Integration of Semi-Blind Speech Source Separation and Voice Activity Detection for Flexible Spoken Dialogue

Masaya Wake*‡, Masahito Togami†, Kazuyoshi Yoshii* and Tatsuya Kawahara*
* Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
E-mail: masaya.wake@linecorp.com, yoshii@kuis.kyoto-u.ac.jp and kawahara@i.kyoto-u.ac.jp
† LINE Corporation, Shinjuku-ku, Tokyo, Japan
E-mail: masahito.togami@linecorp.com
‡ Currently at LINE Corporation

***Abstract*—Conventionally, speech separation (SS) and voice activity detection (VAD) have been investigated separately with a different criteria. In natural dialogue systems such as conversational robots, however, it is critical to accurately separate and detect user utterances even while system's speaking. This study addresses the integration of semi-blind source separation (SS) and voice activity detection (VAD) using a single recurrent neural network under the condition that the speech source and voice activity of the system are given. This study investigates three methods of integrated networks where SS and VAD are processed simultaneously or sequentially prioritizing each. The proposed methods input a single-channel microphone observation spectrum, a speech source spectrum, and voice activity of the system, and then output a speech source spectrum and voice activity of the user. Each network adopts long short-term memory (LSTM) to take the dependency of speech into account. An experimental evaluation using a dataset of recorded dialogues between a user and the android ERICA shows the proposed method that conducts two tasks sequentially with SS first achieves the best performance for both SS and VAD.**

I. Introduction

The use of smartphones, smart speakers, and robots are becoming widespread for voice conversations with users. These systems suppose a situation where one is speaking and the other is not. Thus, users need to tap a screen or speak a specific "magic" word when they make an utternace. On the other hand, in a natural dialogue between humans, there are no restrictions on who can speak, therefore both utterances may overlap. To enable a user and the system to have smooth voice conversations as humans usually do, source separation (SS) [1], [2], [3] and voice activity detection (VAD) [4], [5] are required. SS enables the system to correctly recognize a user's speech even when the user and system speak at the same time. VAD enables the system to correctly identify when the user starts speaking and to immediately stop speaking itself.

In this paper, we propose a method that integrates semi-blind SS and VAD and processes them using a single recurrent neural network (RNN). In a spoken dialogue system, the system's speech signal and voice activity can be easily obtained. This known information leads to semi-blind setting for SS and VAD, therefore overall performance improvement is expected compared with blind setting that does not use such information. The proposed method also uses multi-task learning [6] that processes both SS and VAD in a single neural network. SS and VAD should extract the common feature of a user's speech from microphone observation. Processing these tasks in a single network can improve each of their performances by mutually helping each other. This paper is outlined as follows. Section 2 describes related studies on SS and VAD. Section 3 describes the proposed method, which integrates SS and VAD. Section 4 reports the evaluation experiment. Section 5 presents our conclusion and future tasks.

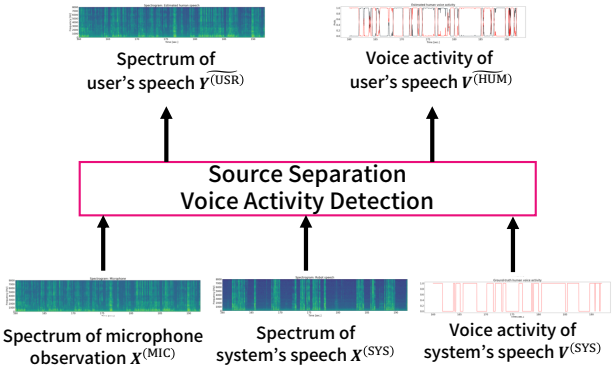

Fig. 1. The process flow of the proposed method

II. Related Studies

A. Semi-blind source separation

This section describes the related methods of semi-blind SS using a known speech source in addition to microphone observation.

