Diffuse noise suppression with asynchronous microphone array based on amplitude additivity model

Yoshikazu Murase*, Hironobu Chiba*, Nobutaka Ono†, Shigeki Miyabe*, Takeshi Yamada*, and Shoji Makino*
* University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
E-mail: {murase, chiba}@mmlab.cs.tsukuba.ac.jp, {miyabe, maki}@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp
† National Institute of Informatics / SOKENDAI (The Graduate University for Advanced Studies)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan, E-mail: onono@nii.ac.jp

***Abstract*—In this paper, we propose a method for suppressing a large number of interferences by using multichannel amplitude analysis based on nonnegative matrix factorization (NMF) and its effective semi-supervised training. For the point-source interference reduction of an asynchronous microphone array, we propose amplitude-based speech enhancement in the time-channel domain, which we call transfer-function-gain NMF. Transfer-function-gain NMF is a robust method against drift, which disrupts an inter-channel phase analysis. We use this method to suppress a large number of sources. We show that a mass of interferences can be modeled by a single basis assuming that the noise sources are sufficiently far from the microphones and the spatial characteristics become similar to each other. Since the blind optimization of the NMF parameters does not work well with merely sparse observation contaminated by the constant heavy noise, we train the diffuse noise basis in advance of the noise suppression using a speech absent observation, which can be obtained easily using a simple voice activity detection technique. We confirmed the effectiveness of our proposed model and semi-supervised transfer-function-gain NMF in an experiment simulating a target source that was surrounded by a diffuse noise.**

I. Introduction

An asynchronous microphone array is a new framework that can expand the applicability of microphone array signal processing [1]–[6]. This framework treats simultaneous recording with independent recording devices as a multichannel observation approach for array signal processing. The use of asynchronous recording devices has various advantages. First, we can easily construct a multi-channel microphone array, because we can use familiar portable recording devices such as smartphones, voice recorders and laptop computers. Second, we can record with high signal-to-noise ratios (SNRs) because we can select the number and placement of the microphones flexibly.

Unfortunately, an asynchronous microphone array also poses problems in addition to these benefits. One of the largest problems is the drift caused by sampling frequency mismatches between the channels. The drift changes the time differences of arrival of each source according to time, and degrades the noise suppression performance of array signal processing based on phase analysis [3], [4]. One straightforward approach is to synchronize the recording devices [5], [6]. However, these methods suffer from high computational costs or constraints as regards the recording manner. Thus for computational efficiency and unconstrained recording, we adopt another approach to employ noise suppression in the amplitude-spectrum domain discarding the phase to achieve robustness as regards synchronization error. Togami *et al.* proposed a method that uses nonnegative matrix factorization (NMF) [7] to estimate the transfer function gain (hereafter referred to as transfer-function-gain NMF) [8]. Since NMF has high potential for blind processing and also has considerable freedom in terms of modifying the model and training manner, we have proposed supervised [9] transfer-function-gain NMF to improve the interference suppression performance.

While the methods introduced above are for the suppression of point sources, in this paper we employ this method to suppress a large number of sources. We show that a mass of interferences can be modeled by a single basis assuming that the noise sources are sufficiently far from the microphones and the spatial characteristics become similar to each other. Since the blind optimization of the NMF parameters does not work well with only a sparse observation contaminated by a constant heavy noise, we train the diffuse noise basis in advance of the noise suppression using the speech absent observation, which can be obtained easily using a simple voice activity detection technique. We confirmed the effectiveness of our proposed model and semi-supervised transfer-function-gain NMF in an experiment simulating a target source surrounded by a diffuse noise.

II. Observed signal model of diffuse noise

A. Problem statement

Before discussing our proposed asynchronous observed signal model, let us begin with the signal modeling of synchronized observation assuming that one target source and K $(K \gg M)$ noise sources are recorded by M microphones. We indicate the components of the target and noise signals with the superscript symbols $\mathbf{S}(Signal)$ and $\mathbf{N}(Noise)$, respectively. Assuming that a simultaneous mixing model holds true in the time-frequency domain, the observed signal can be expressed by the sum of the target and noise signals:

$$\mathbf{X}(\omega) = \mathbf{X}^{\mathbf{S}}(\omega) + \mathbf{X}^{\mathbf{N}}(\omega), \tag{1}$$

