Journal of Signal Processing, Vol.27, No.1, pp.1-6, January 2023

PAPER

Noise Suppression Using Beamformer and Transfer-Function-Gain Nonnegative Matrix Factorization with Distributed Stereo Microphones

Yutaro Matsui[1], Shoji Makino[1,2], Nobutaka Ono[3] and Takeshi Yamada[1]

[1]University of Tsukuba, 1–1–1 Tennodai, Tsukuba, Ibaraki 305–8577, Japan
[2]Waseda University, 2–7 Hibikino, Wakamatsu, Kitakyushu, Fukuoka 808–0135, Japan
[3]Tokyo Metropolitan University, 6–6 Asahigaoka, Hino, Tokyo 191–0065, Japan
E-mail: s.makino@waseda.jp

Abstract **In this paper, we propose a novel approach to noise suppression using multiple distributed recording devices with stereo microphones. In the proposed method, noise suppression based on phase information is applied to the synchronous stereo signals captured by each recording device and then output signals are utilized for transfer-function-gain nonnegative matrix factorization (NMF) as extra input signals. We intended to estimate the target signal more accurately by transfer-function-gain NMF. Experiments using impulse responses measured in a meeting room have shown that the proposed method outperformed conventional methods using transfer-function-gain NMF in terms of the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR).**

Keywords: noise suppression, MVDR beamformer, transfer-function-gain NMF, asynchronous microphone array, stereo microphone

1. Introduction

An asynchronous distributed microphone array is one of the frameworks of array signal processing for speech enhancement or noise suppression. Since we can use portable recording devices such as voice recorders, smartphones and laptops, it is easy to construct an asynchronous distributed microphone array. However, the performance of array signal-processing approaches based on phase information is degraded by phase drift between observed signals recorded by an asynchronous microphone array [1, 2]. Phase drift is caused by the difference in the start time of recording or the mismatch of sampling frequency among recording devices. On the other hand, the approaches utilizing amplitude information for noise suppression can still be effective. One powerful framework of such approaches is the noise suppression method using transfer-function-gain nonnegative matrix factorization (NMF) [3–5]. In this method, we apply transfer-function-gain NMF to the observed signals in the time-channel domain. This method is robust against phase drift because it utilizes not the phase information but the amplitude information of the observed signals. However, if the number of microphones is not sufficiently large compared with the number of sources, the performance of noise suppression using transfer-function-gain NMF is degraded.

When we use voice recorders, smartphones or laptops as recording devices in our daily lives, we can obtain two-channel signals from one device if it has stereo microphones. We can apply array signal processing based on phase information such as that employing a beamformer to the signals from one device because they are synchronous. Thus, we propose a new approach that applies noise suppression based on phase information to the signals recorded by each microphone before applying transfer-function-gain NMF to all the signals recorded by an asynchronous microphone array. Employing the noise suppression based on phase information as the preprocessing is considered effective because it can improve the signal-to-noise ratio (SNR) of the input signals of transfer-function-gain NMF. Our main purpose is to show that on the assumption that the recording devices have stereo microphones, applying the beamformer to

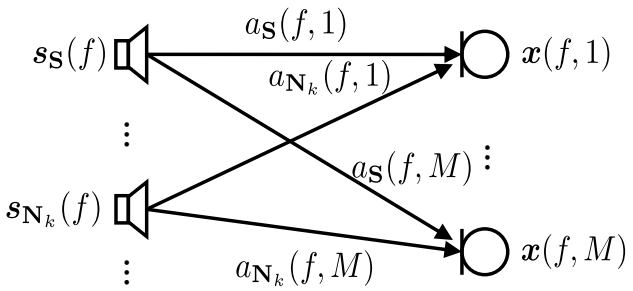


Fig. 1 Setup of observation

each stereo signal as a preprocessing step for the input signals of transfer-function-gain NMF contributes to the improvement of the total performance of noise suppression. In the experiments using measured impulse responses in a room ($T_R = 600$ ms), the proposed method outperformed the conventional methods of noise suppression using transfer-function-gain NMF in terms of the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR).

2. Conventional Method

Let us assume that one target speech signal and $K-1$ interfering speech signals are captured by M recording devices, each having stereo microphones, as shown in Fig. 1. The observed signals are expressed by

$$\boldsymbol{X}(f) = [\boldsymbol{x}(f,1), \cdots, \boldsymbol{x}(f,2M)]^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{x}(f,m') = [x(1,f,m'), \cdots, x(N,f,m')]^{\mathrm{T}} \tag{2}$$

where $x(n,f,m')$ denotes the short-time Fourier transform (STFT) coefficient of the observed signal obtained by the m'th microphone ($1 \leq m' \leq 2M$) at time n ($1 \leq n \leq N$) and frequency f ($1 \leq f \leq F$). The observed signals captured by the mth recording device are denoted by $\boldsymbol{x}(f,2m-1)$ and $\boldsymbol{x}(f,2m)$. The superscript T denotes non conjugate transposition. In the following, we first describe the conventional method of noise suppression using transfer-function-gain NMF.

By assuming the time invariance of the transfer function gain, namely, the amplitude of the transfer function, we express the mixture model in the amplitude domain by

$$\bar{\boldsymbol{X}}(f) \approx \bar{\boldsymbol{A}}(f)\bar{\boldsymbol{S}}(f) \tag{3}$$

$$\bar{\boldsymbol{A}}(f) = \left[\bar{\boldsymbol{a}}_{\mathbf{S}}(f), \bar{\boldsymbol{a}}_{\mathbf{N}_1}(f), \cdots, \bar{\boldsymbol{a}}_{\mathbf{N}_{K-1}}(f)\right] \tag{4}$$

$$\bar{\boldsymbol{a}}_j(f) = [\bar{a}_j(f,1), \cdots, \bar{a}_j(f,2M)]^{\mathrm{T}} \quad (j = \mathbf{S}, \mathbf{N}_k) \tag{5}$$

$$\bar{\boldsymbol{S}}(f) = \left[\bar{\boldsymbol{s}}_{\mathbf{S}}(f), \bar{\boldsymbol{s}}_{\mathbf{N}_1}(f), \cdots, \bar{\boldsymbol{s}}_{\mathbf{N}_{K-1}}(f)\right]^{\mathrm{T}} \tag{6}$$

$$\bar{\boldsymbol{s}}_j(f) = [\bar{s}_j(1,f), \cdots, \bar{s}_j(N,f)]^{\mathrm{T}} \quad (j = \mathbf{S}, \mathbf{N}_k) \tag{7}$$

where $\bar{a}_j(f,m')$ is the transfer function gain between source j and the m'th microphone at frequency f, and $\bar{s}_j(n,f)$ is the activation of the source j at time n and frequency f, with $\bar{\cdot}$ being the amplitude of the element. The bold characters $\mathbf{S}$ and $\mathbf{N}_k$ denote the target speech and the kth interfering speech ($1 \leq k \leq K-1$), respectively. We can obtain the matrices of the transfer function gain $\bar{\boldsymbol{A}}(f)$ and activations $\bar{\boldsymbol{S}}(f)$ of each source by applying transfer-function-gain NMF. NMF approximates one nonnegative matrix as two low-rank nonnegative matrices as

$$\bar{\boldsymbol{X}}(f) = \tilde{\boldsymbol{X}}(f) \approx \tilde{\boldsymbol{A}}(f)\tilde{\boldsymbol{S}}(f) \tag{8}$$

where $\tilde{\cdot}$ denotes the component estimated by NMF in the amplitude domain. In transfer-function-gain NMF, the parameters $\tilde{\boldsymbol{A}}(f)$ and $\tilde{\boldsymbol{S}}(f)$ are estimated by minimizing the distance between $\bar{\boldsymbol{X}}(f)$ and $\tilde{\boldsymbol{A}}(f)\tilde{\boldsymbol{S}}(f)$ in accordance with a certain distance regulation. In this method, we employ I-divergence as the distance regulation and the parameters are estimated with the multiple update rules as

$$\tilde{a}_j(f,m) \leftarrow \tilde{a}_j(f,m)\frac{\sum_n \frac{\bar{x}(n,f,m)\tilde{s}_j(n,f)}{\tilde{x}(n,f,m)}}{\sum_n \tilde{s}_j(n,f)} \quad (j = \mathbf{S}, \mathbf{N}_k) \tag{9}$$

$$\tilde{s}_j(n,f) \leftarrow \tilde{s}_j(n,f)\frac{\sum_m \frac{\bar{x}(n,f,m)\tilde{a}_j(f,m)}{\tilde{x}(n,f,m)}}{\sum_m \tilde{a}_j(f,m)} \quad (j = \mathbf{S}, \mathbf{N}_k) \tag{10}$$

When the m'th microphone is placed closest to the target source, the enhanced target signal $z(n,f)$ is obtained as follows by Wiener filtering:

$$z(n,f) = \lambda(n,f,m')x(n,f,m') \tag{11}$$

$$\lambda(n,f,m') = \frac{\tilde{x}^2_{\mathbf{S}}(n,f,m')}{\tilde{x}^2_{\mathbf{S}}(n,f,m') + \sum_k \tilde{x}^2_{\mathbf{N}_k}(n,f,m')} \tag{12}$$

$$\tilde{x}_{\mathbf{S}}(n,f,m') = \tilde{a}_{\mathbf{S}}(f,m')\tilde{s}_{\mathbf{S}}(n,f) \tag{13}$$

$$\tilde{x}_{\mathbf{N}_k}(n,f,m') = \tilde{a}_{\mathbf{N}_k}(f,m')\tilde{s}_{\mathbf{N}_k}(n,f) \tag{14}$$

where $\lambda(n,f,m')$ is the Wiener filter that enhances the target signal.

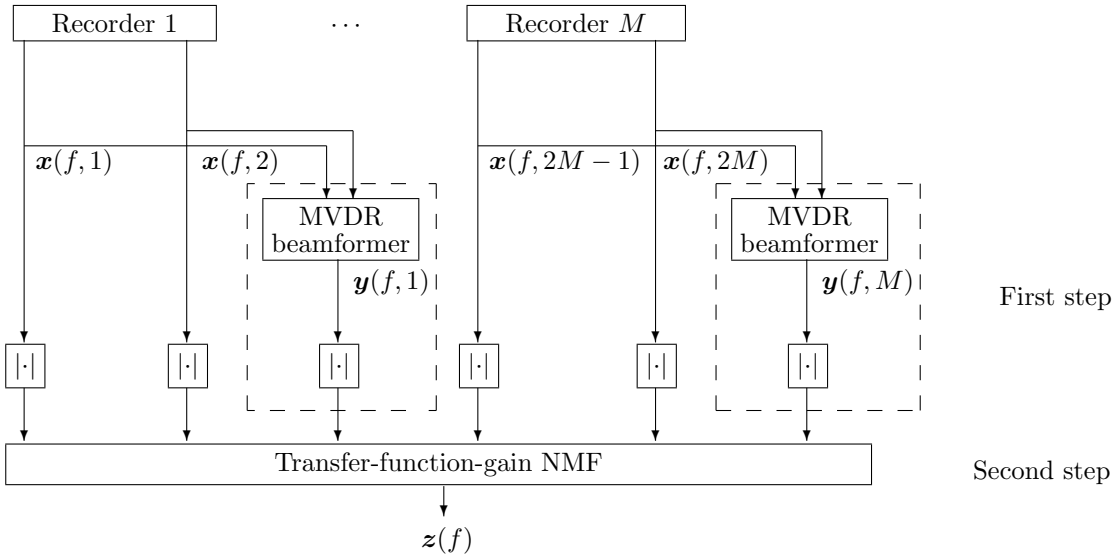


Fig. 2 Process flow of proposed noise suppression method

3. Proposed Noise Suppression Method

Figure 2 shows the process flow of the proposed noise suppression method. The proposed noise suppression method consists of two steps. In the first step, we apply noise suppression based on phase information to the synchronous signals captured by each recording device. Specifically, in this study, a maximum variance distortionless response (MVDR) beamformer [6,7] is employed. The signal of the mth recording device enhanced with an MVDR beamformer, $\boldsymbol{y}(f,m)$, is expressed by

$$\boldsymbol{y}(f,m) = [y(1,f,m),\cdots,y(N,f,m)]^{\mathrm{T}} \tag{15}$$

$$y(n,f,m) = \boldsymbol{w}^{\mathrm{H}}(f)\,[x(n,f,2m-1),x(n,f,2m)]^{\mathrm{T}} \tag{16}$$

$$\boldsymbol{w}(f) = [w(f,2m-1),w(f,2m)]^{\mathrm{T}} \tag{17}$$

where $\boldsymbol{w}(f)$ is the spatial filter estimated by the MVDR beamformer, and the superscript H denotes conjugate transposition. The spatial filter $\boldsymbol{w}(f)$ is computed to minimize the response of nontarget signals such that the target signal is never distorted. We consider that the MVDR beamformer is effective for preprocessing in transfer-function-gain NMF because it can enhance the target signal with a distortionless response. In the second step, we obtain the enhanced target signal by applying transfer-function-gain NMF. The input signals of transfer-function-gain NMF $\bar{\boldsymbol{X}}'(f)$ are expressed by

$$\bar{\boldsymbol{X}}'(f) = \underbrace{[\bar{\boldsymbol{x}}(f,1),\cdots,\bar{\boldsymbol{x}}(f,2M),\bar{\boldsymbol{y}}(f,1),\cdots,\bar{\boldsymbol{y}}(f,M)]^{\mathrm{T}}}_{3M \text{ channels}} \tag{18}$$

The amplitude spectrograms of the signals enhanced by beamformers, $\bar{\boldsymbol{y}}(f,m)$, are utilized as the additional input signals in transfer-function-gain NMF. In the proposed approach, we obtain the target signal by applying Wiener filtering (described in the previous section) to $\boldsymbol{y}(f,m)$, where the mth recording device is placed closest to the target source.

The proposed approach is expected to improve the performance of noise suppression in two ways.

- Application of the Wiener filter to the signals enhanced by beamformers, which are clearer signals than those captured by the recording devices (Effect A)
- Increase in the number of input signals in NMF and the information of the transfer function by utilizing the signals enhanced by beamformers (Effect B)

In our previous research, it was confirmed that the greater the number of input signals of transfer-function-gain NMF is, the greater the performance of the noise suppresion is [5]. Thus, we use not only the signals enhanced by the MVDR beamformer but also the obtained signals for applying transfer-function-gain NMF. Note that both the noise suppression approaches using transfer-function-gain NMF and an MVDR beamformer require single-source sections of the target and interferer signals; hence, the preprocessing of the MVDR beamformer requires no additional information.

Fig. 3 Experimental situation

4. Experimental Evaluation

We conducted experiments on noise suppression to evaluate the performance of our proposed approach. The mixture signals were generated by convolving clean speech [8] with the impulse responses measured in a real environment.

4.1 Experimental condition

In this experiment, we recorded time-stretched pulse (TSP) signals in a real environment to generate impulse responses in the room shown in Fig. 3. Figure 4 illustrates the arrangements of the loudspeakers and recording devices. We set up two situations where interferers were placed in the near field (Setting A) and in the far field (Setting B). The single speech is emitted from each loudspeaker in both settings. For Setting B, we applied the model of diffuse noise [5].

We evaluated the performance characteristics of the two methods, namely, the conventional supervised transfer-function-gain NMF (SNMF) described in [4,5] and the proposed method (Proposed), in terms of the SDR and SIR [9]. By comparing those two methods, we can confirm how Effect A described in the previous section affects the performance of noise suppression. In addition, we evaluated MVDR beamformers applied to the signals recorded by the device placed closest to the target source (MVDR) to examine how the performance of the MVDR beamformers affected the performance of the proposed method. Note that an MVDR beamformer constructed with stereo microphones can steer a spatial null in only one direction, which means that each MVDR beamformer suppresses the noise arriving from one direction. Table 1 shows the sampling frequency of each recording device. The setting of the sampling frequency takes into consideration the individual differences of the analog to digital converters [4, 5]. The other hyperparameters were set as shown in Table 2.

The enhanced signals of S1, S2, S3, and S4 were obtained by the Wiener filtering of the mixture signals of R1, R2, R3, and R4, respectively. In the experiments with Setting B, MVDR beamformers were applied to enhance the target signal, and the enhanced signal was obtained from the observed signals recorded by device R1 placed closest to the target source.

Table 1 Sampling frequency of each recording device

Device R1	16,000 Hz
Device R2	16,001 Hz
Device R3	16,002 Hz
Device R4	16,003 Hz

Table 2 Experimental conditions

Frame length of STFT	4,096
Frame overlap	50%
Signal length for noise suppression	10 s
Signal length for supervised training	10 s
Number of NMF iterations	50
Reverberation time (T_{60})	0.6 s

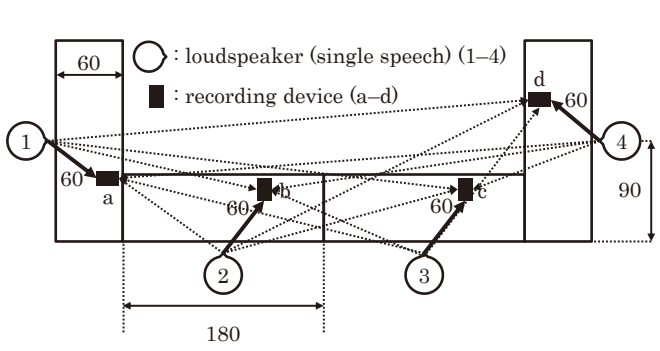
4.2 Evaluation results

Figures 5 and 6 show the SDRs and SIRs obtained in the experiments with Setting A and Setting B, respectively. The "Unproc" scores denote the SDR and SIR of the mixture signals without processing. The results show that the proposed approach outperformed the conventional noise suppression using transfer-function-gain NMF under both the near-field interferer and far-field diffuse noise conditions. Moreover, the greater the improvement of the SDR by the MVDR beamformer, the better the noise suppression performance of the proposed method. This indicates that the proposed method improved the performance of noise suppression because the SNR of the input signals was increased by applying the MVDR beamformer to the mixture signals. In other words, Effect A is crucial for improving the performance of noise suppression. Note that the effect of the MVDR beamformer is limited under underdetermined conditions such as in these experiments; the effect of applying the beamformer in the proposed method is reliable even if the effect of the MVDR beamformer is limited. These results confirmed that the proposed method is capable of improving the noise suppression performance for mixture signals observed by an asynchronous distributed microphone array consisting of recording devices having stereo microphones.

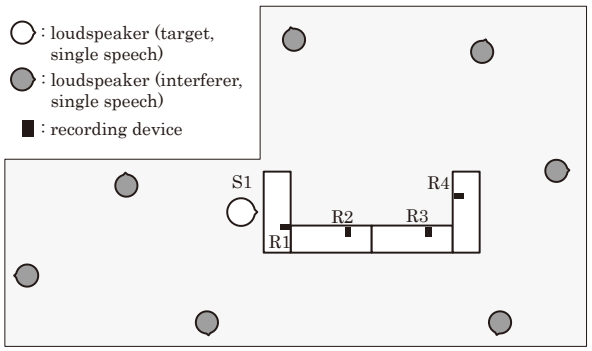
4.3 Confirmation of effectiveness of Effect B

To confirm how Effect B described in the previous section affects the performance of noise suppression, we conducted another experiment. In this experiment, we compared two patterns of the utilization of enhanced signals by MVDR beamformers as below.

- Prop (1): Utilize all enhanced signals by MVDR beamformers as additional input signals in



Setting A: Interferer in near field



Setting B: Interferer in far field

Fig. 4 Arrangements of loudspeakers and recording devices

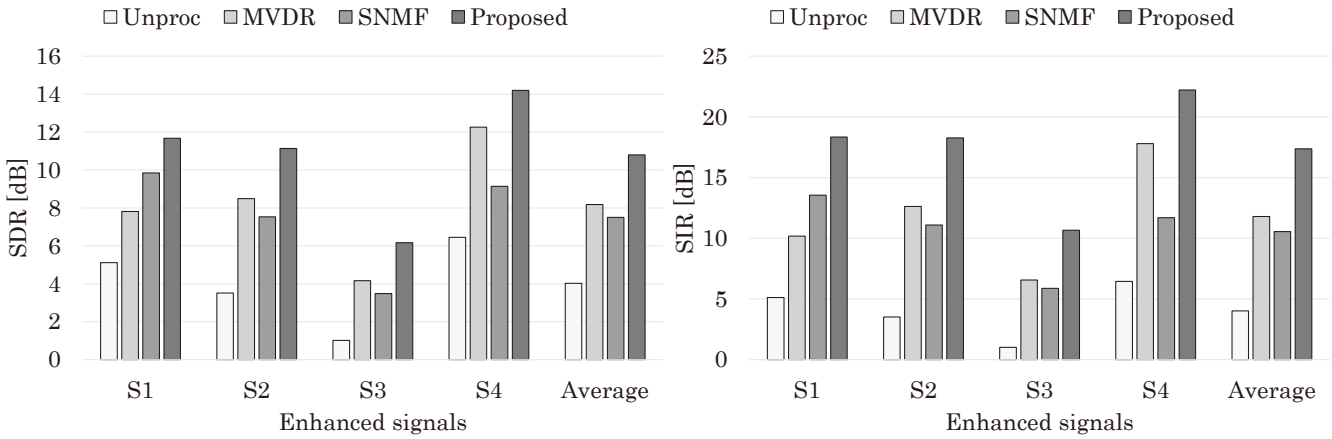


Fig. 5 SDR and SIR for Setting A

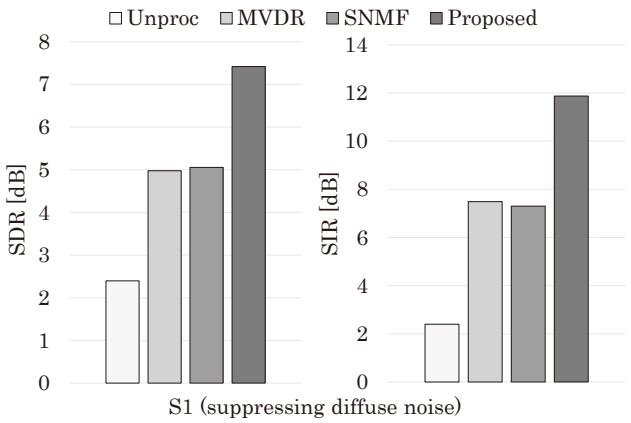


Fig. 6 SDR and SIR for Setting B

transfer-function-gain NMF ($+M$ input signals: $M = 4$ in this experiment)

- Prop (2): Only utilize the enhanced signal obtained from the recording device placed closest to the target source as additional input signals in transfer-function-gain NMF (+1 input signal)

Other conditions are the same as in the previous experiments.

Table 3 Results with Prop (1) and Prop (2)

	SDR [dB]		SIR [dB]	
Setting	Prop (1)	Prop (2)	Prop (1)	Prop (2)
(A)-S1	11.67	11.14	18.34	17.37
(A)-S2	11.14	10.94	18.27	17.74
(A)-S3	6.17	6.03	10.66	10.36
(A)-S4	14.20	14.21	22.21	22.10
(B)-S1	7.42	7.35	11.88	11.58

Table 3 shows the SDRs and SIRs obtained in the experiments with Setting A and Setting B, respectively. The results show that the performance of noise suppression was slightly improved by increasing the number of input signals in NMF. We confirmed that Effect B is not very crucial, but can improve the performance of noise suppression.

5. Conclusion

In this paper, we proposed a new approach to noise suppression using distributed recording devices with stereo microphones. The critical idea is to ap-

ply an MVDR beamformer to the synchronous stereo signals captured by each recording device to improve the SNR of the input signals of transfer-function-gain NMF so that transfer-function-gain NMF can more accurately estimate the signals. The signals enhanced by MVDR beamformers are utilized in transfer-function-gain NMF as extra input signals. Experiments using the mixture signals generated by convolving the impulse responses measured in a meeting room showed that the proposed method greatly outperformed the conventional noise suppression using transfer-function-gain NMF.

Acknowledgement

This work was supported by the Promotion of the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JP19H04131 and JP20H00613.

References

[1] E. Robledo-Arnuncio, T. S. Wada and B.-H. Juang: On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 34–37, Oct. 2007.

[2] Z. Liu: Sound source separation with distributed microphone arrays in the presence of clock synchronization errors, Proc. International Workshop on Acoustic Signal Enhancement, pp. 1–4, Sept. 2008.

[3] M. Togami, Y. Kawaguchi, H. Kokubo and Y. Obuchi: Acoustic echo suppressor with multichannel semi-blind nonnegative matrix factorization, Proc. Asia-Pacific Signal and Information Processing Association, pp. 522–525, Dec. 2010.

[4] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada and S. Makino: Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording, Proc. International Workshop on Acoustic Signal Enhancement, pp. 204–208, Sept. 2014.

[5] Y. Murase, H. Chiba, N. Ono, S. Miyabe, T. Yamada and S. Makino: Diffuse noise suppression with asynchronous microphone array based on amplitude additivity model, Proc. Asia-Pacific Signal and Information Processing Association, pp. 599–603, Dec. 2015.

[6] H. L. Van Trees, Ed. : Optimum Array Processing, Wiley, May 2002.

[7] O. L. Frost: An algorithm for linearly constrained adaptive array processing, Proc. IEEE, Vol. 60, No. 8, pp. 926–935, Aug. 1972.

[8] NTT Advanced Technology Corporation: Multilingual speech database, 1994.

[9] E. Vincent, R. Gribonval and C. Févotte: Performance measurement in blind audio source separation, IEEE Trans. on Audio, Speech, and Language Processing, Vol. 14, No. 4, pp. 1462–1469, July 2006.

Yutaro Matsui received his B.Sc. degree in computer and information science and M.E. degree from University of Tsukuba Japan, in 2017 and 2019, respectively. His research interests include array signal-processing and speech enhancement.

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.

Nobutaka Ono received his B.E., M.S., and Ph.D. degrees from the University of Tokyo, Japan, in 1996, 1998, and 2001, respectively. He became a research associate in 2001 and a lecturer at 2005 in the University of Tokyo. He moved to the National Institute of Informatics in 2011 as an associate professor, and then to Tokyo Metropolitan University in 2017 as a full professor. His research interests include acoustic signal processing, machine learning, and optimization algorithms for them. He is a member of the IEEE, IEICE, IPSJ, and ASJ.

Takeshi Yamada received his B. Eng. degree from Osaka City University, Japan, in 1994, and his M. Eng. and Dr. Eng. degrees from 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, and media quality assessment. He is a member of the IEEE, IEICE, IPSJ, and ASJ.

(Received September 14, 2021; revised June 29, 2022)