Journal of Signal Processing, Vol.25, No.6, pp.245-250, November 2021

RESEARCH NOTE

VMInNet: Interpolation of Virtual Microphones in Optimal Latent Space Explored by Autoencoder

Riki Takahashi[1], Li Li[1,2], Shoji Makino[1,3] and Takeshi Yamada[1]

[1]Graduate School of Science and Technology, University of Tsukuba, Ibaraki 305-8577, Japan
[2]Nippon Telegraph and Telephone Corporation, Atsugi 243-0198, Japan
[3]Graduate School of Information, Production and Systems, Waseda University, Fukuoka 808-0135, Japan
E-mail: s.makino@waseda.jp, takeshi@cs.tsukuba.ac.jp

Abstract **In this paper, we propose a new method for the interpolation of virtual signals between two real microphones to improve speech enhancement performance in underdetermined situations. The virtual microphone technique is a recently proposed technique that can virtually increase the number of channels of observed signals by linearly interpolating the phase and nonlinearly interpolating the amplitude based on β-divergence in the short-time Fourier transform (STFT) domain. This technique has been shown to be effective in improving the speech enhancement performance of beamforming in underdetermined situations. It is reasonable to linearly interpolate the phase based on the sound propagation model and nonlinearly interpolate the amplitude to increase the information content of the observed signals. However, there is no theoretical proof that β-divergence is the optimal criterion for amplitude interpolation due to the complexity of the physical model of amplitude. In this paper, we propose the use of an autoencoder to search for the optimal interpolation domain in a data-driven manner. We perform amplitude interpolation in the latent space, a low-dimensional representation space of observed mixture signals that is trained so that the interpolated virtual signals are optimal for conducting beamforming with high performance. Experimental results revealed that the proposed method achieved higher speech enhancement performance than conventional methods.**

Keywords: speech enhancement, virtual microphones, neural network

1. Introduction

In recent years, with the development of automatic speech recognition (ASR) and robot hearing, the importance of speech enhancement has considerably increased. Owing to the wide usage of small devices with built-in stereo microphones, such as smartphones and voice recorders, speech enhancement that serves dual-channel signals is a particular requisite. Beamforming and blind source separation (BSS) are the two main methods used to deal with this problem. Beamforming and BSS methods using spatial filtering [1, 2, 3] are noteworthy in the low distortion of the enhanced target speech. However, their performance degrades when there are fewer microphones than sources, i.e., underdetermined conditions. To achieve satisfactory speech enhancement performance with such devices having a small microphone array, many methods have been proposed to enhance speech in underdetermined situations such as time-frequency masking [4, 5, 6], multichannel Wiener filtering [7, 8], and nonnegative matrix factorization (NMF) [9]. Although these methods are noteworthy in that they can significantly improve speech intelligibility in underdetermined situations, they face a tradeoff between low signal distortion and high noise reduction performance.

On the other hand, the virtual microphone technique [10] allows the well-studied methods for determined situations to be applied to the signals recorded in underdetermined situations by virtually increasing the number of channels. Using two real microphone signals, the virtual microphone technique estimates the observed signal at a position where there is no real microphone placed by interpolating the phase and amplitude independently. Based on the W-disjoint orthogonality (W-DO) [4, 11] assumption, phases of virtual signals can be obtained using linear interpolation by approximately modeling propagating waves

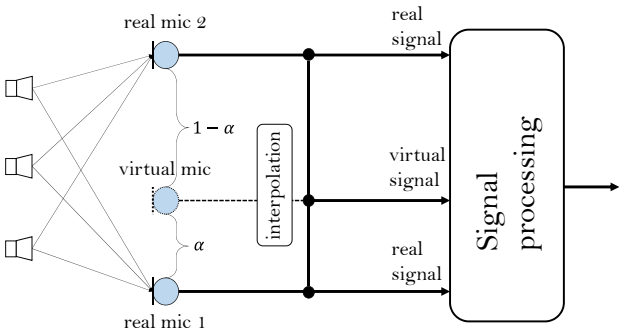

Fig. 1 Microphone array signal processing with virtual microphone technique

as plane waves. For amplitude estimation, since modeling the amplitudes of propagating waves is difficult due to complicated acoustic environments, complex logarithmic interpolation and a generalized version, where interpolation rules are derived as closed-form solutions of an optimization problem formulated using β-divergence, have been proposed and experimentally shown to be effective in improving speech enhancement performance using a maximum signal-to-noise (maxSNR) beamformer. Furthermore, the results in [10] indicated that speech enhancement performance tends to increase when the improved nonlinearity is applied to amplitude interpolation.

However, there is no theoretical proof that β-divergence is the optimal criterion for formulating the amplitude interpolation problem. Because the physical model of amplitude is complicated, it is challenging to design appropriate rules manually. To overcome this problem, in this paper, the use of an autoencoder to search for the optimal interpolation domain in a data-driven manner is proposed. Specifically, we perform amplitude interpolation in the latent space and train the autoencoder so that the interpolated virtual signals become optimal for conducting beamforming. We use two-dimensional fully convolutional neural networks (CNNs) to design the autoencoder, which allows the network to handle inputs with arbitrary length and perform interpolation by taking the whole spectrogram into account. This is different from the conventional method using β-divergence, where the interpolation is performed in a time-frequency bin-wise manner.

Related work: An attempt has already been made to employ a CNN to estimate the amplitude of virtual signals [12], where a CNN is trained to output the amplitude directly given the amplitudes of the observed signals and the interpolation position. Our method differs from this method in that an autoencoder is trained to constrain the estimated amplitude to be a mixture signal by explicitly performing interpolation in the latent space and then passing it through the decoder, which is trained to reconstruct the mixture signals. This design makes the method more intuitive and the network more interpretable. To overcome the strong assumption made for phase and amplitude interpolation, a method that trains a time-domain TasNet to estimate virtual signals in the time domain has been recently proposed [13]. Although this supervised method has shown high performance in improving speech enhancement performance, training networks using real microphone signals limits the flexibility of the location to place the virtual microphone. Moreover, collecting training data in various acoustic conditions has a high cost.

2. Virtual Microphone Technique Based on β-Divergence

We model the microphone signals in the short-term Fourier transform (STFT) domain. Here, let $x_m(\omega, t)$, $m = 1, 2$ be the mth real microphone signal at the angular frequency ω in the tth time frame. The amplitudes of these signals are denoted as $A_m(\omega, t) = |x_m(\omega, t)|$ and the phases are denoted as $\phi_m(\omega, t) = \angle x_m(\omega, t)$. A virtual microphone signal $v(\omega, t, \alpha, \beta)$ is defined as an observation at the point α obtained by internally dividing the line joining two real microphones in the ratio $\alpha : (1 - \alpha)$ (i.e. Fig. 1). β is the hyperparameter of β-divergence, which is used as the metric for amplitude interpolation. Hereafter, we omit ω, t, α, and β for notation simplicity.

We first introduce the interpolation of the phase. We assume W-DO [4, 11] for mixed signals so that each time-frequency bin is dominated by at most one sound source, which means that the observed signal in each time-frequency bin can be regarded as a single wave. Based on this assumption, the physical model of propagating waves can then be approximated as that of a plane wave. The phase $\phi_{\mathrm{v}} = \angle v$ of a virtual microphone signal v can then be interpolated linearly on the basis of this model as

$$\phi_{\mathrm{v}} = (1 - \alpha)\,\phi_1 + \alpha\phi_2 \tag{1}$$

Since the observed phase has an aliasing ambiguity given by $\phi_i \pm 2n_i\pi$ with integer n_i, we need to set the constraint

$$|\phi_1 - \phi_2| \leq \pi \tag{2}$$

to eliminate the ambiguity of the interpolated phase.

For the interpolation of the amplitude, since there are many acoustic conditions such as the distance between the sound sources and microphones, and the direction of arrival (DOA) of sources, it is difficult to faithfully model the amplitude of a propagating wave. Therefore, an interpolation rule based on β-divergence

has been used instead of considering physical models [10]. The β-divergence between the amplitude of a virtual microphone, $A_{\mathrm{v}} = |v|$, and that of the ith real microphone, A_i, is defined as

$$\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_i) = \begin{cases} A_{\mathrm{v}}(\log A_{\mathrm{v}} - \log A_i) + (A_i - A_{\mathrm{v}}) & (\beta = 1) \\ \dfrac{A_{\mathrm{v}}}{A_i} - \log \dfrac{A_{\mathrm{v}}}{A_i} - 1 & (\beta = 0) \\ \dfrac{A_{\mathrm{v}}^\beta}{\beta(\beta-1)} + \dfrac{A_i^\beta}{\beta} - \dfrac{A_{\mathrm{v}} A_i^{\beta-1}}{\beta - 1} & (\text{otherwise}) \end{cases} \tag{3}$$

where $\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_i)$ is continuous at $\beta = 0$ and $\beta = 1$. The interpolation rule of A_{v} is then given as the closed-form solution of the optimization problem that minimizes $\sigma_{\boldsymbol{D}_\beta}$, the sum of $\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_i)$ weighted by the hyperparameter of the virtual microphone interpolation position α:

$$\sigma_{\boldsymbol{D}_\beta} = (1-\alpha)\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_1) + \alpha \boldsymbol{D}_\beta(A_{\mathrm{v}}, A_2) \tag{4}$$

$$A_{\mathrm{v}} = \mathrm{argmin}_{A_{\mathrm{v}}} \sigma_{\boldsymbol{D}_\beta} \tag{5}$$

By differentiating $\sigma_{\boldsymbol{D}_\beta}$ with respect to A_{v} and setting it to 0, the interpolated amplitude is obtained as

$$A_{\mathrm{v}} = \begin{cases} \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right) & (\beta = 1) \\ \left((1-\alpha)\, A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (\text{otherwise}) \end{cases} \tag{6}$$

Note that the phase can be interpolated with an arbitrary real number α, whereas the amplitude interpolation is defined only in the domain of $0 \leq \alpha \leq 1$ when $\beta \neq 1$. The extrapolation of a virtual microphone in the domains $\alpha < 0$ and $\alpha > 1$ was considered in [14].

It has shown that the virtual microphone technique based on β-divergence is effective in improving speech enhancement performance in underdetermined situations [10]. However, there is no proof that β-divergence is the optimal criterion for amplitude interpolation since we cannot analyze the physical model of amplitude.

3. Proposed Method: VMInNet

In this paper, we propose the use of an autoencoder to search for the optimal interpolation domain in a data-driven manner to overcome the above-mentioned problem. The flowchart of the proposed method is shown in Fig. 2. In the proposed method, the two-channel observed signals are separated into the amplitude spectrogram A_m and phase spectrogram ϕ_m as in the conventional method. The amplitude spectrogram A_m is then embedded into latent variables $\mathbf{z}_m = E_\theta(A_m)$ by an encoder network $E_\theta(\cdot)$. We interpolate the amplitude of the virtual microphone signal

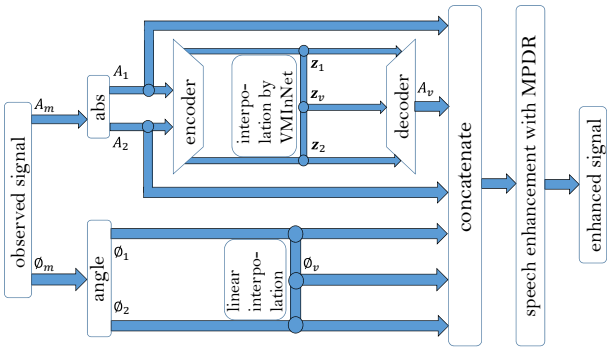

Fig. 2 Flowchart of proposed method

in the latent space as

$$\mathbf{z}_{\mathrm{v}} = (1-\alpha)\mathbf{z}_1 + \alpha \mathbf{z}_2 \tag{7}$$

The latent variables $\mathbf{z}_1$, $\mathbf{z}_2$, and $\mathbf{z}_{\mathrm{v}}$ are then converted to the amplitudes $A_m = D_\psi(\mathbf{z}_m)$ and $A_{\mathrm{v}} = D_\psi(\mathbf{z}_{\mathrm{v}})$ by a decoder network $D_\psi(\cdot)$. Here, θ and ψ respectively denote trainable parameters of the encoder and decoder networks. The phase of a virtual signal, ϕ_{v}, is linearly interpolated using (1). The virtual signal v is obtained by concatenating the estimated amplitude and phase spectrogram as

$$v = A_{\mathrm{v}} \exp\{j\phi_{\mathrm{v}}\} \tag{8}$$

We call the network "VMInNet", an abbreviation of "interpolation network for virtual microphone".

To maximize the potential of the data-driven method, where a task-dependent loss function is allowed, we train the network using a minimum power distortionless response (MPDR) beamformer-based loss function. Specifically, the loss function is designed to minimize the mean squared error between the target signal $s(\omega, t)$ and the output of the MPDR beamformer to force the generated amplitude to be optimal for constructing the MPDR beamformer. For the observed signal consisting of two real and I virtual microphone signals, we use $\boldsymbol{x}(\omega, t) = [x_1(\omega, t), v_1(\omega, t), \cdots, v_I(\omega, t), x_2(\omega, t)]^{\mathrm{T}}$ to denote the mixture signal vector. A beamformer that enhances the source of interest is given by

$$y(\omega, t) = \boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega, t) \tag{9}$$

$$\boldsymbol{w}(\omega) = [w_1(\omega) \cdots w_M(\omega)]^{\mathrm{T}} \tag{10}$$

where $y(\omega, t)$ is the output signal of the beamformer, $\boldsymbol{w}(\omega)$ denotes the spatial filter vector, $(\cdot)^{\mathrm{T}}$ denotes the transpose, $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose, and $M = 2 + I$ denotes the number of channels of the observed signals. The spatial filter $\boldsymbol{w}(\omega)$ derived from

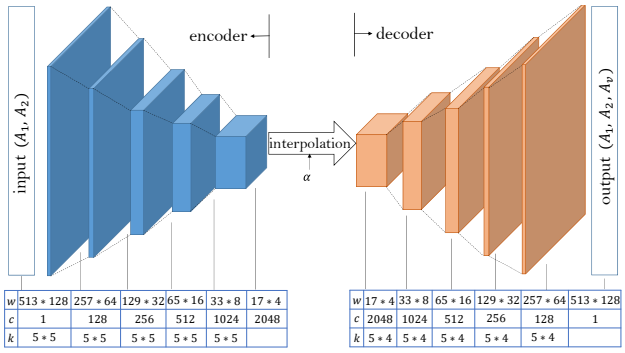

Fig. 3 Network architecture of VMInNet: "w", "c", and "k" denote feature size, channel number, and kernel size, respectively.

the MPDR beamformer is expressed as

$$\boldsymbol{w}(\omega) = \frac{\boldsymbol{\Phi}(\omega)^{-1}\boldsymbol{a}(\omega)}{\boldsymbol{a}^{\mathrm{H}}(\omega)\boldsymbol{\Phi}(\omega)^{-1}\boldsymbol{a}(\omega)} \tag{11}$$

$$\boldsymbol{\Phi}(\omega) = \mathbb{E}[\boldsymbol{x}(\omega,t)\boldsymbol{x}(\omega,t)^{\mathrm{H}}] \tag{12}$$

Here, $\boldsymbol{\Phi}(\omega)$ is the covariance matrix of the signals observed at frequency ω, and $\boldsymbol{a}(\omega)$ is the relative transfer function (RTF) of the target, which is defined as the ratio of the acoustic transfer functions $\boldsymbol{h}(\omega) = [h_1(\omega)\cdots h_M(\omega)]^{\mathrm{T}}$ from the target source to the microphone array, i.e., $\boldsymbol{a}(\omega) = \left[1 \;\; \frac{h_2(\omega)}{h_1(\omega)} \;\; \cdots \;\; \frac{h_M(\omega)}{h_1(\omega)}\right]^{\mathrm{T}}$.

The encoder and decoder parameters θ and ψ, respectively, are trained by minimizing the following loss function:

$$\mathcal{L}(\theta,\psi) = \lambda \sum_{m,\omega,t} |D_\psi(E_\theta(A_m(\omega,t))) - A_m(\omega,t))|^2 + \sum_{\omega,t} |\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t) - s(\omega,t)|^2 \tag{13}$$

Here, the first term of (13) is the restoration error of the autoencoder to ensure that the virtual amplitude latent variables generated in the latent space are restored to the amplitude space. The second term is the MPDR beamformer loss. λ denotes a parameter that weighs the importance of the restoration error. This training criterion allows the network to search for the optimal interpolation space where the interpolated virtual signal is optimal for constructing an MPDR beamformer. Note that an other task-dependent loss function can also be considered as long as the loss function is differentiable.

The network architecture is shown in Fig. 3. We use two-dimensional CNNs to carefully design the network architecture, which allows the network to handle inputs with arbitrary length and perform interpolation by taking the whole spectrogram into account. After each CNN layer, we employ batch normalization to stabilize the training process.

4. Experiments

4.1 Experimental conditions

To evaluate the effectiveness of the proposed method, we conducted speech enhancement experiments in underdetermined situations. The training dataset consisted of three speakers excerpted from the Wall Street Journal (WSJ0) corpus [15] and the audio files for each speaker were 1 minute. The test dataset comprised 10 speakers, where the audio files for each speaker were about 9 minutes. The observed mixture signals were generated by adding reverberant signals of the three speakers together, where the reverberant signals were generated by convolving the impulse responses simulated by a room impulse response generator [16] with clean speech signals. The DOA of the target was set to 90°, and those of interferers were set to 50° and 150° with a reverberation time of 120 ms. All signals were resampled at 8 kHz. Spectrograms with 513 frequency bins were calculated using a Hamming window with a length of 1024 samples and half overlap. The experimental conditions are listed in Table 1. We used data with a fixed length of 128 frames for training by segmenting the training data, while we used data having an arbitrary length for testing. The power of the training data was normalized to unity. λ was set at 20 based on a preliminary experiment. We used the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR) [17] as evaluation criteria to quantify the speech enhancement performance. Here, we used source images of signals as reference signals.

We chose the method interpolating the amplitude based on β-divergence (beta) [10] and that using a CNN to directly estimate the amplitude of virtual signals (cnn) [12] as baseline methods. We assumed that the RTF was known for all the methods. The RTF at the position of the virtual microphone was estimated using the conventional β-divergence-based method, where the amplitude was estimated using (6) with $\beta = 20$, and the phase was estimated using (1). The RTF can also be estimated using a neural network, which is one direction of our future work since we would like to focus on amplitude interpolation in this work.

4.2 Results and discussion

Table 2 shows the results obtained by each method. The proposed method achieved the best score in terms

Table 1 Experimental conditions

Number of real microphones, M	2 or 3
Number of sound sources, N	3
Distance between microphones	4 cm
Reverberation time	120 ms
Sampling rate	8 kHz
Input SNR	0 dB
Window length/shift	1024 / 512 samples

Table 2 SDR, SIR, and SAR [dB] achieved with each method

Conditions	SDR	SIR	SAR
2 real mic + 1 vir mic (beta)	5.97	8.58	**10.06**
2 real mic + 1 vir mic (cnn)	5.34	**15.54**	5.92
2 real mic + 1 vir mic (prop.)	**7.04**	13.32	8.36
2 real mic	2.18	2.86	12.56
3 real mic	16.07	19.49	18.85

of SDR, which confirmed its effectiveness in improving speech enhancement performance. This result indicates that interpolation in the latent space is more effective than using β-divergence and directly estimating interpolated amplitudes with a CNN. Compared with performing interpolation with a rule-based method, the use of neural networks significantly improved SIR but decreased SAR, which is due to the nonlinearity of neural networks. Since the baseline method using a CNN has no constraint on the estimated amplitude, it is possible to estimate the amplitude including fewer interfering signals, where the CNN reduced the interference. However, this led to more distortions. The proposed method addressed this problem by introducing a constraint on the estimated amplitude into the training criterion. Although the constraint reduced the SIR by about 2 dB, it increased the SAR, resulting in a better SDR. We considered that the proposed method achieved a better balance in estimating the amplitude with high nonlinearity and realism.

We also compared the results of using the phase or RTF of real microphones and VMInNet for amplitude interpolation to investigate the importance of accurately interpolating the amplitude. The results are shown in Table 3. From these results, we found that although an accurate phase and RTF could improve the performance, the improvement was slight. Comparing the last row in Table 3 and that in Table 2, we found that the accuracy of the amplitude strongly influences the speech enhancement performance. One possible reason could be that the estimated amplitude is not consistent with the phase of the estimated or real signal. This means that there is no signal in the time domain that corresponds to the estimated spectrogram. A future work is to perform virtual microphone estimation considering the spectrogram consistency.

Table 3 SDR, SIR, and SAR [dB] achieved with each method: "prop.", "linear", "beta", and "real" indicate the proposed VMInNet-based amplitude estimation, linear phase estimation, RTF estimation using β-divergence, and the phase or RTF observed from real microphones, respectively.

Conditions	SDR	SIR	SAR
amp (prop.), phase (linear), RTF (beta)	7.04	13.32	8.36
amp (prop.), phase (real), RTF (beta)	7.38	13.98	8.75
amp (prop.), phase (linear), RTF (real)	7.37	13.94	8.74
amp (prop.), phase (real), RTF (real)	7.41	14.02	8.77

5. Conclusion

In this paper, we proposed VMInNet, an autoencoder network for interpolating the amplitude of virtual microphone signals, which aims to improve speech enhancement performance in underdetermined situations. We trained the network to search for the interpolation domain that is optimal for conducting MPDR beamforming. The experimental results revealed that the proposed method outperformed conventional virtual microphone techniques by achieving an SDR improvement of about 1.1 dB.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JPH04131 and JP19J20420.

References

[1] P. Smaragdis: Blind separation of convolved mixtures in the frequency domain, Neurocomputing, Vol. 22, Nos. 1–3, pp. 21–34, 1998.

[2] D. Kitamura et al.: Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization, IEEE/ACM TASLP, Vol. 24, No. 9, pp. 1622–1637, 2016.

[3] T. Higuchi et al.: Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR, IEEE/ACM TASLP, Vol. 25, No. 4, pp. 780–793, 2017.

[4] O. Yılmaz et al.: Blind separation of speech mixtures via time-frequency masking, IEEE TSP, Vol. 52, No. 7, pp. 1830–1847, 2004.

[5] S. Rickard: The DUET blind source separation algorithm, Blind Speech Separation, S. Makino et al., Eds., pp. 217–241, Springer, 2007.

[6] H. Sawada et al.: Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment, IEEE TASLP, Vol. 19, No. 3, pp. 516–527, 2010.

[7] N. Q. K. Duong et al.: Under-determined reverberant audio source separation using a full-rank spatial covariance model, IEEE TASLP, Vol. 18, No. 7, pp. 1830–1840, 2010.
[8] R. V. Rompaey et al.: GEVD based speech and noise correlation matrix estimation for multichannel Wiener filter based noise reduction, Proc. EUSIPCO, pp. 2562–2566, 2018.
[9] A. Ozerov et al.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation, IEEE TASLP, Vol. 18, No. 3, pp. 550–563, 2010.
[10] H. Katahira et al.: Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer, EURASIP JASP, Vol. 2016, No. 1, pp. 1–8, Jan. 2016.
[11] A. Jourjine et al.: Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures, Proc. ICASSP, pp. 2985–2988, 2000.
[12] K. Yamaoka et al.: CNN-based virtual microphone signal estimation for MPDR beamforming in underdetermined situations, Proc. EUSIPCO, pp. 1–5, 2019.
[13] T. Ochiai et al.: Neural network-based virtual microphone estimator, arXiv preprint arXiv:2101.04315, 2021.
[14] R. jinzai et al.: Microphone position realignment by extrapolation of virtual microphone, Proc. APSIPA, pp. 367–372, 2018.
[15] Garofolo, John, et al.: CSR-I (WSJ0) Complete LDC93S6A, Web Download, Philadelphia: Linguistic Data Consortium, 1993.
[16] E. A. P. Habets: Room impulse response (RIR) generator. https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, 2008
[17] E. Vincent et al.: Performance measurement in blind audio source separation, IEEE TASLP, Vol. 14, No. 4, pp. 1462–1469, 2006.

Riki Takahashi received his B.E. and M.E. degrees from University of Tsukuba, Japan, in 2019 and 2021, respectively. He is currently a systems engineer at Canon IT Solutions Inc. His research interests include audio and speech signal processing and machine learning.

Li Li received her B.E. degree from the Shanghai University of Finance and Economics, China, in 2014, and her M.S. and Ph.D. degrees from University of Tsukuba, Japan, in 2018 and 2021, respectively. She is currently a postdoctoral researcher with Nagoya University and an adjunct researcher with NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. She was a Research Fellow of the Japan Society of Promotion of Science. Her research interests include audio and speech signal processing, source separation, and machine learning. She received the Student conference Presentation Award from the Acoustical Society of Japan and the Student Paper Award from the Research Institute of Signal Processing, Japan.

Shoji Makino received his B.E., M.E., and Ph.D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981 and University of Tsukuba in 2009. He is now a professor at Waseda University, Japan. His research interests include adaptive filtering technologies, the realization of acoustic echo cancellation, the blind source separation of convolutive mixtures of speech, and acoustic signal processing for speech and audio applications. He is an IEEE Fellow, an IEICE Fellow, a council member of the ASJ, and a member of the EURASIP.

Takeshi Yamada received his B.E. degree from Osaka City University, Japan, in 1994, and his M.E. and Dr. Eng. degrees from the Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. He is presently an associate professor with the Faculty of Engineering, Information and Systems, University of Tsukuba, Japan. His research interests include speech recognition, sound scene understanding, multichannel signal processing, media quality assessment, and e-learning. He is a member of the IEEE, the IEICE, the IPSJ, and the ASJ.

(Received May 12, 2021; revised June 26, 2021)