where, $\mathbf{X}(\omega)$, $\mathbf{X}^{\mathbf{S}}(\omega)$ and $\mathbf{X}^{\mathbf{N}}(\omega)$ constitute a matrix with a size $M \times N$ and have the complex-values $X_{mn}(\omega)$, $X^{\mathbf{S}}_{mn}(\omega)$ and $X^{\mathbf{N}}_{mn}(\omega)$, respectively, in the (m, n) element. ω and N represent the frequency index and the number of time frames, respectively. Here, $\mathbf{X}^{\mathbf{S}}(\omega)$ is expressed by

$$\mathbf{X}^{\mathbf{S}}(\omega) = \mathbf{A}^{\mathbf{S}}(\omega)\mathbf{S}^{\mathbf{S}}(\omega), \tag{2}$$

where, $\mathbf{A}^{\mathbf{S}}(\omega)$ is a column vector with a size $M \times 1$ and the element of the vector $A^{\mathbf{S}}_m(\omega)$ shows the transfer function from the target source to the mth microphone. $\mathbf{S}^{\mathbf{S}}(\omega)$ is a row

vector with a size $1 \times N$ and the element of the vector $S_n^{\mathbf{S}}(\omega)$ shows the time-frequency component of the target source in the nth time frame.

In addition, the noise signal $\mathbf{X}^{\mathbf{N}}(\omega)$ can be expressed by the sum of the K noise sources as below.

$$\mathbf{X}^{\mathbf{N}}(\omega) = \sum_{k=1}^{K} \mathbf{A}^{\mathbf{N}_k}(\omega)\mathbf{S}^{\mathbf{N}_k}(\omega) \tag{3}$$

where, $\mathbf{A}^{\mathbf{N}_k}(\omega)$ is a column vector with a size $M \times 1$ and the element of the vector $A_m^{\mathbf{N}_k}(\omega)$ shows the transfer function from the kth noise source to the mth microphone. $\mathbf{S}^{\mathbf{N}_k}(\omega)$ is a row vector with a size $1 \times N$ and the element of the vector $S_n^{\mathbf{N}_k}(\omega)$ shows the time-frequency component of the kth noise source in the nth time frame.

The above model is valid for the case with the synchronous microphone array but invalid for that with the asynchronous microphone array. This is because, $X_{mn}(\omega)$ is affected by the sampling frequency mismatch among the recording devices as follows [10].

$$Y_{mn}(\omega) \approx X_{mn}(\omega)\exp(-\jmath\omega\epsilon_m n), \tag{4}$$

where, $\jmath$ and ϵ_m represent an imaginary number ($\jmath = \sqrt{-1}$) and the value of the drift from the 1st microphone $\epsilon_1 = 0$, respectively. For this reason, the asynchronous microphone array causes phase drift and the process becomes complicated. Therefore, we show a model that works without the phase information, and which can even be applied to asynchronous recording.

B. Mixing model in amplitude domain

Assuming the additivity of the amplitude in the frequency domain, the mixing model can be expressed by the product sum of the amplitude spectrum omitting the phase;

$$|\mathbf{X}(\omega)| \approx |\mathbf{X}^{\mathbf{S}}(\omega)| + |\mathbf{X}^{\mathbf{N}}(\omega)|. \tag{5}$$

Moreover, the target source $|\mathbf{X}^{\mathbf{S}}(\omega)|$ and the noise source $|\mathbf{X}^{\mathbf{N}}(\omega)|$ in the amplitude domain are expressed by

$$|\mathbf{X}^{\mathbf{S}}(\omega)| = |\mathbf{A}^{\mathbf{S}}(\omega)||\mathbf{S}^{\mathbf{S}}(\omega)|, \tag{6}$$

$$|\mathbf{X}^{\mathbf{N}}(\omega)| \approx \sum_{k=1}^{K} |\mathbf{A}^{\mathbf{N}_k}(\omega)||\mathbf{S}^{\mathbf{N}_k}(\omega)|, \tag{7}$$

where, $|\mathbf{A}^{\mathbf{S}}(\omega)|$ and $|\mathbf{A}^{\mathbf{N}_k}(\omega)|$ show the transfer function gain of the target and noise source, respectively. $|\mathbf{S}^{\mathbf{S}}(\omega)|$ and $|\mathbf{S}^{\mathbf{N}_k}(\omega)|$ show the absolute values of the amplitude of the target and noise source, respectively. This model is the conventional observed signal model [9] for the transfer-function-gain NMF and is only valid if the numbers of target and noise sources are already known. However, if the noise sources become diffuse noise that contains an indefinite number of noises, we cannot suppress the diffuse noise using the NMF described below with the conventional observed signal model. Therefore, we propose the observed signal model for suppressing the diffuse noise.

If the K noise sources act like the diffuse noise sources that arrive from far off and are scattered, the average energies of the noise sources in the amplitude domain are similar [11]. Therefore, we assume that the transfer function gain vectors of noise sources can be expressed as one common transfer function gain vector. And then we assume that the observed signal, the transfer function gain and the absolute value of the amplitude are expressed as follows.

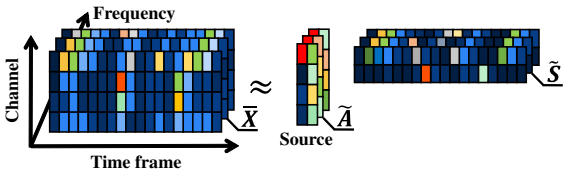

Fig. 1. Channel time-frequency domain representation of observed signals for each frequency bin

$$|\hat{\mathbf{X}}^{\mathbf{N}}(\omega)| \approx |\hat{\mathbf{A}}^{\mathbf{N}}||\hat{\mathbf{S}}^{\mathbf{N}}|, \tag{8}$$

$$|\hat{\mathbf{A}}^{\mathbf{N}}(\omega)| \approx |\mathbf{A}^{\mathbf{N}_1}(\omega)| \approx \ldots \approx |\mathbf{A}^{\mathbf{N}_K}(\omega)|, \tag{9}$$

$$|\hat{\mathbf{S}}^{\mathbf{N}}(\omega)| \approx \sum_{k=1}^{K} |\mathbf{S}^{\mathbf{N}_k}(\omega)|, \tag{10}$$

where, $|\hat{\mathbf{A}}^{\mathbf{N}}(\omega)|$ is a column vector with a size $M \times 1$ and the element of the vector $|\hat{A}_m^{\mathbf{N}}(\omega)|$ shows the transfer function gain from the diffuse noise to the mth microphone. $|\hat{\mathbf{S}}^{\mathbf{N}}(\omega)|$ is a row vector with a size $1 \times N$ and the element of the vector $|\hat{S}_n^{\mathbf{N}}(\omega)|$ shows the absolute value of the amplitude of the diffuse noise in the nth time frame. Accordingly, the observed signal model in the amplitude domain in a diffuse noise environment can be expressed by

$$|\mathbf{X}(\omega)| \approx |\mathbf{A}(\omega)||\mathbf{S}(\omega)|, \tag{11}$$

$$|\mathbf{A}(\omega)| \approx \left[|\mathbf{A}^{\mathbf{S}}(\omega)| \quad |\hat{\mathbf{A}}^{\mathbf{N}}(\omega)|\right], \tag{12}$$

$$|\mathbf{S}(\omega)| \approx \begin{bmatrix} |\mathbf{S}^{\mathbf{S}}(\omega)| \\ |\hat{\mathbf{S}}^{\mathbf{N}}(\omega)| \end{bmatrix}. \tag{13}$$

The purpose of this study is to suppress the diffuse noise signal and enhance the target signal. In particular, we assume that the absolute value of $A_1^{\mathbf{S}}(\omega)$ is at its highest in $A_j^{\mathbf{S}}(\omega), j = 1, \ldots, M$ because the 1st microphone can be placed closest to the target source utilizing the flexibility of asynchronous recording. Thus the diffuse noise suppression signal is given by suppressing the diffuse noise from the highest SNR signal with the first microphone. In the following, all the modeling and processing can be carried out at each frequency bin. Therefore, we omit ω for simplicity.

III. Diffuse noise suppression with NMF

A. Diffuse noise suppression with transfer-function-gain NMF

In this section, we describe diffuse noise suppression with the proposed observed signal model in the amplitude domain, which uses NMF to estimate the parameters of the model. The parameterization of the NMF is shown in Fig. 1: Typically, the decomposition of NMF in audio and acoustic signal processing [12]–[14] such as decomposition into spectral patterns and activations, is not used. NMF approximates a nonnegative matrix as two low rank nonnegative matrices as follows.

$$|\mathbf{X}| \approx \tilde{\mathbf{X}} = \tilde{\mathbf{A}}\tilde{\mathbf{S}}, \tag{14}$$

where, the tilde represents the decorated matrices or elements that are the values estimated by NMF in the amplitude domain. Here the distance measure between $|\mathbf{X}|$ and $\tilde{\mathbf{A}}\tilde{\mathbf{S}}$ can be customized for the task by choosing from a variety of functions that NMF can minimize. A low-rank approximation with a minimum distance means that the solution of $\tilde{\mathbf{S}}$ will be sparse due to the non-negative constraint. As a result, the estimation

of the source amplitude is accompanied by the identification of the transfer function gain and the source activation. In other words, $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{S}}$ estimate $|\mathbf{A}|$ and $|\mathbf{S}|$.

We cannot decide the size of the initial matrices of NMF using the conventional observed signal model because the number of sources is unknown in a diffuse noise environment. Even if we can set the initial matrices, we cannot expect to estimate NMF because the sources outnumber the microphones. By contrast, using the proposed observed signal model, we can decide the size of the initial matrices for NMF because the only sources are the target and diffuse noise. Moreover, if we use more than two microphones, we can avoid underdetermination and estimate the parameters with NMF.

The procedure for suppressing diffuse noise by transfer-function-gain NMF using the proposed observed signal model is as follows. First, we estimate the parameters $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{S}}$ with NMF. After that, a diffuse noise suppression signal Y is obtained with the observed signal X_{1n} of the $m = 1$ microphone and a Wiener mask as

$$Y = X_{1n} \frac{\left(\tilde{A}_1^{\mathbf{S}} \tilde{S}_n^{\mathbf{S}}\right)^2}{\left(\tilde{A}_1^{\mathbf{S}} \tilde{S}_n^{\mathbf{S}}\right)^2 + \left(\tilde{A}_1^{\mathbf{N}} \tilde{S}_n^{\mathbf{N}}\right)^2}. \quad (15)$$

This Wiener filtering reduces the error in $\tilde{\mathbf{S}}$ caused by the model mismatch of the linear modeling in the amplitude spectrum domain. We set the initial value of $\tilde{\mathbf{A}}^{\mathbf{S}}$ as

$$\tilde{A}_m^{\mathbf{S}} = \begin{cases} 1-(M-1)\alpha & (m=1) \\ \alpha & (m \neq 1) \end{cases} \quad (16)$$

because the assumed absolute value of $A_1^{\mathbf{S}}$ is the highest in $A_j^{\mathbf{S}}, j = 1, \ldots, M$. Here, α is an arbitrary positive number that satisfies $0 < \alpha < 1/(M-1)$. Also, we set the initial value of $\tilde{\mathbf{S}}^{\mathbf{N}}$ as

$$\tilde{A}_m^{\mathbf{N}} = \frac{1}{M} \ (m = 1, \ldots, M), \quad (17)$$

due to the diffusibility.

If the numbers of microphones and sources are similar, the estimation accuracy of NMF is degraded and higher performance cannot be expected. Thus, some methods for appending restrictions to the NMF have been proposed. Therefore, we propose a method for applying the observed signal model for the diffuse noise to the conventional method. Moreover, we propose the semi-supervised transfer function gain NMF training of only the diffuse noise transfer-function gain.

B. Supervised transfer-function-gain NMF

Supervised transfer-function-gain NMF is a noise suppression method that can estimate activation closer to the optimum solution by using the trained transfer function gain. We employ I-divergence as the distance regulation and each parameter is estimated with the following multiplicative update rules.

$$\tilde{A}_m^i \leftarrow \tilde{A}_m^i \frac{\sum_n \frac{|X_{mn}|\tilde{S}_n^i}{\tilde{A}_m^{\mathbf{S}}\tilde{S}_n^{\mathbf{S}} + \tilde{A}_m^{\mathbf{N}}\tilde{S}_n^{\mathbf{N}}}}{\sum_n \tilde{S}_n^i}, \quad (i = \mathbf{S}, \mathbf{N}) \quad (18)$$

$$\tilde{S}_n^i \leftarrow \tilde{S}_n^i \frac{\sum_m \frac{|X_{mn}|\tilde{A}_m^i}{\tilde{A}_m^{\mathbf{S}}\tilde{S}_n^{\mathbf{S}} + \tilde{A}_m^{\mathbf{N}}\tilde{S}_n^{\mathbf{N}}}}{\sum_m \tilde{A}_m^i}. \quad (i = \mathbf{S}, \mathbf{N}) \quad (19)$$

The procedure for suppressing diffuse noise by supervised transfer-function-gain NMF is as follows. The first step involves learning the basis vectors of the transfer function gain. The transfer function gain vectors $\tilde{\mathbf{A}}^{\mathbf{S}}$ and $\tilde{\mathbf{A}}^{\mathbf{N}}$ are obtained by pre-estimations with NMF. The pre-estimations require two observations containing only the single source duration of the target and the diffuse noise source, respectively. The second step is to estimate the activation vectors $\tilde{\mathbf{S}}^{\mathbf{S}}$ and $\tilde{\mathbf{S}}^{\mathbf{N}}$ of the observed signal. The activation vectors $\tilde{\mathbf{S}}^{\mathbf{S}}$ and $\tilde{\mathbf{S}}^{\mathbf{N}}$ are obtained by updating (19). Meanwhile, the transfer function gain vectors $\tilde{\mathbf{A}}^{\mathbf{S}}$ and $\tilde{\mathbf{A}}^{\mathbf{N}}$ are fixed to the trained transfer function gain vectors. Finally, we suppress the diffuse noise with (15) derived from the trained transfer function gain and estimated activation.

C. Semi-supervised transfer-function-gain NMF

In this section, we propose using the semi-supervised transfer-function-gain NMF, which trains only the transfer function gain vector of diffuse noise. Supervised transfer-function-gain NMF requires the single source duration of all sources. However, with a constantly noisy space such as outside, it is difficult to obtain the single source durations of the target source. Therefore, we propose semi-supervised transfer-function-gain NMF where we apply semi-supervised NMF in the time-frequency domain [16] to the transfer-function-gain NMF. In particular, the transfer function gain vector of the diffuse noise $\tilde{\mathbf{A}}^{\mathbf{N}}$ is only trained by the diffuse noise duration. After that, $\tilde{\mathbf{A}}^{\mathbf{N}}$ is fixed when (18) and (19) are updated.

D. Penalized transfer-function-gain NMF

Penalized transfer-function-gain NMF is a noise suppression method that introduces a sparse constraint into the source activation. The objective function is as follows.

$$\mathcal{J}(|\mathbf{X}|, \tilde{\mathbf{A}}\tilde{\mathbf{S}}) = \mathcal{D}_I(|\mathbf{X}||\tilde{\mathbf{A}}\tilde{\mathbf{S}}) + \lambda g(\tilde{\mathbf{S}}), \quad (20)$$

where $g(\tilde{\mathbf{S}})$ is a function for measuring the spatial sparseness of $\tilde{\mathbf{S}}$ and we employ $L_{0.5}$ norm. λ shows the non-negative weight and it is calibrated in conformance with to the observed signal. The penalized multiplicative update rule with I-divergence is given by

$$\tilde{A}_m^i \leftarrow \tilde{A}_m^i \frac{\frac{\Sigma_n |X_{mn}|\tilde{S}_n^i}{\tilde{A}_m^i \tilde{S}_n^i}}{\Sigma_n \tilde{S}_n^i}, \quad (i = \mathbf{S}, \mathbf{N}) \quad (21)$$

$$\tilde{S}_n^i \leftarrow \tilde{S}_n^i \frac{\Sigma_m \frac{|X_{mn}|\tilde{A}_k^i}{\tilde{A}_m^i \tilde{S}_n^i}}{\Sigma_m \tilde{A}_m^i + \lambda \nabla g(\tilde{S}_n^i)}, \quad (i = \mathbf{S}, \mathbf{N}) \quad (22)$$

where the $\nabla g(\tilde{S}_n^i)$ shows the gradient of $g(\tilde{S}_n^i)$. Penalized transfer-function-gain NMF suppresses the diffuse noise from the observed signal using parameters estimated with the above update rules.

IV. Experimental evaluation

A. Experimental condition

In this experiment, to confirm the effectiveness of the proposed model for the diffuse noise, we investigated the performance of methods where the proposed model was applied to conventional transfer-function-gain NMF. Moreover, we investigated the effectiveness of the semi-supervised transfer-function-gain NMF. Table 1 shows the experimental conditions. Each signal from each source to each microphone was given as a convolutive mixture of clean speech and impulse responses by using the image method [17]. To obtain the

TABLE I
EXPERIMENTAL CONDITIONS

Sampling frequency for synchronous recording	16,000 Hz
Frame length	4096 samples
Frame shift	2048 samples
Signal length for evaluation	10 sec
Signal length for supervised and semi-supervised NMF training	10 sec
Divergence	I-divergence
α (initialization parameter)	0.25
Number of NMF iterations	200
Reverberation time	0.3 sec
Signal to diffuse noise ratio	0, 5 dB

TABLE II
SAMPLING FREQUENCIES ON EACH PATTERN

	Patt. 1	Patt. 2	Patt. 3
16,000 Hz	Mic 1	Mic 1, 4	Mic 1, 4, 7
16,001 Hz	Mic 2	Mic 2, 5	Mic 2, 5, 8
16,002 Hz	Mic 3	Mic 3, 6	Mic 3, 6, 9

impulse response, we assumed all the microphones to be omnidirectional. Moreover, the experiment was conducted with 3, 6 and 9 microphones. The observed signals, recorded with the asynchronous microphone array, were given by the resampling of the synchronous data according to Table II. