SEMI-BLIND SPEECH ENHANCEMENT BASED ON RECURRENT NEURAL NETWORK FOR SOURCE SEPARATION AND DEREVERBERATION

Masaya Wake, Yoshiaki Bando, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

This paper describes a semi-blind speech enhancement method using a semi-blind recurrent neural network (SB-RNN) for human-robot speech interaction. When a robot interacts with a human using speech signals, the robot inputs not only audio signals recorded by its own microphone but also speech signals made by the robot itself, which can be used for semi-blind speech enhancement. The SB-RNN consists of cascaded two modules: a semi-blind source separation module and a blind dereverberation module. Each module has a recurrent layer to capture the temporal correlations of speech signals. The SB-RNN is trained in a manner of multi-task learning, i.e., isolated echoic speech signals are used as teacher signals for the output of the separation module in addition to isolated unechoic signals for the output of the dereverberation module. Experimental results showed that the source to distortion ratio was improved by 2.30 dB on average compared to a conventional method based on a semi-blind independent component analysis. The results also showed the effectiveness of modularization of the network, multi-task learning, the recurrent structure, and semi-blind source separation.

Index Terms— Semi-blind source separation, Blind dereverberation, Recurrent neural network

1. INTRODUCTION

Speech enhancement is indispensable for realizing smooth speech interaction between a human and a robot. When a human and a robot interacts using speech, the observation signals recorded by the microphone of the robot is composed of not only a direct sound of the human speech but also that of the robot's speech because the current robot often fails to perform smooth turn-taking and it cannot stop speaking once speech output is generated. The observation also contains reverberations of the human and robot's speeches. It is necessary to enhance the speech by separating human speech and removing the reverberation.

The speech enhancement task we address consists of two steps. The first step is semi-blind source separation: an observation sound which is a mixture of human and robot's echoic

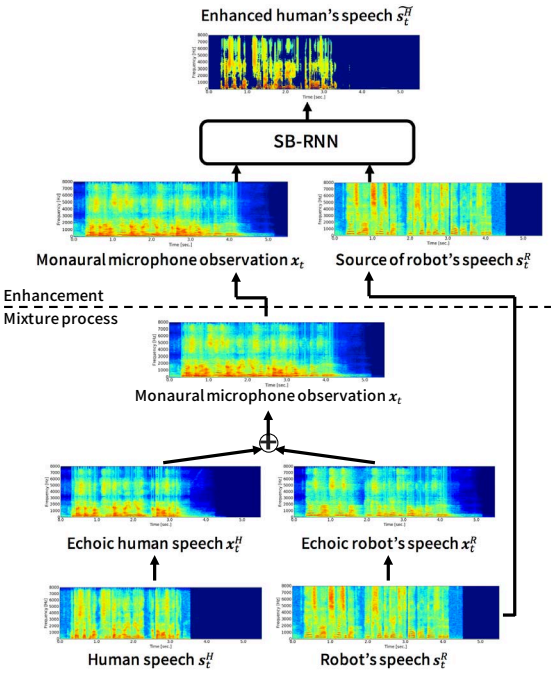


Fig. 1. The overview of the proposed method

speech signals is separated to their isolated speech signals while an anechoic signal of the robot's speech as auxiliary information is known. Since one of the speech signal to be separated is unknown and the other is known, this source separation is *semi-blind*. We assume that the anechoic robot speech signal can be obtained via a spoken dialogue system that controls the robot. Due to the reverberation of the robot speech signal, the separation cannot be achieved simply by subtracting the anechoic signal from the mixed signal. Therefore, we need the second step of blind dereverberation: the reverberation component of the separated human speech signal is removed. This step outputs an anechoic signal of human's speeches. Since any prior information affecting the reverberation, such as the size of the room and the position of the human or the robot, is not given, this dereverberation is *blind*.

This paper describes a newly-designed neural network, named semi-blind recurrent neural network (SB-RNN). The SB-RNN has two modules: a semi-blind source separation

978-1-5090-6341-3/17/$31.00

module and a blind dereverberation module. These modules are concatenated to form a single network to conduct a source separation and a dereverberation. When the network for source separation and that for the dereverberation are trained independently, the overall performance is not optimized and the network for dereverberation assumes inputs of completely separated signals. Therefore, we introduce joint training of the two module networks. Furthermore, we employ multi-task learning [1] to clarify what each module should learn. Thus, the two modules are tuned to separation and dereverberation respectively.

The organization of this paper is as follows. In Section 2 we refer to related works and Section 3 describes the proposed SB-RNN method. In Section 4 we describe the experimental setting and the results and Section 5 concludes this study.

2. RELATED WORK

For the speech enhancement task addressed in this paper, semi-blind independent component analysis (SB-ICA) was proposed [2]. The SB-ICA uses the observation of the microphone and source signals of robot's speech as inputs to estimate source signals of human speech. The SB-ICA is an extension of independent component analysis (ICA) [3, 4]. The methods using ICA, including the SB-ICA method, do not require learning in advance. However, these methods require the same number of channels of microphone array as the number of the signals. The SB-ICA method can operate in real time but it takes few seconds to converge the outputs.

Some other methods predict masks for the source separation. Masks have two types: a hard mask and a soft mask. Methods with a hard mask (binary mask) [5] assume the only one signal affects the observation at a certain time and frequency. On the other hands, methods with a soft mask [6] assume multiple signals affect the observation at any point. The hard mask is a matrix whose values are just 0 or 1, while the soft mask is a matrix with values between 0 and 1. The values of each element of the estimated mask present how much the target source signal affects the observation. The output is an element-wise product of the microphone observation and the estimated mask. Hard-mask based methods normally contain errors that some time-frequency bins are attributed wrongly when two speakers speak simultaneously.

Methods dividing observations into many components are also proposed for source separation. Non-negative matrix factorization (NMF) or probabilistic latent component analysis (PLCA) [7,8] are investigated since the spectrum of the voice is low rank matrix. However, these methods require assigning each component to proper sources. Therefore, these methods need learning components of the signals in advance.

Many methods for dereverberation has also been investigated. One method assumes exponentially decay of the reverberation and subtracts reverberation components in the spectra [9]. Another method estimates a filter to reconstruct envelope modulations of anechoic signals [10]. A method using features that anechoic signals have high kurtosis and maximize kurtosis of an input is also proposed [11]. These methods assume how the impulse responses are and how the anechoic signals are.

Recently, neural networks are also used as the solution for the source separation and dereverberation the [12,13]. Especially, many methods have been proposed along with the growth of the deep neural network. A basic multi-layer perceptron shows performances equivalent or superior to other methods [14]. Recently advanced networks such as recurrent neural network (RNN) performs well in the speech separation tasks and the dereverberation tasks [15–17]. Deep neural networks can learn any model of the mixture and the reverberation. Preparing huge amount of data for learning networks enables the networks to conduct the task robustly.

3. PROPOSED METHOD

This section describes a semi-blind speech enhancement method using a newly-designed semi-blind recurrent neural network (SB-RNN). The overview of the proposed method is shown in Fig. 1.

- Input: a spectrum of the audio signal recorded by the monaural microphone. The audio signal is a mixture of human and robot's echoic speech signals.
- Output: a spectrum of the enhanced (separated and anechoic) human speech signal.

Let $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tF})$ be a spectrum of the input audio signal at t-th time frame where F is the number of the frequency bins and let $\boldsymbol{x}_t^R$ and $\boldsymbol{x}_t^H$ be spectra of the echoic robot's and user's speech, respectively. $\boldsymbol{s}_t^R$ and $\boldsymbol{s}_t^R$ are the sources of human speech and robot's speech, respectively. The observation process is described as

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{x}_t^R + \boldsymbol{x}_t^H, \\ \boldsymbol{x}_t^R &= \boldsymbol{h}^R \odot \boldsymbol{s}_t^R, \\ \boldsymbol{x}_t^H &= \boldsymbol{h}^H \odot \boldsymbol{s}_t^H, \end{aligned} \tag{1}$$

where $\boldsymbol{h}^R$ and $\boldsymbol{h}^H$ represent the frequency transfer functions to the microphone from the robot's loudspeaker and the user, respectively. The operator $\odot$ represents the Hadamard (element-wise) product of two vectors.

We assume that equation (1) holds for the amplitude spectra, i.e., we assume the additivity for the amplitude spectra:

$$\begin{aligned} |\boldsymbol{x}_t| &= |\boldsymbol{x}_t^R| + |\boldsymbol{x}_t^H|, \\ |\boldsymbol{x}_t^R| &= |\boldsymbol{h}^R| \odot |\boldsymbol{s}_t^R|, \\ |\boldsymbol{x}_t^H| &= |\boldsymbol{h}^H| \odot |\boldsymbol{s}_t^H|. \end{aligned} \tag{2}$$

The inputs and outputs of typical neural networks are real numbers. We use the amplitude spectra as the inputs and outputs of the proposed SB-RNN. For the convenience, we omit

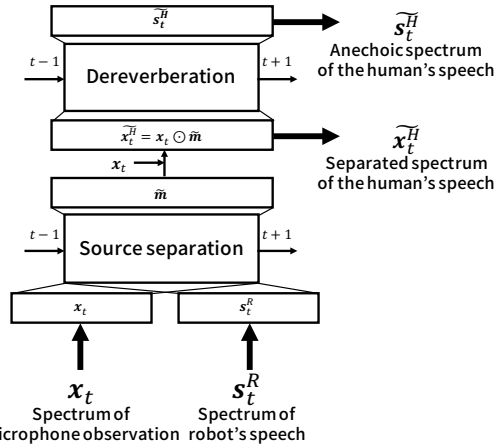


Fig. 2. A structure of the proposed network

the notation of the absolute value, e.g., simply $\boldsymbol{x}_t$ represents the amplitude spectrum of the recorded signal and other symbols are used in the same manner. When we reconstruct a time domain signal from predicted amplitude spectrum $\tilde{\boldsymbol{s}}^H$, we use phases from the microphone observation i.e. $\arg(\boldsymbol{x}_t)$.

Fig. 2 illustrates the structure of the proposed SB-RNN. The SB-RNN is composed of serially cascaded two modules: a source separation module and a dereverberation module. The following part of this section explains the architecture of each module of the network.

3.1. Source separation module

The source separation module estimates a spectral mask $\boldsymbol{m}_t = (m_{t1}, \ldots, m_{tF})$ that separates an amplitude spectrum $\boldsymbol{x}_t$, the mixture of an echoic spectrum of the user $\boldsymbol{x}_t^H$ and that of the robot $\boldsymbol{x}_t^R$, into each isolated speech signal. The inputs of the separation module are the observation signal $\boldsymbol{x}_t$ and the anechoic speech signal of the robot $\boldsymbol{s}_t^R$ and the output is an estimated mask $\tilde{\boldsymbol{m}}_t$. By using the estimated mask $\tilde{\boldsymbol{m}}_t$, a separated spectra of the user $\tilde{\boldsymbol{x}}_t^H$ and that of the robot $\tilde{\boldsymbol{x}}_t^R$ are described as

$$\begin{aligned} \tilde{\boldsymbol{x}}_t^H &= \tilde{\boldsymbol{m}}_t \odot \boldsymbol{x}_t, \\ \tilde{\boldsymbol{x}}_t^R &= (\boldsymbol{1} - \tilde{\boldsymbol{m}}_t) \odot \boldsymbol{x}_t. \end{aligned} \tag{3}$$

$\boldsymbol{1}$ represents an F-dimensional vector of all ones.

The separation module is constructed of a five-layer network: an input layer, three hidden layers and an output layer. The numbers of nodes in the input, each of the hidden, and the output layers are $2F$, 500, and F, respectively. The input of the module is a concatenated vector of the observation spectrum $\boldsymbol{x}_t$ and the anechoic robot speech spectrum $\boldsymbol{s}_t^R$. The output of the module is a spectral mask $\boldsymbol{m}_t$. The middle of the hidden layers is designed as a recurrent layer because improvement of the enhancement performance is expected by handling a strong correlation between adjacent frames of the audio spectra. A rectified linear unit (ReLU) is used as an activation function for the input and hidden layers. For the output layer, a sigmoid function is used as an activation function because each element of the spectral mask is defined in 0 to 1. The structure of the separation module is designed by reference to a conventional method of source separation [16].

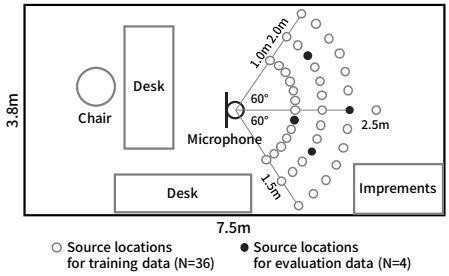


Fig. 3. A room where the experiment was conducted

In this paper, the separation module refers 32 past frames during training and evaluation. If we consider more frames, the network can model longer-term features of the signals, but it would be difficult to train.

3.2. Dereverberation module

The dereverberation module estimates an anechoic amplitude spectrum $\tilde{\boldsymbol{s}}_t^H$ from the echoic spectrum $\tilde{\boldsymbol{x}}_t^H$ separated by the separation module. The input of the dereverberation module is $\tilde{\boldsymbol{x}}_t^H$ and the output is $\tilde{\boldsymbol{s}}_t^H$. Teacher signals for the dereverberation module are anechoic signals $\boldsymbol{s}_t^H$.

The dereverberation module is also constructed of a five-layer network: an input layer, three hidden layers, and an output layer. The numbers of nodes in the input, each of the hidden, and the output layers are F, 500, and F, respectively. Because reverberation components have strong correlations between adjacent frames, the second hidden layer is designed as a recurrent layer as in the separation module. The difference from the separation module is in an activation function in the output layer; the separation module employs a sigmoid function but the dereverberation module employs the ReLU function for an output to be $0 \leq S_{tf}^H < \infty$ in each frequency bin f and time bin t.

3.3. Multi-task learning

The SB-RNN uses teacher signal for the separation module $\boldsymbol{x}_t^H$ and teacher signal for the dereverberation module $\boldsymbol{s}_t^H$. The teacher signal $\boldsymbol{x}_t^H$ is used against the output of the separation module $\tilde{\boldsymbol{x}}_t^H$. On the other hand, teacher signal $\boldsymbol{s}_t^H$ is used against the output of the separation module $\tilde{\boldsymbol{s}}_t^H$. The training of the network is multi-task learning with these two teacher signals so that the separation module is mainly tuned to separate signals and the dereverberation module are mainly tuned to dereverberate the separated signals. We use the mean squared error as a cost function of the SB-RNN. Therefore, the cost between the teacher signal and the estimated output

Table 1. Results of the experiments, SDR (dB)

		SNR (dB) of the datasets						
Method	Figure 4	-6.0	-3.0	0.0	3.0	6.0	9.0	Avg.
No processing	–	-7.63	-5.31	-3.49	-0.67	0.18	1.13	-2.65
SB-ICA	–	-2.42	-1.36	-0.31	1.96	1.96	2.23	0.34
Blind RNN	A	-1.78	0.06	1.25	1.68	2.16	2.52	0.99
Semi-blind MLP	B	-1.91	-0.11	0.76	1.69	1.77	2.11	0.71
Single-task SB-RNN	C	0.12	1.49	2.21	2.99	3.22	2.97	2.17
Semi-blind separated RNNs	D	0.55	1.63	2.42	3.37	3.37	**3.68**	2.51
Proposed SB-RNN	E	**0.67**	**1.97**	**2.65**	**3.45**	**3.46**	**3.68**	**2.64**

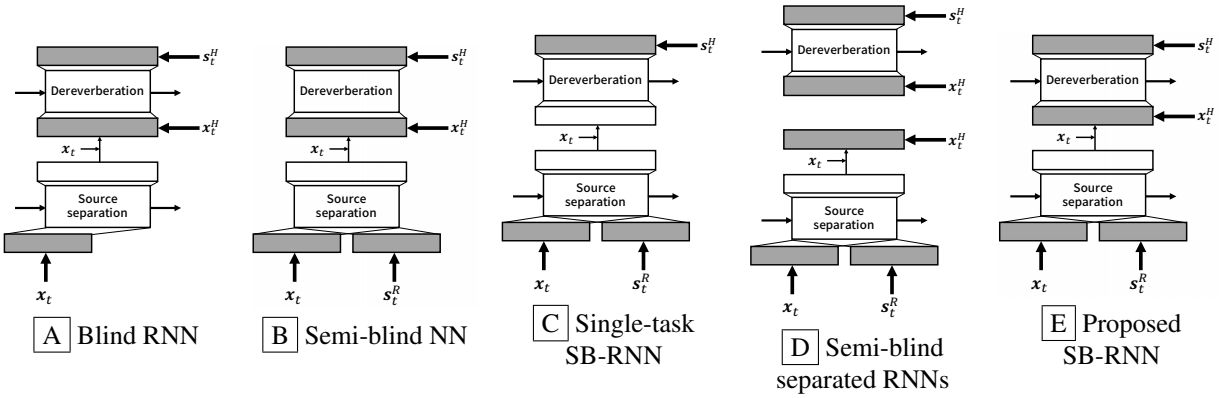


Fig. 4. Networks compared to the proposed SB-RNN.

of the SB-RNN is

$$\begin{aligned} J &= J_S + J_D, \\ J_S &= \|\boldsymbol{x}_t^H - \tilde{\boldsymbol{x}}_t^H\|_2^2, \\ J_D &= \|\boldsymbol{s}_t^H - \tilde{\boldsymbol{s}}_t^H\|_2^2. \end{aligned} \tag{4}$$

Adam [18] is used as a method for optimization of the network. We use the batch normalization [19] at all layers except the output layers of the separation module and the dereverberation module.

4. EXPERIMENTAL EVALUATION

We conducted an experiment to evaluate the proposed SB-RNN. The source to distortion ratio (SDR) [20] is used for the evaluation measure.

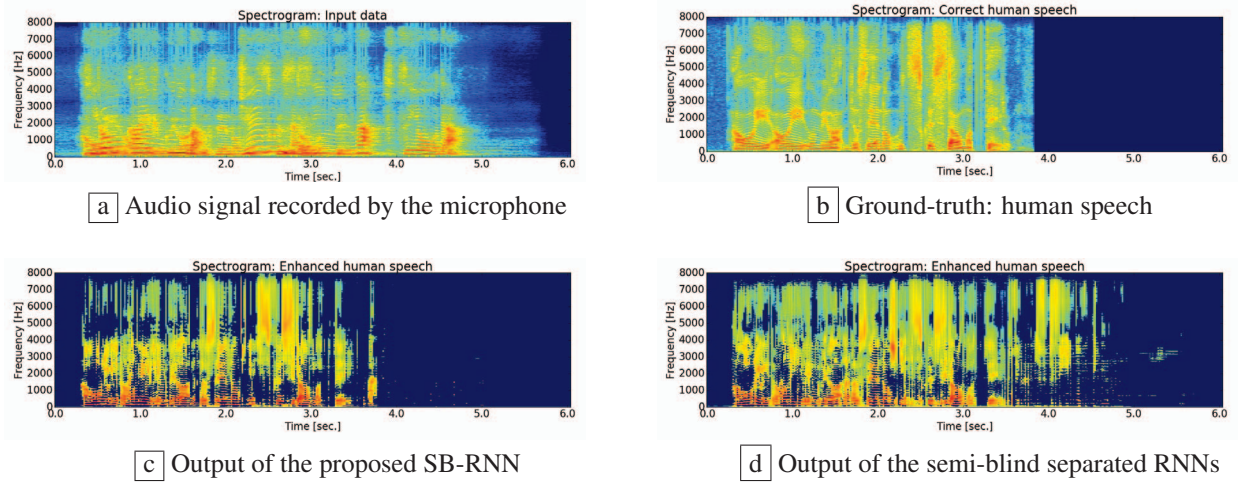
4.1. Experimental conditions

In this experiment, a Japanese large speech database ASJ-JNAS [21] is used for training and evaluation. Speeches of male speakers are used as "human speech" and those of female speakers are used as "robot's speech". For each, 3012 speeches of 60 speakers are used for training and 200 speeches of 4 speakers are used for evaluation.

Signals observed by microphones are simulated by impulse responses. Fig. 3 shows a room where the experiment was conducted. Impulse responses are measured at 40 points in the room. These 40 measured impulse responses are divided into 36 impulse responses for training and 4 impulse responses for evaluation, which make the information of places blind. The signal to noise ratio (SNR) is randomly set to one of -6.0dB, -3.0dB, 0.0 dB, 3.0dB, 6.0dB and 9.0dB for each speech.

We also evaluated the performances of other methods: 'No processing', 'Semi-blind ICA', 'Semi-blind separated RNNs', 'Single-task SB-RNN', 'Semi-blind MLP' and 'Blind RNN'. The 'No processing' method does not execute any process, so $\boldsymbol{x}_t$ is used as an output of this method. The 'Semi-blind ICA' method is taken from Robot Audition System HARK.

Remaining four methods are based on neural network. Fig. 4 shows structure of the networks used in these methods. The 'Blind RNN' method uses the same network to the proposed SB-RNN, but does not employ source signals of robot's speeches $\boldsymbol{s}_t^R$ as an input of the network. The 'Semi-blind MLP' method uses almost the same network as the proposed SB-RNN without recurrent structures. The 'Single-task SB-RNN' method uses the same networks as the proposed SB-RNN, but only $\boldsymbol{s}_t^H$ is used as a teacher signal of the network. Therefore, all layers of the network are to conduct source separation and dereverberation. The cost function of this method is $J^{ST} = \|\boldsymbol{s}_t^H - \tilde{\boldsymbol{s}}_t^H\|_2^2$. The 'Semi-blind separated RNNs' method uses the same modules to the proposed network, but these modules are trained individually. The separation module of this method is learned with the input $\boldsymbol{x}_t$ and the output



a Audio signal recorded by the microphone

b Ground-truth: human speech

c Output of the proposed SB-RNN

d Output of the semi-blind separated RNNs

Fig. 5. An example of the results.

$\boldsymbol{x}_t^H$, and the dereverberation module of this method is learned with the input $\boldsymbol{x}_t^H$ and the output $\boldsymbol{s}_t^H$.

4.2. Experimental results

Table 1 shows the results of the experiment. The SDRs with the proposed SB-RNN is improved by 2.3dB on average compared to the SB-ICA. The SB-RNN succeeded in gaining robustness against the different impulse responses through training of the network, since the locations of the source signals are different between training and evaluation.

Other methods based on neural networks perform better than the SB-ICA but not as well as the proposed SB-RNN. Blind RNN (A) resulted in poor performance in SDR, confirming the effectiveness of the semi-blind scheme. Semi-blind MLP (B) also did not perform well, confirming that the recurrent structure is essential in capturing temporal information. Compared with Single-task SB-RNN (C), the proposed SB-RNN (E) adopts multi-task learning and shows better performance in all SNR conditions. The result demonstrates that the multi-task learning is effective for the complex task of separation and dereverberation.