1) Independent component analysis: Takeda et al. [7] proposed a method that processes semi-blind SS and dereverberation by extending the independent component analysis (ICA)

[8]. Given an M-channel microphone observation and a given sound source, the target speech source is obtained assuming that the user's speech and the system's speech are statistically independent.

2) Neural network: Wake et al. [9] proposed a method that processes semi-blind SS and dereverberation using an RNN. By adopting multi-task learning [6] that processes the SS and dereverberation in a single neural network and employing two modules for them in the network, the performance of each process is improved.

B. Voice activity detection

This section describes the related VAD methods.

1) Sohn's method: For a robust VAD under high noise, Sohn et al. [10] proposed a method that uses the likelihood ratio by modeling speech and noise. This method assumes that the spectra of the speech and noise follow a Gaussian distribution, and determines the voice activity by calculating the likelihood ratio between the speech and noise. This method enables a robust VAD performance, even when the signal-to-noise ratio (SNR) is low, if the variance of the Gaussian distribution can be correctly estimated.

2) Neural network: Recently, methods for VAD using neural networks have also been proposed [5], [11], [12], [13]. VAD is conducted without setting a threshold or estimating a source model. These methods adopt a number of neural network types such as RNNs [5], [11], [12] and convolutional neural networks (CNNs) [13].

III. Proposed method

A. Problem setting

In a voice dialogue between a user and a system whose speech source and voice activity are known, the proposed method estimates the user's speech source and voice activity using single-channel microphone observations and known information. The problem is formulated below (See Figure 1).

- Inputs
  $\boldsymbol{X}^{(\mathrm{MIC})}$: Power spectrum of the observation,
  $\boldsymbol{X}^{(\mathrm{SYS})}$: Power spectrum of the system's speech,
  $\boldsymbol{V}^{(\mathrm{SYS})}$: Voice activity of the system's speech
- Outputs
  $\tilde{\boldsymbol{Y}}^{(\mathrm{USR})}$: Estimated power spectrum of the user's speech,
  $\tilde{\boldsymbol{V}}^{(\mathrm{USR})}$: Estimated voice activity of the user's speech

The proposed method does not directly estimate the power spectrum of the user's speech source, but estimates the SS mask $\tilde{\boldsymbol{m}}$ to process the SS by calculating the element product of the microphone observation and estimated mask.

B. Designing network

This section describes the design of a single neural network for the SS and VAD. Three types of network can be considered, depending on whether to process the SS and VAD simultaneously or sequentially, and also depending on whether to process first.

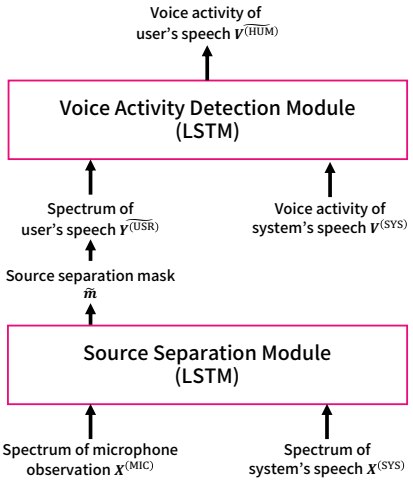

Fig. 2. Outline of sequential type (SS-VAD)

1) Sequential type (SS-VAD): The sequential type (SS-VAD) is designed to process the SS module followed by the VAD module. Figure 2 shows an outline of this neural network.

The SS module parameters are expressed as θ_{SS} and the VAD module parameters are expressed as θ_{VAD}. The network described in this subsection can be expressed as follows:

$$\begin{aligned}
\tilde{\boldsymbol{m}} &= \mathrm{DNN}_{\theta_{\mathrm{SS}}}\left(\boldsymbol{X}^{(\mathrm{MIC})}, \boldsymbol{X}^{(\mathrm{SYS})}\right) \\
\tilde{\boldsymbol{Y}}^{(\mathrm{USR})} &= \tilde{\boldsymbol{m}} \odot \boldsymbol{X}^{(\mathrm{MIC})} \\
\tilde{\boldsymbol{V}}^{(\mathrm{USR})} &= \mathrm{DNN}_{\theta_{\mathrm{VAD}}}\left(\tilde{\boldsymbol{Y}}^{(\mathrm{USR})}, \boldsymbol{V}^{(\mathrm{SYS})}\right)
\end{aligned}$$

When training this network, the loss between the output and ground-truth of the VAD module propagates to the SS module, while the loss between the output and the ground-truth of the SS module does not propagate to the VAD module. Therefore, it is considered that the SS module assists the function of the VAD module.

2) Sequential type (VAD-SS): The sequential type (VAD-SS) is designed to process the VAD module followed by the SS module. Figure 3 shows an outline of this neural network.

Using θ_{SS} and θ_{VAD} described previously, the network described in this subsection can be expressed as follows:

$$\begin{aligned}
\tilde{\boldsymbol{V}}^{(\mathrm{USR})} &= \mathrm{DNN}_{\theta_{\mathrm{VAD}}}\left(\boldsymbol{X}^{(\mathrm{MIC})}, \boldsymbol{V}^{(\mathrm{SYS})}\right) \\
\tilde{\boldsymbol{m}} &= \mathrm{DNN}_{\theta_{\mathrm{SS}}}\left(\tilde{\boldsymbol{V}}^{(\mathrm{USR})}, \boldsymbol{X}^{(\mathrm{MIC})}, \boldsymbol{X}^{(\mathrm{SYS})}\right) \\
\tilde{\boldsymbol{Y}}^{(\mathrm{USR})} &= \tilde{\boldsymbol{m}} \odot \boldsymbol{X}^{(\mathrm{MIC})}
\end{aligned}$$

When training this network, the loss between the output and the ground-truth of the SS module propagates to the VAD module, while the loss between the output and the ground-truth of the VAD module does not propagate to the SS module. Therefore, it is considered that the VAD module assists the function of the SS module.

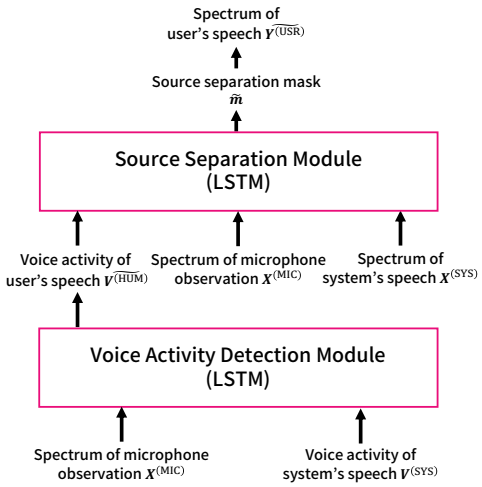

Fig. 3. Outline of sequential type (VAD-SS)

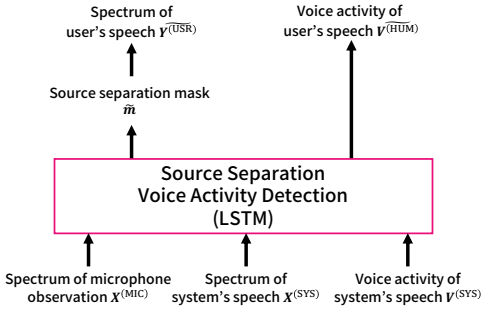

Fig. 4. Outline of Simultaneous type

3) Simultaneous type: The simultaneous type is designed to process the VAD and SS modules simultaneously in the whole network. Figure 4 shows an outline of this neural network.

The parameters of the network are expressed as θ_{SV} and the network described in this subsection can be expressed as follows:

$$\begin{aligned}\left[\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{V}}^{(\mathrm{USR})}\right] &= \mathrm{DNN}_{\theta_{\mathrm{SV}}}\left(\boldsymbol{X}^{(\mathrm{MIC})}, \boldsymbol{X}^{(\mathrm{SYS})}, \boldsymbol{V}^{(\mathrm{SYS})}\right) \\ \tilde{\boldsymbol{Y}}^{(\mathrm{USR})} &= \tilde{\boldsymbol{m}} \odot \boldsymbol{X}^{(\mathrm{MIC})}\end{aligned}$$

When training this network, the loss between the output and the ground-truth of both the SS and VAD modules propagates to the whole network. Therefore, the parameters that can simultaneously process the SS and VAD modules are trained.

C. Designing loss function

In the proposed method, a loss function for the SS and VAD modules is designed since both are performed in a single network.

In this study, the loss function of the network L is divided into the following:

TABLE I
NUMBER OF DATASET FOR TRAINING AND EVALUATION

	# of dialogue	Time
Training	31	5.3 hrs.
Evaluation	8	1.4 hrs.

- L_{SS}: Loss of SS
- L_{VAD}: Loss of VAD

For the L_{SS}, the square error of the logarithmic spectrum is used to consider the scale of the power of the speech and to perform well even in low power frames. For the L_{VAD}, the cross-entropy is used since the VAD output is the probability of speech for a certain frame. Weighting these losses with λ_{VAD}, the loss L to train a neural network is designed as follows.

$$\begin{aligned}L_{\mathrm{SS}} &= \sum_{t,f}\left(\log X_{tf}^{(\mathrm{USR})} - \log \tilde{X}_{tf}^{(\mathrm{USR})}\right)^2 \\ L_{\mathrm{VAD}} &= -\sum_{t,l}\left(V_{tl}^{(\mathrm{USR})} \log \tilde{V}_{tl}^{(\mathrm{USR})}\right) \\ L &= L_{\mathrm{SS}} + \lambda_{\mathrm{VAD}} L_{\mathrm{VAD}}\end{aligned}$$

In this formula, t is the time bin, f is the frequency bin and l is the VAD label.

D. Pre-training

For the sequential types, pre-training each module as an independent network before training the whole network is considered. Pre-training is expected to be more efficient than training the parameters of the entire network from initial random values.

IV. Evaluation Experiment

This section describes the experiment conducted to evaluate the performance of the proposed method.

A. Experiment setting

The recorded dialogues between a user and the android ERICA [15] were used as the dataset of this experiment. 39 users in their 20s to 60s had a dialogue with ERICA under the task of speed dating, interviewing, or attentive listening. The role of ERICA was performed by using the Wizard of Oz method, with a female operator speaking remotely. The speech of ERICA was recorded with a close-up microphone and played using speakers installed at its ear and waist. A 16-channel microphone array was placed on the desk between the user and ERICA and a directional single-channel gun microphone was placed at the user's feet to record the dialogue.

Figure 5 shows the positional relationship between the user, ERICA, and the microphones. Transcripts with timestamps were also created for the speech of the user and ERICA.

The dataset was divided into training data and evaluation data as shown in Table I. In this experiment, a randomly selected single-channel signal among the 16-channel microphone observation was used from the microphone array as a

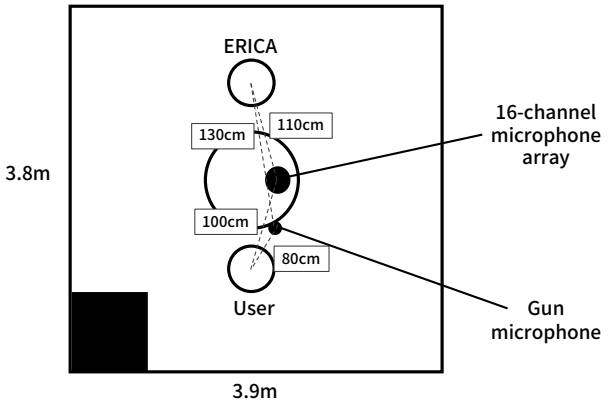

Fig. 5. Positional relationship between user, ERICA and microphones

TABLE II
HYPER PARAMETERS

Sample rate	16000Hz
STFT window size	1024 samples (64msec.)
STFT shift size	256 samples (16msec.)
Input length	3750 frames (Approx. 60sec.)
# of units in LSTM	512
Weight L_{VAD}	1.0
Learning rate α	0.001
Size for a minibatch	16

single-channel microphone observation, and a single-channel signal from the directional gun microphone was used as a ground-truth user speech source. The power spectrum input to the neural network is calculated by a short-time Fourier transform (STFT) using a Hamming window. Timestamps of the transcript are used as ground-truths for VAD.

In constructing the network, long short-term memory (LSTM), which can represent the long-term dependence of speech in the time axis, was adopted. The number of the hidden layers are 3 LSTM + 1 dense for the SS module and 1 LSTM + 1 dense for the VAD module for the sequential types, and 4 LSTM + 1 dense for the SS module and 1 fully-connected for the VAD module for the simultaneous type.

In this experiment, Adam [16] was adopted as the optimization algorithm, which can stabilize training more efficiently compared with the stochastic gradient descent (SGD) by using past gradient information. Other hyperparameters used in the experiment are shown in Table II.

Figure 6 shows an outline of each examined network. The numbers in parentheses in Figure 6 match the numbers in the experimental results. The logarithmic spectral distance (LSD) and the accuracy rate of voice activity are used as evaluation measures for SS and VAD, respectively.

B. Experimental results

Table III shows the experimental results.

Comparing the sequential type (SS-VAD), sequential type (VAD-SS), and simultaneous type at the top of Table III, the sequential type (SS-VAD) shows the highest performance

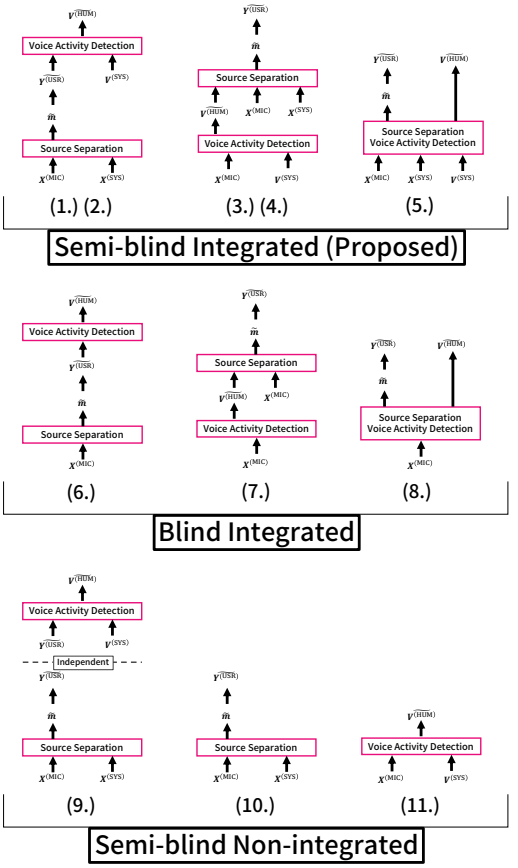

Fig. 6. Outline of compared networks

regardless of the presence or absence of pre-training. When the input to the VAD module is a microphone observation spectrum that is not separated, it is necessary to estimate the number of speakers in the sequential type (VAD-SS) and to train the parameters for both the SS and VAD modules jointly in the simultaneous type. These factors make training them more difficult than the sequential type (SS-VAD).

Comparing the methods with and without pre-training, those with pre-training show better performance in both sequential types (SS-VAD) and (VAD-SS). Training both the SS and VAD modules before they are integrated has a positive effect on the performance of training compared with the case where the whole network is trained from random values.

Comparing the semi-blind and blind methods with those in the middle of Table III, the proposed semi-blind methods show better performance in both SS and VAD in the sequential type (SS-VAD), sequential type (VAD-SS), and simultaneous type.

TABLE III
EXPERIMENTAL RESULT

Setting	LSD	VAD Acc.
Semi-blind Integrated Training (Proposed)		
(1) Semi-blind Integrated Seq. (SS-VAD)	1.953	93.6
(2) Semi-blind Integrated Seq. (SS-VAD) + Pre-training	1.953	94.0
(3) Semi-blind Integrated Seq. (VAD-SS)	1.953	93.0
(4) Semi-blind Integrated Seq. (VAD-SS) + Pre-training	1.953	93.3
(5) Semi-blind Integrated Simul.	1.953	92.6
Blind Integrated Training		
(6) Blind Integrated Seq. (SS-VAD)	1.981	92.5
(7) Blind Integrated Seq. (VAD-SS)	1.967	87.6
(8) Blind Integrated Simul.	2.050	46.1
Semi-blind Non-Integrated Training		
(9) Semi-blind Non-Integrated Seq. (SS-VAD)	1.953	93.3
(10) Only source separation	1.953	—
(11) Only voice activity detection	—	91.0

This result shows that using a known speech source and voice activity of the system is reasonable.

Comparing the integrated methods and non-integrated ones with those at the bottom of Table III, we can observe the effect of integration in the VAD accuracy.

Figures 7 and 8 show the results of the LSD and accuracy rate of VAD, by the sequential type (SS-VAD) for each value of λ_{VAD}. The LSD exceeded 1.96 by $\lambda_{\mathrm{VAD}} = 10^4$, and thereafter the performance of the SS module declined as λ_{VAD} increased. On the other hand, the accuracy rate of VAD was not significantly different depending on the value of λ_{VAD}.

Figure 9 shows an example of voice activity estimation with the sequential type (SS-VAD) + pre-training. It shows the voice activity of the system robot, ground-truth data of the voice activity of the user, and estimated voice activity of the user from the top. The area surrounded by the dotted line in Figure 9 is the section where the user and the system speak at the same time, and the figure shows that the voice activity of the user are correctly estimated even in these sections.

V. CONCLUSION

This report proposed a method that integrates SS and VAD modules and trains them with a single neural network to enable a smooth speech dialogue between a user and system. Three types of networks were investigated in accordance with the order of the SS and VAD, and multi-task learning is adopted. Moreover, since the speech source and voice activity of the system can be acquired in the system, the proposed methods adopt this information appropriately to train the network efficiently in an end-to-end manner. Experimental results have shown that the proposed method improves the performance of SS and VAD compared with the methods of training the SS and VAD networks independently.

Future work includes adding labels called social signals [17] such as backchannels and fillers to the outputs of VAD and estimating these labels by the neural network to estimate the emotions and interests of the user.

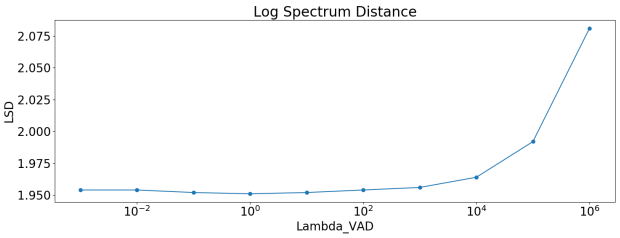

Fig. 7. Logarithmic spectral distance (LSD) comparison by λ_{VAD}

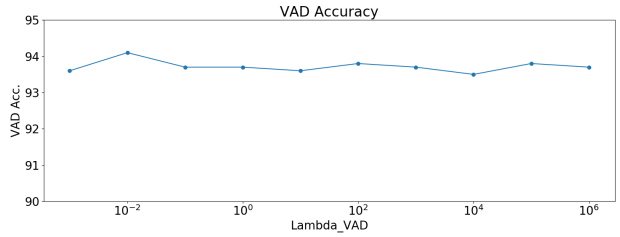

Fig. 8. Accuracy rate of VAD comparison by λ_{VAD}

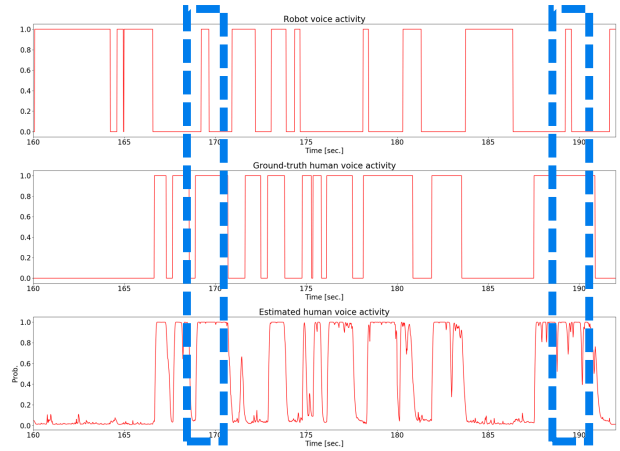

Fig. 9. Example output of semi-blind integrated seq. type (SS-VAD)

ACKNOWLEDGMENT

This research was supported by NII CRIS collaborative research program operated by NII CRIS and LINE Corporation.

REFERENCES

[1] J. Cho, H. Park, and C. D. Yoo, "Blind speech separation and recognition system for human robot interaction in reverberant environment," *2012 9th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI),* November 2012, pp. 584–585.

[2] S. C. Lee, B. W. Chen, and J. F. Wang, "Noisy environment-aware speech enhancement for speech recognition in human-robot interaction application," *2010 IEEE International Conference on Systems, Man and Cybernetics,* October 2010, pp. 3938–3941.

[3] J. M. Valin, J. Rouat, and F. Michaud, "nhanced robot audition based on microphone array source separation with post-filter," *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* September 2014, pp. 2123–2128.

[4] R. Brueckmann, A. Scheidig, and H. Gross, "Adaptive noise reduction and voice activity detection for improved verbal human-robot interaction using binaural data," *Proceedings 2007 IEEE International Conference on Robotics and Automation,* April 2007, pp. 1782–1787.

[5] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing,* May 2013, pp. 7378–7382.

[6] R. Caruana, "Multitask learning," *Machine learning,* vol. 28, no. 1, pp. 41–75, 1997.

[7] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "Step-size parameter adaptation of multi-channel semi-blind ica with piece-wise linear model for barge-in-able robot audition," *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems,* October 2009, pp. 2277–2282.

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

[9] M. Wake, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Semi-blind speech enhancement based on recurrent neural network for source separation and dereverberation," *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP),* September 2017, pp. 1–6.

[10] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters,* vol. 6, no. 1, pp. 1–3, January 1999.

[11] P. Sertsi, S. Boonkla, V. Chunwijitra, N. Kurpukdee, and C. Wutiwiwatchai, "Robust voice activity detection based on lstm recurrent neural networks and modulation spectrum," *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* December 2017, pp. 342–346.

[12] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing,* May 2013, pp. 483–487.

[13] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* May 2014, pp. 2519–2523.

[14] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* Vol. 9, No. 8, pp. 1735–1780, 1997.

[15] D. F. Glas, T. Minato, C. T. Ishi, T. Kawahara, and H. Ishiguro, "Erica: The erato intelligent conversational android," *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN),* August 2016, pp. 22–29.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* pp. 1–15, 2014.

[17] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social Signals, Their Function, and Automatic Analysis: A Survey," *Proceedings of the 10th International Conference on Multimodal Interfaces,* pp. 61–68, 2008.