)$$

$$w_{N,l} = \frac{\sum_t P_{N,t,l}}{\sum_{t,l} P_{N,t,l}} \quad (13)$$

$$\boldsymbol{\mu}_{N,l} = \frac{\sum_t P_{N,t,l} \cdot \tilde{\boldsymbol{N}}_t}{\sum_t P_{N,t,l}} \quad (14)$$

$$\boldsymbol{\Sigma}_{N,l} = \frac{\sum_t P_{N,t,l} \cdot \tilde{\boldsymbol{N}}_t \tilde{\boldsymbol{N}}_t^{\mathrm{T}}}{\sum_t P_{N,t,l}} - \boldsymbol{\mu}_{N,l}\boldsymbol{\mu}_{N,l}^{\mathrm{T}} \ . \quad (15)$$

By iterating the MMSE estimation of Eqs. (7) and (8) and the parameter estimation of Eqs. (12) to (15) until convergence, the accuracies of the MMSE estimates and the model parameter will be mutually improved. The result of feature enhancement is obtained as MMSE estimate $\tilde{\boldsymbol{S}}_t$ at the final iteration.

4. FEATURE ENHANCEMENT BASED ON GENERATIVE-DISCRIMINATIVE HYBRID APPROACH

4.1. Generative-discriminative hybrid approach

The method described in Sec. 3 consists of three modules, i.e., MMSE estimation, parameter estimation, and posterior probability computation. Each module utilizes the GMMs of the clean speech, the noise, and the observed signal. Therefore, this method is based on the framework of the generative model approach. By utilizing GMMs, we can easily provide suitable parameters for MMSE estimation, e.g., the second term on the right side of Eqs. (7) and (8), by using the expectations of the PDF or some sampling algorithms. In addition, we can also easily implement the unsupervised parameter estimation given by Eqs. (12) to (15). These properties prove that this generative model (GMM-based) approach is suitable for model-based feature enhancement.

In the method, the MMSE estimation and parameter estimation modules both require the posterior probability $P_{O,t,k,l}$ given by Eq. (9). To achieve accurate feature enhancement, the computation of the posterior probability $P_{O,t,k,l}$ is a critical factor. In this problem, $P_{O,t,k,l}$ indicates the class discrimination probability of a given observed signal $\boldsymbol{O}_t$. The class referred to here means a latent variable of a GMM, i.e., the index of a Gaussian component that generates $\boldsymbol{O}_t$ in the GMM data generation process. Thus, the computation of $P_{O,t,k,l}$ is equivalent to the class discrimination problem of $\boldsymbol{O}_t$.

With this discrimination problem, the discriminative model approach is usually superior to the generative model approach, because it focuses on the discrimination of given data by utilizing supervised learning. In consideration of this model property, the use of the discriminative model is a reasonable way to realize the accurate class discrimination of $\boldsymbol{O}_t$. Thus, we investigate a generative-discriminative hybrid approach to incorporate the benefits of both generative and discriminative models. This hybrid approach applies a generative model approach to MMSE estimation and parameter estimation by using GMMs, and computes the posterior probability with a discriminative model approach by using DNNs.

4.2. Discriminative posterior probability

With Eqs. (10) and (11), posterior probability $P_{O,t,k,l}$ is represented as joint probabilities of a clean speech GMM and a noise GMM as follows:

$$P_{O,t,k,l} = P_{S,t,k} \cdot P_{N,t,l} \ . \tag{16}$$

The proposed method aims to estimate the noise model with utterance-wise processing, however, it is currently difficult to estimate the DNN of noise using only a given observation. Thus, the proposed method employs the GMM posterior probability $P_{N,t,l}$ for noise (see Eq. (11)). On the other hand, the posterior probability w.r.t. speech is given by the following DNN output instead of GMM posterior probability.

$$P^{(DNN)}_{S,t,k} = \mathrm{softmax}_k \left(\boldsymbol{W} \boldsymbol{x}^{(M)}_t + \boldsymbol{v} \right) \ , \tag{17}$$

where $P^{(DNN)}_{S,t,k}$ denotes the class discrimination probability given by the softmax output of a DNN with M hidden layers. $\mathrm{softmax}_k(\cdot)$ denotes the softmax function that gives the softmax output of the k-th node at the output layer. $\boldsymbol{W}$, $\boldsymbol{v}$, and $\boldsymbol{x}^{(M)}_t$ denote the weight matrix and the bias vector of the output layer and the output of the M-th hidden layer, respectively. The DNN is trained by using alignment labels Lab_t given by Eq. (18). By using these alignment labels, each output node of the DNN corresponds to each Gaussian component contained in a clean speech GMM.

$$Lab_t = \arg\max_k P_{S,t,k} \tag{18}$$

By using $P^{(DNN)}_{S,t,k}$ and $P_{N,t,l}$, the discriminative posterior probability $P^{(DNN)}_{O,t,k,l}$ is derived as:

$$P^{(DNN)}_{O,t,k,l} = P^{(DNN)}_{S,t,k} \cdot P_{N,t,l} \ . \tag{19}$$

Then, MMSE estimation and parameter estimation are carried out by using the discriminative posterior probability $P^{(DNN)}_{O,t,k,l}$ instead of the GMM posterior probability $P_{O,t,k,l}$.

4.3. Processing flow

The following algorithm summarizes the proposed method, and is applied to each utterance.

Algorithm 1 Proposed feature enhancement with GMMs and DNNs

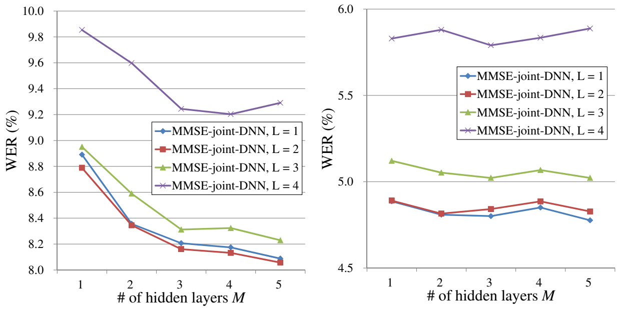
```
Feature extraction of $\boldsymbol{O}_t$ for all $t$
Initialization (Eqs. (5) and (6))
repeat
    Model compensation (Eqs. (2), (3), and (4))
    Compute posterior probabilities of noise GMM (Eqs. (9) and (11))
    Compute softmax outputs of DNN (Eq. (17))
    Compute discriminative posterior probabilities (Eq. (19))
    Estimate $\tilde{\boldsymbol{S}}_t$ and $\tilde{\boldsymbol{N}}_t$ for all $t$ (Eqs. (7) and (8))
    Estimate parameters (Eqs. (12) to (15))
until convergence is achieved
Output $\tilde{\boldsymbol{S}}_t$ at final iteration
```

5. EXPERIMENTS

5.1. Experimental setup

We evaluated the proposed method on the AURORA2 task [25]. AURORA2 consists of three evaluation sets, i.e., set A (four types of known additive noises with the same channel characteristic), set B (four types of unknown additive noises with the same channel characteristic), and set C (one known and one unknown additive noise with the different channel characteristics). In this evaluation, set A was used as the development set and sets B and C were used as the evaluation sets.

The feature parameters for feature enhancement were 24 LMFBs that were extracted by using a Hamming window with a 25 ms frame length and a 10 ms frame shift. The SI clean speech GMM was trained by using the AURORA2 clean training data. The GMM had $K = 512$ Gaussian components. The number of Gaussian components of the noise GMM was set at $L = 1, \cdots, 4$. The parameter U was set at $U = 10$. Then, we also trained a DNN for the proposed feature enhancement by using multi-condition training data. The feature parameters of the DNN were utterance-wise mean and variance normalized 24 LMFBs and their first and second order derivatives. A context window with 11 frames was applied to each utterance. We trained five DNNs by changing the number of hidden layers with $M = 1, \cdots, 5$. Each hidden layer had 2048 nodes and



(a) Clean training

(b) Multi-condition training

Fig. 1. Average WERs for the proposed method with the development set

the output layer had $K = 512$ nodes, which correspond to the Gaussian components contained in the clean speech GMM. We trained each hidden layer with restricted Boltzmann machine (RBM)-based unsupervised pre-training. After the pre-training, the DNNs were obtained by supervised fine-tuning with state alignment labels of Eq. (18).

For comparative evaluation, we also trained a DAE that employs pre-trained RBMs of DNNs for feature enhancement. The number of hidden layers of DAE is same as the number of DNNs used for feature enhancement.

The ASR evaluations were carried out by using a DNN-HMM system. In the training stage, we firstly built a GMM-HMM system with the clean training data of AURORA2. The GMM-HMMs consisted of the 16-state word HMMs and the 3-state silence HMM. Thus, there were a total of 179 HMM states. Each state had 20 Gaussian components. The feature parameters of the GMM-HMMs consisted of 13 MFCCs (including the zero-th MFCC) and their first and second order derivatives. Mean and variance normalization was applied to each utterance. With these GMM-HMMs, we applied a state alignment to clean training data. Since the utterance content of AURORA2 is the same for clean and multi-condition training data, we used the state alignment labels of the clean training data for both clean and multi-condition DNNs. After state alignment, we built a DNN-HMM system for both clean and multi-condition training data. Each DNN consisted of five hidden layers. The feature parameters and topology of the hidden layer were same as DNNs for feature enhancement; however, the output layer had 179 nodes that correspond to the GMM-HMM states.

The evaluation criterion for ASR was the word error rate (WER).

5.2. Experimental results

In the evaluations, we compared three methods, namely DAE, our previous method with the GMM posterior probability described in Sec. 3 ("MMSE-joint-GMM"), and our proposed method with the DNN discriminative posterior probability described in Sec. 4 ("MMSE-joint-DNN"). The adjustable parameters of each method, i.e., the numbers of hidden layers M and Gaussian components in the noise GMM L, were adjusted by using the development set (set A). Fig. 1 shows the average WER of the proposed method with various parameter values. As seen in the figure, the WER tends to improve when M is increased, whereas it degrades when L is increased. With $L = 4$, the WER seriously degraded due to the over-fitting of the noise GMM. Thus, a deep structured DNN and a simple structured noise GMM are suitable for the proposed method.

Table 1. Adjusted parameters for each method

Training condition	Clean		Multi-condition	
Parameter	M	L	M	L
DAE	1	—	1	—
MMSE-joint-GMM	—	3	—	1
MMSE-joint-DNN	5	2	5	1

Table 2. ASR results for the clean training task with the average WER (%)

Data set	Set A	Set B	Set C	Avg.
Baseline	21.34	18.69	21.18	20.25
DAE	9.19	9.85	9.67	9.55
MMSE-joint-GMM	12.78	11.87	14.66	12.79
MMSE-joint-DNN	**8.06**	**9.19**	**9.13**	**8.72**

Table 3. ASR results for the multi-condition training task with the average WER (%)

Data set	Set A	Set B	Set C	Avg.
Baseline	5.23	6.42	5.61	5.78
DAE	5.95	7.29	6.58	6.62
MMSE-joint-GMM	6.08	7.08	6.41	6.54
MMSE-joint-DNN	**4.78**	**6.37**	**5.38**	**5.53**

Table 1 shows the values of the adjusted parameters for each method and each training condition. We used these values to perform ASR experiments with evaluation sets (sets B and C).

Tables 2 and 3 show the average WER of each training condition and each data set. As seen in Table 2, the proposed method "MMSE-joint-DNN" indicated remarkable improvements compared with the previous method "MMSE-joint-GMM" with the clean training. By this comparison, we can confirm that the use of discriminative posterior probability is indispensable for accurate feature enhancement and model parameter estimation. On the other hand, the proposed method also showed improvement compared with DAE. DAE trains the denoising transformation with stereo-data in advance. However, it is difficult to adapt the DAE to the utterance-wise fluctuations of a speaker and noise environment, because the DNN-based approach including DAE has many more parameters than a GMM. In contrast, the proposed method has utterance-wise unsupervised joint speaker adaptation and a noise GMM estimation scheme through its use of generative model approach. The equipment of this architecture is a great advantage for DAE. Therefore, these results justify the effectiveness of the proposed generative-discriminative hybrid approach.

