*INTERSPEECH 2016*
September 8–12, 2016, San Francisco, USA

# DNN-based Amplitude and Phase Feature Enhancement for Noise Robust Speaker Identification

*Zeyan Oo*[1], *Yuta Kawakami*[1], *Longbiao Wang*[1], *Seiichi Nakagawa*[2], *Xiong Xiao*[3], *Masahiro Iwahashi*[1]

[1]Nagaoka University of Technology,
[2]Toyohashi University of Technology,
[3]Nanyang Technological University
wang@vos.nagaokaut.ac.jp

## Abstract

The importance of the phase information of speech signal is gathering attention. Many researches indicate system combination of the amplitude and phase features is effective for improving speaker recognition performance under noisy environments. On the other hand, speech enhancement approach is taken usually to reduce the influence of noises. However, this approach only enhances the amplitude spectrum, therefor noisy phase spectrum is used for reconstructing the estimated signal. Recent years, DNN based feature enhancement is studied intensively for robust speech processing. This approach is expected to be effective also for phase-based feature. In this paper, we propose feature space enhancement of amplitude and phase features using deep neural network (DNN) for speaker identification. We used mel-frequency cepstral coefficients as an amplitude feature, and modified group delay cepstral coefficients as a phase feature. Simultaneous enhancement of amplitude and phase based feature was effective, and it achieved about 24% relative error reduction comparing with individual feature enhancement.

**Index Terms**: speaker identification, feature enhancement, deep neural network, phase information

## 1. Introduction

Today, the performance of speaker recognition system is extremely high in clean conditions. However, in the real conditions, the performance is significantly degraded by environmental noise. Speech enhancement approach (i.e. Wiener filtering) is taken usually for noise robust speech processing. However, the phase spectrum cannot be enhanced by such methods, unlike the amplitude spectrum, therefore this approach has not been applied to the phase based processing [1][2].

In recent years, the importance of the phase information is attracting attention [1]. Because of its complicated structure, the phase spectrum of the speech is ignored in many applications such as speaker recognition. Nakagawa et al. and Wang et al. proposed phase normalization method which expresses the phase difference from base-phase value [3]-[8], and this is called relative phase. Relative phase features were effective for speaker recognition under noisy environments with combination with amplitude feature (Mel-Frequency Cepstral Coefficients: MFCC) [9] because of its complementarity. To manipulate the phase information more simply, the group delay which is defined as the frequency derivative of the phase spectrum is often used. Hegde *et al.* proposed modified group delay cepstrral coefficients (MGDCC) [10]-[15]. They reported the MGDCC was effective for speaker recognition under noisy environments. As stated above, the phase information is considered significant even in the noisy environments.

However, the phase information had been ignored at the enhancement approach. For example, even in the state-of-the-art speech enhancement method, the phase spectrum of the noisy speech is used for signal reconstruction [2][17]. In this context, the iterative phase estimation method called Griffin and Lim algorithm was proposed by Griffin *et al.* for signal reconstruction [22][23]. This algorithm requires a huge number of iterative FFT, hence this approach is not realistic. On the other hand, the feature space enhancement method has been developed which is based on deep neural network technique [16]-[20]. DNN can learn the nonlinear transformation from a noisy feature vectors to clean ones. Zhang *et al.* applied DNN-based feature transformation for reverberant speaker recognition [18]. They transformed reverberant MFCC to clean MFCC, then the speaker recognition performance was improved. However, MFCC only contains amplitude information and ignores the phase, therefore the DNN enhancement might be incomplete. Evidently, Weninger *et al.* proposed a phase-sensitive error function for deep LSTM speech enhancement, and the method was effective [21]. However, they did not estimate phase of clean signal.

In this paper, we propose feature space enhancement using DNN for phase based feature. The phase based features could not be used effectively in noisy environments so far, however, DNN based feature enhancement approach might be effective because of its nonlinearity. In addition, we propose joint feature enhancement by DNN. The DNN is expected to be able to use both amplitude and phase information simultaneously in one network. By covering each information, the feature enhancement is expected to be more accurate.

The remainder of this paper is organized as follows: Section 2 presents the method of joint feature enhancement using DNN. Section 3 introduces the modified group delay feature extraction. The experimental setup and results are described in Section 4, and Section 5 presents our conclusions.