Now we discuss the difference between the proposed SB-RNN and the semi-blind separated RNNs qualitatively. Fig. 5-c and Fig. 5-d show examples of the spectrogram of the enhanced speech signals obtained by the SB-RNN and the separated RNNs, respectively. In the segment where the human keeps silent and the robot speaks, the SB-RNN correctly outputs silent but the semi-blind separated RNNs often output some kind of noises. In the multi-task learning of the SB-RNN, the spectrum including noise and distortion caused by source separation is given to the input of the dereverberation module. The dereverberation module could acquire a noise-reduction capability in addition to the dereverberation capability. As a result, even if the input spectrum is noisy and/or distorted, the dereverberation module can suppress them. On the other hand, in the learning of the semi-blind separated RNNs, only the speech spectrum without noise and distortion is used as the input of the dereverberation module. The module acquired the very dereverberation capability and thus the module cannot suppress them.

5. CONCLUSION

This paper has presented a speech enhancement method using a semi-blind recurrent neural network (SB-RNN) for human-robot interaction. The experimental results show that the proposed method achieved better SDR by 2.3dB compared to the SB-ICA. We also compared with other networks to confirm the effectiveness of the proposed method.

Future works include the extension of the proposed network. The proposed SB-RNN does not consider noises from neither a robot nor a human. Therefore, the SB-RNN should be extended to deal with these noises. We also need to evaluate the proposed method in automatic speech recognition. Without reconstructing phase information, artificial distortions arise in the reconstructed time-domain signals of human speech [22]. Thus, the improvement of the SDRs is limited. One solution is employing complex valued neural networks [23]. Complex valued neural networks can treat phase information of the signals recorded by a single-channel microphone. Other solution is using multi-channel microphone to learn the arrival delay of the signal.

6. REFERENCES

[1] Rich Caruana, "Multitask learning," in *Learning to learn*, pp. 95–133. Springer, 1998.

[2] Ryu Takeda, Kazuhiro Nakadai, Kazunori Komatani,

Tetsuya Ogata, and Hiroshi G. Okuno, "Barge-in-able robot audition based on ICA and missing feature theory under semi-blind situation," in *IROS 2008*, pp. 1718–1723.

[3] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[4] Noboru Murata, Shiro Ikeda, and Andreas Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

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

[6] Aarthi M. Reddy and Bhiksha Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1766–1776, 2007.

[7] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *ICASSP 2006*, vol. 5, pp. 621–624.

[8] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[9] Katia Lebart, Jean-Marc Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[10] Carlos Avendano and Hynek Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *ICSLP 1996*, vol. 2, pp. 889–892.

[11] Bradford W. Gillespie, Henrique S. Malvar, and Dinei A.F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *ICASSP 2001*, vol. 6, pp. 3701–3704.

[12] Juha Karhunen, Erkki Oja, Liuyue Wang, Ricardo Vigario, and Jyrki Joutsensalo, "A class of neural networks for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 486–504, 1997.

[13] Ying Tan, Jun Wang, and Jacek M. Zurada, "Nonlinear blind source separation using a radial basis function network," *IEEE Transactions on Neural Networks*, vol. 12, no. 1, pp. 124–134, 2001.

[14] Kun Han, Yuxuan Wang, DeLiang Wang, William S. Woods, Ivo Merks, and Tao Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[15] Felix Weninger, Florian Eyben, and Bjorn Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *ICASSP 2014*, 2014, pp. 3709–3713.

[16] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[17] Andrew L. Maas, Quoc V. Le, Tyler M. O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Interspeech 2012*, 2012, pp. 22–25.

[18] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[21] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano, and Shuichi Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research.," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[22] Jonathan Le Roux, Nobutaka Ono, and Shigeki Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Interspeech 2008*, pp. 23–28.

[23] Md. Faijul Amin and Kazuyuki Murase, "Single-layered complex-valued neural network for real-valued classification problems," *Neurocomputing*, vol. 72, no. 4, pp. 945–955, 2009.