With multi-condition training, as seen in Table 3, the results obtained with the conventional methods deteriorated compared with baseline system results, whereas the those obtained with the proposed method improved slightly. To achieve a significant improvement, we will investigate ways of introducing noise adaptive training [23], noise aware training [24], or another additional technique.

6. CONCLUSIONS

This paper presented a generative-discriminative hybrid approach for model-based feature enhancement. The proposed method applies the generative model approach to feature enhancement and parameter estimation by using GMMs, and computes the posterior probability with a discriminative model approach by using DNNs. The evaluation results showed that the proposed method provides significant improvements compared with the conventional technique with the generative model approach. In this paper, posterior probabilities w.r.t. the noise model were given by a generative model approach. In future, we plan to introduce a discriminative approach to unsupervised noise modeling.

7. REFERENCES

[1] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[2] K. Ishizuka and T. Nakatani, "A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1447–1457, November 2006.

[3] J. C. Segura, M.C. Benítez, A. de la Torre, A. M. Peinado, and A. Rubio, "Non-linear transformations of the feature space for robust speech recognition," in *Proc. of ICASSP '02*, May 2002, vol. I, pp. 401–404.

[4] V. Digalakis, D. Ritischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. on SAP*, vol. 3, no. 5, pp. 357–366, September 1995.

[5] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. of Interspeech '06*, September 2006, pp. 2286–2289.

[6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on ASSP*, vol. 27, no. 2, pp. 113–120, April 1979.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on ASSP*, vol. 32, pp. 1109–1121, December 1984.

[8] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. of ICASSP '96*, May 1996, vol. II, pp. 733–736.

[9] M. Fujimoto and T. Nakatani, "A reliable data selection for model-based noise suppression using unsupervised joint speaker adaptation and noise model estimation," in *Proc. of ICSPCC '12*, August 2012, pp. 4713–4716.

[10] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Extracting deep neural network bottleneck features using low-rank matrix factorization," in *Proc. of ICASSP "14*, May 2014, pp. 185–189.

[11] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. of Interspeech '13*, August 2013, pp. 436–440.

[12] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencorders for noisy reverberant speech recognition," in *Proc. of ICASSP '14*, May 2014, pp. 1778–1782.

[13] A. L. Maas, Q. V. Le, T. M. O 'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. of Interspeech '12*, September 2012.

[14] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of ICASSP '13*, May 2013, pp. 7092–7096.

[15] B. Li and K. C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Proc. of ASRU '13*, December 2013, pp. 279–284.

[16] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on SAP*, vol. 4, no. 5, pp. 352–359, May 1996.

[17] R. C. van Dalen and M. J. F Gales, "Extended VTS for noise-robust speech recognition," *IEEE Trans. on SAP*, vol. 19, no. 4, pp. 733–743, May 2011.

[18] K. K. Chin, H. Xu, M. J. F. Gales, C. Breslin, and K. Knill, "Rapid joint speaker and noise compensation for robust speech recognition," in *Proc. of ICASSP '11*, May 2011, pp. 5500–5503.

[19] Y. Q. Wang and M. J. F. Gales, "Speaker and noise factorisation on the AURORA4 task," in *Proc. of ICASSP '11*, May 2011, pp. 4584–4587.

[20] C. L. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, April 1995.

[21] O. Siohan, T. Myrvoll, and C. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech & Language*, vol. 16, no. 1, pp. 5–24, January 2002.

[22] S. Watanabe, A. Nakamura, and B. H. Juang, "Bayesian linear regression for hidden Markov model based on optimizing variational bounds," in *Proc. of MLSP '11*, December 2011, pp. 1–6.

[23] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. of ICASSP'14*, May 2014, pp. 2523–2527.

[24] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP '13*, May 2013, pp. 7398–7402.

[25] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR'00*, September 2000, pp. 18–20.

[26] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," in *Proc. of Interspeech '05*, September 2005, pp. 989–992.

[27] Y. Kashiwagi, D. Saito, N. Minematsu, and K. Hirose, "Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition," in *Proc. of ASRU '13*, December 2013, pp. 330–335.

[28] M. G. Rahim and B. H. Juang., "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on SAP*, vol. 4, no. 1, pp. 19–30, January 1996.