EXPLORING MULTI-CHANNEL FEATURES FOR DENOISING-AUTOENCODER-BASED SPEECH ENHANCEMENT

Shoko Araki† *Tomoki Hayashi*†,‡ *Marc Delcroix*† *Masakiyo Fujimoto*† *Kazuya Takeda*‡ *Tomohiro Nakatani*†

† NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
‡ Department of Media Science, Nagoya University
1 Furo-cho, Chikusa-ku, Nagoya-shi, Aichi 464-0814, Japan

ABSTRACT

This paper investigates a multi-channel denoising autoencoder (DAE)-based speech enhancement approach. In recent years, deep neural network (DNN)-based monaural speech enhancement and robust automatic speech recognition (ASR) approaches have attracted much attention due to their high performance. Although multi-channel speech enhancement usually outperforms single channel approaches, there has been little research on the use of multi-channel processing in the context of DAE. In this paper, we explore the use of several multi-channel features as DAE input to confirm whether multi-channel information can improve performance. Experimental results show that certain multi-channel features outperform both a monaural DAE and a conventional time-frequency-mask-based speech enhancement method.

Index Terms— Deep learning, denoising autoencoder, multi-channel noise suppression, PASCAL 'CHiME' challenge

1. INTRODUCTION

Speech enhancement techniques including noise suppression, separation, and dereverberation, have been developed over many years [1–4]. In Particular, multi-channel approaches including beamforming (e.g., [2]), source separation (e.g., [3, 5]) and multi-channel Wiener filter [6, 7], have been widely researched and employed in many practical applications. One reason for this is that the multi-channel approaches generally perform better than single-channel (monaural) approaches due to their ability to utilize information about speech source location, in addition to the speech and noise spectral information.

On the other hand, deep-learning-based monaural speech enhancement methods have recently come to the forefront. For example, Wang and Wang [8, 9] and Huang et al., [10] proposed employing deep neural networks (DNNs) for time-frequency mask estimation for speech separation. Moreover, some DNN-based approaches have also worked successfully for noise suppression [11] and dereverberation [12–14] for robust automatic speech recognition (ASR). To obtain enhanced features for ASR, some of these techniques [11, 12, 14] employ a denoising autoencoder (DAE) [15], which is a neural network that is trained to reconstruct a clean speech signal from its own noisy input. The DAE has been shown to have an excellent feature enhancement capability for ASR. In addition, there have been many successful DNN-based trials in the field of robust ASR, e.g., [16–20]. Among these, the authors of [16] have reported a notable result, namely that we can improve the ASR performance when we use the estimated noise features as additional information for DNN inputs. This is known as noise-aware training.

Inspired by the abovementioned successful results for the DAE approaches and noise-aware training, this paper explores the use of multi-channel features for a DAE, and discusses suitable features for multi-channel DAE-based speech enhancement. To the best of the authors' knowledge, almost all existing DAE-based speech enhancement proposals have concentrated on single-channel approaches, and there are only a few papers that discuss multi-channel features as inputs to a DAE. A few papers have tried using multi-channel features for DNN-based ASR systems: Narayanan and Wang [17] have explored a variety of features for monaural ASR, and [18] and [19] have reported the positive effect of certain multi-channel features, such as simple channel-concatenated features and enhancement speech with a beamforming, for DNN-based ASR inputs. On the other hand, we explore multi-channel features including channel-concatenation and noise-aware features for DAE-based speech enhancement, and investigate whether multi-channel information can improve performance.

The rest of this paper is organized as follows: Section 2 describes our multi-channel noise suppression task, and Sec. 3 reviews the DAE. In Sec. 4, we present our proposed multi-channel DAE, which employs multi-channel information for the DAE input. Section 5 reports the results of noise suppression experiments conducted by using the PASCAL 'CHiME' database, and shows the performance with each multi-channel feature. The last section concludes the paper.

2. PROBLEM FORMULATION

This paper considers a multi-channel noise suppression task. An M-channel noisy observation vector $\mathbf{Y}_{t,f} = [Y_{t,f}^{(1)}, \cdots, Y_{t,f}^{(M)}]$ is given as

$$\mathbf{Y}_{t,f} = \log\{\exp(\mathbf{X}_{t,f}) + \exp(\mathbf{N}_{t,f})\}, \quad (1)$$

where $X_{t,f}^{(m)}$ and $N_{t,f}^{(m)}$ are the clean speech and noise log mel filter-bank spectra, respectively. In this paper, we work in the log mel domain, therefore, t and f denote a time frame and a mel frequency. Our objective is to obtain the denoised speech feature $\hat{X}_{t,f}^{(m)}$ at a channel m, and then reconstruct the waveform of the denoised speech. We consider here the stereo case ($M = 2$).

For the sake of simplicity, we use the following two notations:
(i) We denote the first channel feature

$$y_{t,f} = Y_{t,f}^{(1)}$$

unless otherwise noted without loss of generality, and
(ii) a vector

$$\mathbf{y}_t = [y_{t,1}, \cdots, y_{t,f}, \cdots, y_{t,F}],$$

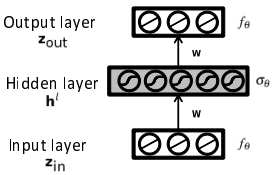

Fig. 1. Basic autoencoder

which denotes the log mel spectral feature vector of the first channel. The same vector notations are used for other variables, including clean speech $\mathbf{x}_t$ and denoised speech $\hat{\mathbf{x}}_t$.

3. DENOISING AUTOENCODER (DAE)

Let us first review autoencoders (AEs) as shown in Fig. 1. An AE is a neural network, which is trained to output the same vector $\mathbf{z}_{\text{out}}$ as its clean input vector $\mathbf{z}_{\text{in}}$. That is, the network learns parameters to minimize the squared error between its outputs and target input features. An AE is employed to pre-train a DNN and as a feature extractor for ASR.

Denoising autoencoder (DAE) [15, 21] is an extended version of an AE, where the input vector $\mathbf{z}_{\text{in}}$ is corrupted by noise, but a DAE is still trained to output a clean vector $\mathbf{z}_{\text{out}}$. That is, a DAE is a neural network that learns the mapping between noisy and clean speech. DAE parameters $\theta = \{\mathbf{W}, \mathbf{b}\}$ are used to calculate the output of the l-th hidden layer as

$$\mathbf{h}^l(\mathbf{z}_{\text{in}}) = \sigma_\theta(\mathbf{W}^l\mathbf{h}^{(l-1)}(\mathbf{z}_{\text{in}}) + \mathbf{b}^l),$$

where $\mathbf{W}^l$ and $\mathbf{b}^l$ are a weight matrix and a bias vector, respectively. The function σ_θ is a nonlinear activation function, and here we utilize a sigmoid function. The first hidden layer $\mathbf{h}^1$ is computed as: $\mathbf{h}^1(\mathbf{z}_{\text{in}}) = \sigma_\theta(\mathbf{W}^1\mathbf{z}_{\text{in}} + \mathbf{b}^1)$, and the output layer has a linear activation, i.e.,: $\mathbf{z}_{\text{out}} = f_\theta(\mathbf{z}_{\text{in}}) = \mathbf{W}^L\mathbf{h}^{(L-1)}(\mathbf{z}_{\text{in}}) + \mathbf{b}^L$.

The input vector $\mathbf{z}_{\text{in}}$ for a single channel (MONAURAL) DAE is obtained by concatenating several frames of the noisy observation features to account for left and right contexts, as

$$\mathbf{z}_{\text{in}} = [\mathbf{y}_{t-T}, \cdots, \mathbf{y}_t, \cdots, \mathbf{y}_{t+T}], \tag{2}$$

where T is the context window size.

The output vector of the DAE consists of clean speech with the same context as the input:

$$\mathbf{z}_{\text{out}} = [\mathbf{x}_{t-T}, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_{t+T}]. \tag{3}$$

A DAE is trained in advance using a large amount of noisy and clean stereo training data. During testing, an enhanced feature $\hat{\mathbf{x}}_t$ is obtained by propagating the noisy features through the DAE.

4. PROPOSED MULTI-CHANNEL DAE

The original DAE described in the previous section usually utilizes MONAURAL noisy observation features as input features $\mathbf{z}_{\text{in}}$ (eq. (2)). However, when we have multiple observations, it is natural to consider the possibility of utilizing additional features that can be obtained from the multi-channel observations. In this section, we explain multi-channel features that can be added to the input features of a conventional MONAURAL DAE, to take advantage of multi-channel information.

Figure 2 shows the proposed multi-channel DAE. As the input vector, in addition to the noisy observation vector $\mathbf{y}$, we also employ

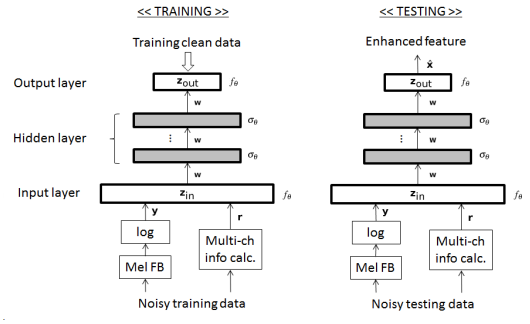

Fig. 2. Multi-channel DAE for speech enhancement

the multi-channel information $\mathbf{r}_t$. The input feature of the DAE thus becomes,

$$\mathbf{z}_{\text{in}} = [(\mathbf{y}_{t-T}, \mathbf{r}_{t-T}), \cdots, (\mathbf{y}_t, \mathbf{r}_t), \cdots, (\mathbf{y}_{t+T}, \mathbf{r}_{t+T})]. \tag{4}$$

We investigate several types of multi-channel information $\mathbf{r}_t = [r_{t,1}, \cdots, r_{t,f}, \cdots, r_{t,F}]$, which will be discussed in the following subsection.

The output vector for the DAE is the same vector as the original monaural DAE output (eq. (3)): that is, the output vector for the DAE training is the monaural clean speech vector.

4.1. Multi-channel information for DAE

Here, we explain several multi-channel features $\mathbf{r}_t$ used as our input vector (eq. (4)). We investigated the following features:

- **Interaural level difference (ILD):** The simplest multi-channel feature is the interaural level difference (ILD). That is, we consider the log mel feature of the second channel:

$$r_{t,f} = -Y^{(2)}_{t,f}. \tag{5}$$

  Because we are working in the log domain, eq. (4) with (5) represents the level ratio between channel observations[1]. Moreover, it is worth noting that the ILD information corresponds to the channel concatenation (e.g., [18]).

- **Interaural phase difference (IPD):** The second multi-channel feature is the interaural phase difference (IPD), which reflects the location information of the target and noise sources. Locational information is usually quite helpful for speech enhancement. In this paper, we utilize the phase difference $\phi_{f'(f)}$ at the center frequency f' of each mel filterbank,

$$r_{t,f} = \phi_{f'(f)} = \cos(\angle \bar{Y}^{(2)}_{t,f'(f)} - \angle \bar{Y}^{(1)}_{t,f'(f)}), \tag{6}$$

  where $f'(f)$ is the center frequency of the f-th bin of the mel filterbank, and $\bar{Y}_{t,f'}$ is the short-time Fourier transform (STFT) coefficient of noisy observation. The cosine function in eq. (6) is employed to wrap the phase information, whose range is limited to $-\pi \leq \phi \leq \pi$.

- **Mask for speech enhancement (SOFTMASK):** By using multi-channel information, we can estimate a time-frequency "mask" for speech enhancement. For example, by clustering the locational information (e.g., IPD) at each time-frequency

[1] $\log(\exp(Y^{(1)}_{t,f})/\exp(Y^{(2)}_{t,f})) = \log(\exp(Y^{(1)}_{t,f})) - \log(\exp(Y^{(2)}_{t,f}))$

slot, we can obtain a time-frequency mask that enhances target speech [5, 22, 23]. In this paper, we consider a mel-frequency domain softmask $M_{t,f}$ [23] which is obtained by transforming a locational feature clustering-based softmask in the STFT domain into the mel-frequency domain (for more detail, see eq. (40) in [23]). This paper defines the SOFTMASK in the log-mel domain as

$$r_{t,f} = \log(M_{t,f})^2. \tag{7}$$

In this paper, however, we do not use this feature directly, as we found that the DAE training with eq. (7) sometimes became unstable. Instead, we use ENHANCE feature defined in the following, which can be shown to be equivalent to the SOFTMASK feature as a DAE input.

- **Pre-enhanced speech with SOFTMASK (ENHANCE):** We can also employ the pre-enhanced speech feature $\hat{x}$, which is estimated with the abovementioned SOFTMASK as shown in footnote 1, that is,

$$r_{t,f} = y_{t,f} + \log(M_{t,f}). \tag{8}$$

It should be noted that the ENHANCE (eq. (8)) and the SOFTMASK (eq. (7)) are equivalent for a DAE when used with $\mathbf{y}_t$ in multi-channel inputs (eq. (4)). This is because the two types of multi-channel inputs can be transformed to the same values by applying an affine projection to the DAE inputs.

- **Estimated noise with SOFTMASK (NOISE):** In contrast to the ENHANCE feature, we can utilize an estimated noise with the SOFTMASK. The noise feature can be estimated by

$$r_{t,f} = y_{t,f} + \log(1 - M_{t,f}). \tag{9}$$

Note that this is equivalent to the multiplication of a noisy observation and a noise enhancement mask in the linear spectral domain. It should also be noted that this feature corresponds to a detailed version of the noise aware training [16], since we use feature level noise estimates whereas [16] uses utterance level noise estimates.

4.2. Enhanced sound reconstruction

After obtaining the enhanced features $\hat{\mathbf{x}}_t$ from the output of the DAE, we reconstruct a waveform of the enhanced target speech signal with the following procedure. First, we use the enhanced speech feature $\hat{x}_{t,f}$ to calculate the enhancement filter

$$W_{t,f} = \frac{\exp(\hat{x}_{t,f})}{\exp(y_{t,f})} \tag{10}$$

in the mel-frequency domain. Then, we transform the enhancement filter $W_{t,f}$ into the filter $\bar{W}_{t,f'}$ in the linear frequency (STFT) domain, where f' represents a linear frequency [24]. Then, the enhanced signal $\bar{x}_{t,f'}$ in STFT is obtained by multiplying the STFT coefficient of noisy observation $\bar{y}_{t,f'}$ and the filter $\bar{W}_{t,f'}$. Finally, the waveform of the enhanced signal is calculated by using the inverse Fourier transform of $\bar{x}_{t,f'}$.

[2] An enhanced speech feature $\hat{x}_{t,f}$ can be represented using the SOFTMASK as $\hat{x}_{t,f} = \log(M_{t,f}\exp(y_{t,f})) = \log(M_{t,f}) + y_{t,f}$. Note that $\hat{x}_{t,f}$ and $y_{t,f}$ are log spectra.

5. EXPERIMENTS

5.1. Experimental setups

We investigated multi-channel DAE performance using the PASCAL 'CHiME' challenge task [25]. The test utterances of the CHiME challenge consist of speech commands in the presence of living room noise that are recorded with two distant microphones. The commands were always spoken in front of the microphones with a fixed distance of 2 meters between the speaker and the microphones. Therefore, the locational variation of the target speech is limited in this task. On the other hand, the noise was collected in a real living room and it includes highly non-stationary noises, such as children's voices, vacuum cleaners and televisions.

5.1.1. DAE evaluation setup

We utilized a DAE with a single hidden layer of 1024 units. The clean and noisy speech features were 40 log mel filterbank coefficients *without* their delta and acceleration. The multi-channel features were also calculated in the mel frequency domain with 40 filterbanks. Here, 5 left and 5 right context frames were employed both for the DAE input and output layers. The window size and frame shift for DAE feature calculation were set at 100 ms and 25 ms, respectively.

The training data consist of utterances spoken by the 34 speakers and 6 hours of background noise data. The clean training data is noise free, but corrupted by reverberation.

We used the development set of the CHiME challenge database as the evaluation set, because the test set does not include clean reference signals and thus cannot be used for speech enhancement evaluation. The development set of the CHiME challenge includes 600 noisy utterances at SNRs ranging from -6 to 9 dB.

5.1.2. Evaluation metrics

As noise-free (reverberant) utterances are available for the development set, we can evaluate both the noise suppression performance and the command (keyword) recognition accuracy [25].

The noise suppression performance was evaluated in terms of the segmental signal-to-noise ratio (SSNR) and the cepstral distortion (CD).

For the keyword recognition, we utilized our DNN-based ASR back-end system [20], where we employed 40 log mel filterbank coefficients with their delta and acceleration, and 5 left and 5 right context windows as the DNN input. Note that the window size and frame shift were different for enhancement and recognition. For recognition we used standard settings of 25 msec and 10 msec for the window size and frame shift, respectively. The DNN structure had 6 hidden layers (2048 units each) and 254 output HMM states. We used two types of acoustic models; one trained with noise free data and one trained with multi-condition training data. In the latter case the amount of training data was 42 times that used for the noise-free case [25]. Note that we did not retrain the DNN used for recognition with the enhanced speech, and consequently the results obtained with the multi-condition model may not be optimal.

5.2. Experimental results

Figure 3 shows the average CD calculated over the 1st to 12th order cepstral coefficients and the average SSNR. Here, the entire training set of 34 speakers was utilized for DAE training, and the development set of all 34 speakers was evaluated. In the figure, bars labelled "SOFTMASK" and "MONAURAL" show the enhancement results with SOFTMASK based speech enhancement *without*

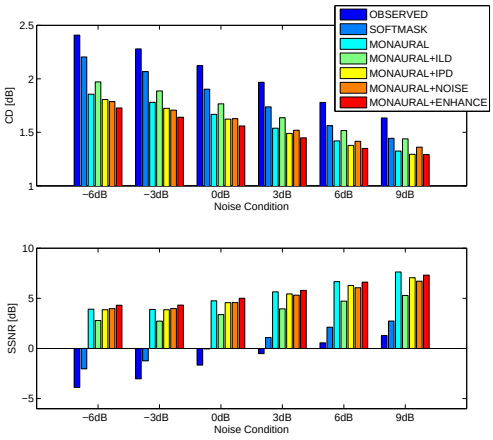

Fig. 3. Average CD and SSNR results.

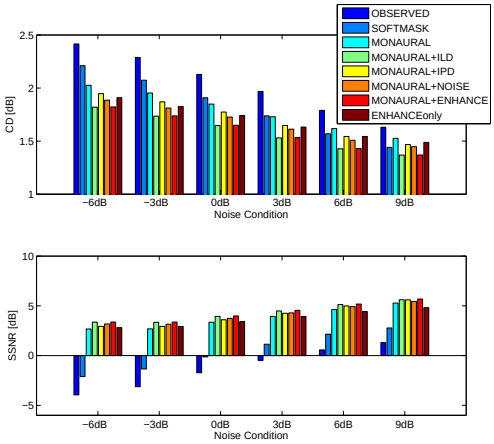

Fig. 4. Result for speaker-open development set.

a DAE, and with the monaural DAE (see Sec. 3), respectively. The results labelled "MONAURAL+ILD", "MONAURAL+IPD", "MONAURAL+NOISE", and "MONAURAL+ENHANCE" represent the multi-channel DAE performance with the corresponding multi-channel features. We can see that the multi-channel DAEs (except with the ILD feature) outperform the monaural DAE, and we obtain the best performance when we employ the "MONAURAL+ENHANCE" feature. In our experiments, the ILD feature did not work well. This is due to the recording setup of the CHiME challenge, where the commands were always spoken in front of the microphones, and the locational variation of the target speech was limited.

Table 1 (a) and (b) summarize the keyword recognition accuracy with (a) a noise-free acoustic model and (b) a multi-condition acoustic model. With a noise-free model (Tab. 1(a)), the DAEs with multi-channel features (except the ILD) provided greater improvement than the conventional SOFTMASK and the monaural DAE. With a multi-condition model (Tab. 1(b)), the superiority of the multi-channel DAE is not as such large as with the noise-free

Table 1. Keyword accuracy in %. "M+" means "MONAURAL".

(a) With noise-free model:

	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg.
OBSERVED	38.17	40.17	51.83	65.42	75.75	85.00	59.39
SOFTMASK	44.67	48.92	61.50	74.08	83.50	87.00	66.61
MONAURAL	63.33	65.08	74.87	81.67	86.50	89.75	76.83
M+ILD	59.00	62.83	71.50	79.58	83.58	87.33	73.97
M+IPD	64.83	68.42	76.25	82.67	88.33	90.25	78.46
M+NOISE	68.00	70.42	78.25	84.83	88.50	90.75	80.13
M+ENHANCE	**69.75**	**74.08**	**81.42**	**86.58**	**89.83**	**91.50**	**82.19**

(b) With multi-condition model:

	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg.
OBSERVED	81.00	86.25	90.08	93.50	94.83	**96.50**	90.36
SOFTMASK	**83.42**	**87.67**	91.00	93.00	95.25	96.00	91.06
MONAURAL	81.17	84.67	87.58	92.17	94.75	95.75	89.35
M+ILD	80.00	83.75	87.92	91.67	93.17	95.50	88.67
M+IPD	81.00	86.17	88.83	92.17	94.75	95.08	89.67
M+NOISE	82.42	86.00	90.75	93.50	95.00	96.25	90.65
M+ENHANCE	82.83	86.92	**91.25**	**93.92**	**95.42**	96.42	**91.13**

model[3]. However, the multi-channel DAE still works better than the monaural DAE, and we also obtained the best performance with the ENHANCE feature.

5.3. Discussion

The CHiME task employs the same speakers for the training and evaluation sets. Consequently, Sec. 5.2 presents results for a speaker closed case. To confirm the applicability of the multi-channel DAE to the speaker independent training, here we looked at the performance for a speaker-open case, where we trained the DAE with only 4 speakers taken from the 34 speakers of the training set and evaluated the performance of the 30 remaining speakers.

Figure 4 shows the average CD and SSNR results. Even for the open speaker test data, the multi-channel DAE still outperforms the monaural DAE. Moreover, we obtained the best performance with the "MONAURAL+ENHANCE" feature.

Next, let us compare the results of a DAE with MONAURAL and ENHANCE input features ("MONAURAL+ENHANCE") with a DAE that only uses ENHANCE features "ENHANCEDonly". As seen in Fig. 4, in our experiment the "ENHANCEDonly" feature performs worse than the "MONAURAL+ENHANCE" feature.

6. CONCLUSION

We investigated several multi-channel input features for a DAE, including the ILD information (=channel-concatenation), IPD, estimated noise (=noise-aware), and pre-enhanced speech signal. We found that some multi-channel features outperform the monaural input features in terms of cepstral distortion, segmental SNR and keyword recognition accuracy. We obtained the best performance when we input the monaural and pre-enhanced speech features together as the DAE input vector. Future work includes exploring the optimal number of hidden layers and units, and the optimal window and shift sizes. Additional evaluations for different databases (e.g., AURORA2/4 and AMI meeting data), and different tasks (e.g., speech separation) also remain for future investigation.

[3]Note that this may be due to the mismatch between the training and test conditions because we use a DNN trained on noisy speech. We expect that performance would increase if we would retrain the acoustic model using training data processed with the DAE front-end.

7. REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, 2005.

[2] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[3] S. Makino, T. W. Lee, and H. Sawada, Eds., *Blind Speech Separation*, Springer, 2007.

[4] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[6] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001.

[7] A. Spriet, M. Moonen, and J. Wouters, "The impact of speech detection errors on the noise reduction performance of multichannel Wiener filtering and generalized sidelobe cancellation," *Signal Processing*, vol. 85, no. 6, pp. 1073–1088, 2005.

[8] Y. Wang and D. L. Wang, "Cocktail party processing via structured prediction," in *Proc. of NIPS2012*, 2012, pp. 224–232.

[9] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[10] P. S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. of ICASSP2014*, 2014, pp. 1581–1585.

[11] 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 Interspeech2012*, 2012.

[12] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *Proc. of ICASSP2014*, 2014, pp. 4623–4627.

[13] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. of ICASSP2014*, 2014, pp. 4661–4665.

[14] X. Feng, Y. Zhang, and J. Grass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. of ICASSP2014*, 2014, pp. 1778–1782.

[15] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc of ICML2008*, 2008, pp. 1096–1103.

[16] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP2013*, 2013, pp. 7398–7402.

[17] A. Narayanan and D. L. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. of ICASSP2014*, 2014, pp. 2523–2527.

[18] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognitions," in *Proc. of ICASSP2014*, 2014, pp. 5579–5583.

[19] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Proc. of HSCMA2014*, 2014.

[20] M. Delcroix, Y. Kubo, T. Nakatani, and A. Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?," in *Proc. of Interspeech2013*, 2013, pp. 2992 –2996.

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[22] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[23] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 12, pp. 2516–2531, 2013.

[24] M. Fujimoto, S. Watanabe, and T. Nakatani, "A robust estimation method of noise mixture model for noise suppression," in *in Proc of Interspeech2011*, 2011, pp. 697–700.

[25] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.