## 2. DNN based Phase Feature Enhancement

### 2.1. Conventional DNN-based amplitude feature enhancement

Neural networks are universal mapping functions that could be used for both classification and regression problem. Deep neural network has been used for speech enhancement scheme for quite some time. Fig. 1(a) shows the basic scheme of feature enhancement using DNN. The network is trained to minimize mean square error function between the output features and the



http://dx.doi.org/10.21437/Interspeech.2016-717

target features.

$$E_r = \frac{1}{N}\sum_{n=1}^{N} ||\hat{\mathbf{X}}_n(\mathbf{Y}_{n-\tau}^{n+\tau}, \mathbf{W}, \mathbf{b}) - \mathbf{X}_n||_2^2. \quad (1)$$

Here, $\mathbf{X}_n$ indicates the reference (clean) feature, $\hat{\mathbf{X}}_n$ denotes the estimated feature, $\mathbf{Y}_{n-\tau}^{n+\tau}$ is input noisy feature which spliced at $\pm\tau$ context frames, $\mathbf{W}$ denotes the weight matrices, $\mathbf{b}$ indicates bias vectors. To predict the clean features from the corrupted features a sequence of feature vectors around the current frame are fed into the DNN. This allows DNN to utilize the context information to predict the clean feature vector. Then, the DNN parameters $\mathbf{W}, \mathbf{b}$ are estimated iteratively by stochastic gradient decent (SGD) using the update equation below.

$$\Delta(\mathbf{W}_{n+1}, \mathbf{b}_{n+1}) = \quad (2)$$
$$-\lambda\frac{\partial E_r}{\partial(\mathbf{W}_n, \mathbf{b}_n)} - \kappa\lambda(\mathbf{W}_n, \mathbf{b}_n) + \omega\Delta(\mathbf{W}_n, \mathbf{b}_n)$$

Here, $n$ denotes the number of update iteration, $\lambda$ indicates the learning rate, $\kappa$ is weight decay coefficient, and $\omega$ is momentum coefficient. This supervised training step often called fine-tuning. To obtain the initial parameters of the network, RBM (restricted Boltzmann machine) based unsupervised pretraining is applied. In [18], the DNN based feature enhancement was successfully applied to MFCC in reverberant robust speaker identification. However, MFCC only contains the amplitude information of the speech, therefore the feature enhancement might be incomplete.

### 2.2. Simultaneous Enhancement of Amplitude and Phase feature

In [10], the robustness of the phase based feature (modified group delay cepstral coefficients: MGDCC) is reported. DNN based feature enhancement is expected to be effective also for phase based feature. However, phase based features contain less (or no) amplitude information, therefor the enhancement would be incomplete same as mentioned at 2.1. On the other hand, augmentation different features with the corresponding speech feature could improve the performance of the DNN training. This can be seen in improvement in performance in noise aware training [12][13]. Another research based in augmentation microphone distant information in speech recognition task has also provide with promising result [14].

With this in mind we have proposed the method in which phase features are augmented into the magnitude feature during the DNN training. Fig. 1(b) briefly shows the concept of the joint feature enhancement DNN. We try to enhance the amplitude and phase features simultaneously by concatenating two features as a input and reference vector, then the network is tuned to minimize the error of both amplitude and phase features. Phase information contain deep relationship with the magnitude feature, therefore we believe that DNN could utilize this deep relationship to improve the performance of the identification.

## 3. Amplitude and Phase-based features

In this work, we use two feature extraction methods to utilize both amplitude and phase information.

### 3.1. Mel-frequency cepstral coefficients (MFCC)

MFCC [9] is the most popular feature extraction method for speech processing including speaker identification. We used

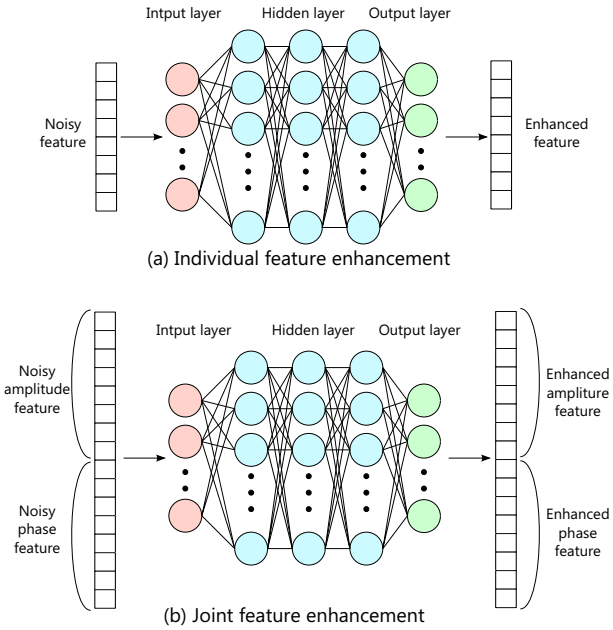


Figure 1: DNN feature enhancement for amplitude and phase features

MFCC as an amplitude feature for the DNN input.

### 3.2. Modified group delay feature

The phase spectrum can be obtained by applying $\tan^{-1}(\cdot)$ function. However, the phase values are stuffed into $(-\pi \leq \theta \leq \pi)$ range by $\tan^{-1}(\cdot)$, and the phase spectrum becomes like a noise. This problem is called phase wrapping. To overcome this problem, several phase processing methods are proposed, and some are applied to speaker identification. The group delay spectrum is the most popular method to manipulate phase information. Group delay $\tau_x(\omega)$ is defined as the frequency differential of the phase spectrum, and it can avoid phase wrapping problem because $\tan^{-1}$ is not required.

$$\tau_x(\omega) = -\frac{d}{d\omega}\angle X(\omega) \quad (3)$$
$$= -\mathrm{Im}\left(\frac{d}{d\omega}\log\left(X(\omega)\right)\right) \quad (4)$$
$$= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \quad (5)$$

Here, $X(\omega)$ is the Fourier transform of the signal $x(n)$, $Y(\omega)$ denotes the Fourier transform of $nx(n)$, footnote "R" and "I" indicates the real and imaginary part of the complex. Focusing on the denominator of eq.(5), the value of $\tau_x(\omega)$ would explode as $|X(\omega)|$ approximating to zero. Instead of $|X(\omega)|$, modified group delay defined as eq.(7) has smoothed $|X(\omega)|$ as the denominator.

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right)(|\tau(\omega)|)^{\alpha} \quad (6)$$
$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}} \quad (7)$$

Table 1: *Analysis conditions for MFCC and MGDCC*

| | MFCC | MGDCC |
|---|---|---|
| Frame length | 25 ms | |
| Frame shift | 5 ms | |
| FFT size | 512 samples | |
| Dimensions | 39 (13 MFCCs, 13 $\Delta$s, and 13$\Delta\Delta$s) | 39 (Lower 39 points of the cepstral coefficients) |

Here, $S(\omega)$ is cepstrally smoothed $X(\omega)$. The range of $\alpha$ and $\gamma$ are $(0 < \alpha \leq 1.0), (0 < \gamma \leq 1.0)$, in this paper, $\alpha = 0.4, \gamma = 0.9$ are used referring [10]. In the experiments, cepstral coefficients of the $\tau_m(\omega)$ (=MGDCC) is used as feature parameter by applying DCT. [10] reported that the MGDCC was effective for speaker identification in noisy environments.

# 4. Experiments

## 4.1. Experimental setup

We evaluate our proposed method for speaker recognition using artificial noisy speech. To obtain the noisy speech, clean speech was added with noise. Speech of the JNAS (Japanese Newspaper Article Sentence) database [25] is used as clean speech. The JNAS corpus consists of the recordings of 270 speakers (135 males and 135 females). The input speech was sampled at 16 kHz. The average duration of the sentences was approximately 3.5 seconds. Noise from JEIDA Noise Database [26] is used as background noise to create artificial noisy speech.. 4 noise kinds (air conductor, station, elevator hall, duct), with 4 SNRs (3, 9, 15, 21 dB) were used for multi-condition training, and 4 noise kinds (computer room, exhibition hall, bubble, road), with 3 SNRs (0, 10, 20 dB) were used for evaluation. Fig. 2 briefly shows the flow of the experiments. Each speaker was modeled as 256 mixture multi-condition GMM. 160 sentences (10 clean sentences $\times$ 16 training conditions) were used as training data for each speaker. 10 other sentences with evaluation noise were used as test data. In total, the test corpus consisted of about 2700 (10$\times$270) trials for each test condition. The GMM likelihood from different kind of features are combined linearly by following equation.

$$L^n_{comb} = \alpha L^n_{MFCC} + (1-\alpha) L^n_{MGDCC}, \quad (8)$$

$$\alpha = \frac{L^n_{MFCC}}{L^n_{MFCC} + L^n_{MGDCC}}.$$

Here, $n$ indicates the speaker index. The feature extraction conditions are shown in Table 1.

For DNN training, multi-condition speech data of all 270 speakers are used. DNN has 3 sigmoid hidden layers and linear output layer, each hidden layer contains 1024 nodes, and input features were spliced $\pm 5$ frames. Sigmoid type hidden layer is used here except for the input layer in which linear hidden unit were used. To train model for speech enhancement approach we have done unsupervised RBM (Restricted Boltzmann Machine) pretraining based on and supervised fine-tuning. To fasten up the training we first perform RBM wise pretraining. Kaldi toolkit is used for the pretraining task. The layers are trained by layer-wise greedy fashions to maximize the likelihood over the training sample. The pretraining only requires the corrupted version of the utterance. For the back propagation

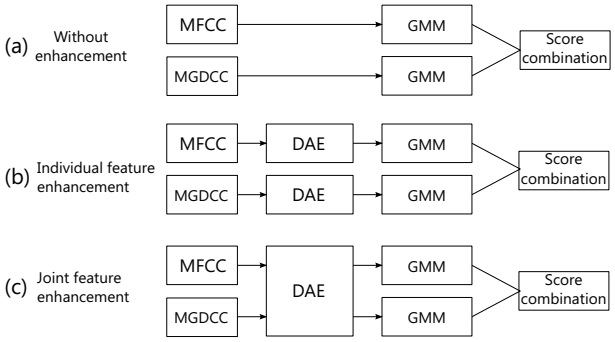


Figure 2: The flow of speaker identification experiments

to train the DNN parallel data consisting clean and distorted version of the same utterance. The objective of this training is to minimize the Mean Square Error (MSE) between the features. Stochastic gradient decent algorithm is used to improve the MSE error function. In the fine-tuning, the learning rate $\lambda$ was 0.01, the weight decay coefficient $\kappa$ was 0.5, and the momentum $\omega$ was 0.5.

## 4.2. Experimental results

Fig. 3 shows the feature spectrograms of MFCC and MGDCC by each enhancement method. Comparing (c) with (d), individual enhancement illustrated its performance for MFCC. Similarly, (g) and (h) shows the effectiveness of MGDCC enhancement. Moreover, comparing (d) with (e), joint method enhanced slightly better, and the same tendency can be found in (h) and (i).

Table 2 shows the experimental results in speaker identification accuracy. Raw indicates no enhancement, enhanced (individual) means individual feature DNN enhancement, and enhanced (joint) means simultaneous enhancement of amplitude and phase feature. MFCC + MGDCC means the speaker identification accuracy by the score combination. Without enhancement, speaker identification accuracy by using MFCC exceeded that of MGDCC, however, the score combination of them was effective. This shows the complementarity of the amplitude and phase features at speaker identification stage.

By applying individual feature enhancement, the speaker identification accuracies using each feature were improved. Therefore the DNN enhancement was effective not only for amplitude-based feature, but also for phase-based feature (MGDCC). However, DNN in this experiment only considers amplitude or phase independently, so we believe the method is not appropriate to use the whole of useful information.

When joint feature enhancement was applied to amplitude and phase based feature, the speaker identification accuracies were greatly improved. Focusing on MFCC, the relative error reduction of individual feature enhancement was about 15% (77.5% to 80.8%), and that of joint feature enhancement was about 37% (77.5% to 85.8%). The similar tendency of accuracy improvement is shown also for MGDCC. This is because the DNN could use both amplitude and phase information for the enhancement, and hence more accurate clean features were estimated. At last, the combination of joint enhanced MFCC and MGDCC achieved the best performance. This result is based on the complementarity of the amplitude and phase features at different stages; speaker modeling and feature enhancement.

Table 2: Speaker identification results by each enhancement method (%)

| | feature | 0 dB | | | | 10 dB | | | | 20 dB | | | | ave. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | bubble | road | server | exhibition | bubble | road | server | exhibition | bubble | road | server | exhibition | |
| raw | MFCC | 81.5 | 31.4 | 6.3 | 63.9 | 97.3 | 88.1 | 83.0 | 94.9 | 97.0 | 94.8 | 95.9 | 95.9 | 77.5 |
| | MGDCC | 33.9 | 17.2 | 7.1 | 54.9 | 95.2 | 95.2 | 80.9 | 97.2 | 96.7 | 97.3 | 97.6 | 96.6 | 72.5 |
| | MFCC+ MGDCC | 66.5 | 8.1 | 29.3 | 73.2 | 97.1 | 92.8 | 96.3 | 98.0 | 98.4 | 98.8 | 98.2 | 98.1 | 79.6 |
| enhanced (individual) | MFCC | 85.9 | 46.1 | 18.7 | 79.5 | 95.9 | 88.9 | 82.6 | 96.1 | 95.5 | 93.4 | 90.6 | 96.4 | 80.8 |
| | MGDCC | 58.2 | 36.7 | 17.2 | 46.8 | 94.1 | 94.8 | 75.2 | 94.1 | 96.1 | 97.4 | 96.4 | 96.7 | 75.3 |
| | MFCC+ MGDCC | 81.4 | 24.1 | 49.6 | 69.7 | 96.1 | 89.5 | 96.1 | 97.3 | 97.7 | 97.6 | 98.3 | 98.2 | 83.0 |
| enhanced (joint) | MFCC | 88.2 | 62.0 | 33.5 | 78.0 | 96.9 | 94.8 | 91.1 | 97.0 | 97.2 | 97.1 | 96.3 | 97.8 | **85.8** |
| | MGDCC | 76.2 | 51.5 | 24.7 | 77.8 | 94.6 | 94.7 | 84.0 | 97.6 | 96.7 | 97.6 | 97.8 | 98.4 | **82.6** |
| | MFCC+ MGDCC | 85.8 | 37.1 | 61.9 | 83.3 | 96.7 | 92.7 | 96.7 | 98.2 | 97.6 | 98.2 | 98.7 | 98.8 | **87.1** |

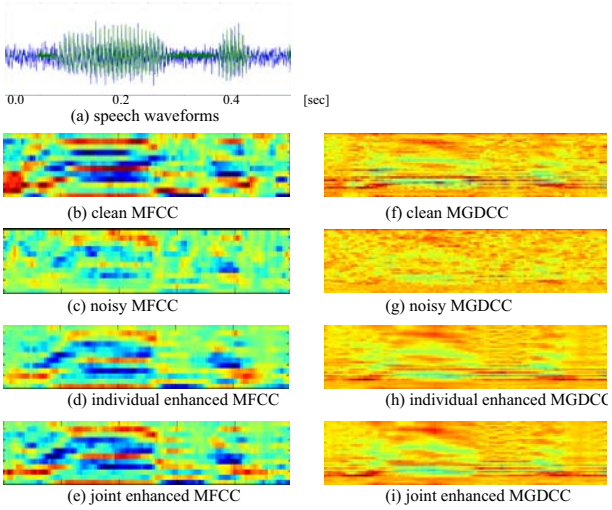


Figure 3: The spectrograms of each enhancement method: (a) green line is clean speech, blue is 0 dB noisy speech

## 5. Conclusions

In this paper, we proposed feature space enhancement using DNN for amplitude and phase based feature. Simultaneous feature enhancement of amplitude and phase features by DNN was evaluated on the experiments. We confirmed the effectiveness of the DNN based feature enhancement for the phase-based feature(MGDCC). In addition, the speaker identification performance by joint feature enhancement exceeded that of the individual enhancement. This is because the feature enhancement got more accurate by covering each information in the network.

In our future work, the more suitable network should be applied for speaker identification task. For example, multi-task training (feature enhancement + speaker identification) of DNN might be effective.

## 6. Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 15K16020.

## 7. References

[1] P. Mowlaee, R. Saeidi, and Y. Stylianou, "INTERSPEECH 2014 Special Session: Phase Importance in Speech Processing Applications", Proc. Interspeech, pp. 1623-1627, 2014.

[2] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase Processing for Single-Channel Speech Enhancement", IEEE Signal Processing Magazine, pp. 55-66, 2015.

[3] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker Recognition by Combining MFCC and Phase Information", Proc. Interspeech, pp. 2005-2008, 2007.

[4] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information", IEEE Trans. on Audio, Speech and Language Processing, vol. 20 no. 4, pp. 1085-1095, 2012.

[5] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker Recognition by Combining MFCC and Phase Information in Noisy Conditions", IEICE Trans. Inf. & Syst. , Vol. E93-D, No.9, pp. 2397-2406, 2010.

[6] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments", Proc. on ICASSP, pp. 4502-4505, 2010.

[7] L. Wang, S. Ohtsuka and S. Nakagawa, "High improvement of speaker identification and verification by combining MFCC and phase information", Proc. on ICASSP, pp. 4529-4532, 2009.

[8] L. Wang, Y. Yoshida, Y. Kawakami and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech", Proc. Interspeech, pp. 2092-2096, 2015.

[9] S. Davis, B. Santa, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 28, Issue 4, pp. 357-366, 1980.

[10] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the Modified Group Delay Feature in Speech Recognition", IEEE Trans. on Audio, Speech, and Language Processing, Vol. 15, No.1, pp. 190-202, 2007.

[11] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief networks", Neural Computation, vol. 18, issue 7, pp. 1527-1554, 2006.

[12] M. L. Seltzer D. Yu, and Y. Wang, "An Investigation of Deep Neural Networks for Noise Robust Speech Recognition", Proc. ICASSP, pp. 7398-7402, 2013.

[13] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic Noise Aware Training for Speech Enhancement based on Deep Neural Networks", Proc. Interspeech, pp. 2970-2974, 2014.

[14] Y. Miao, and Florian Metze, "Distant Aware DNNs for Robust Speech Recognition", Proc. Interspeech, pp. 761-765, 2015.

[15] R. Padmanabhan, S. Parthasarathi, and H. Murthy, "Robustness of phase based features for speaker recognition", Proc. Interspeech, pp.2355-2358, 2009.

[16] X.-G. Lu, Y. Tsao, S. Matsuda, and C. Hori, Speech enhancement based on deep denoising Auto-Encoder," Proc. Interspeech, pp. 436440, 2013.

[17] Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, and M. Iwahashi, "Deep Neural Network-based Bottleneck Feature and Denoising Autoencoder-based Dereverberation for Distant-talking Speaker Identification", Eurasip Journal on Audio, Speech, and Music Processing, 2015:12, 2015.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks", IEEE Trans. on Audio, Speech and Language Processing, vol. 23, No. 1, 2015.

[19] Y. Ueda, L. Wang, A. Kai and B. Ren, "Environment-dependent denoising autoencoder for distant-talking speech recognition", Eurasip Journal on Advances in Signal Processing, 2015:92, 2015.

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

[21] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise Robust ASR", Latent Variable Analysis and Signal Separation, pp. 91-99, 2015.

[22] D. Griffin, and J. Lim, "Signal Estimation from Modified Short-Time Fourier Transform", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No.2, pp. 236-242, 1984.

[23] J. L. Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast Signal Reconstruction From Magnitude STFT Spectrogram Based on Spectrogram Consistency", Proc. of the 13th Int. Conference on Digital Audio Effects, pp. 397-403, 2010.

[24] X. Zhao, Y. Wang, and D. Wang, "Robust Speaker Identification in Noisy and Reverberant Conditions", Proc. ICASSP, pp. 4025-4029, 2014.

[25] K. Itou, M. Yamamoto, K. Takeda, T. Kakezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, " J. Acoust. Soc. Jpn (E), Vol. 20, pp. 199-206, 1999.

[26] I. Shuichi, "On recent speech corpora activities in Japan", Journal of the Acoustical Society of Japan (E), Vol. 20 (1999) No. 3, pp. 163-169, 1999.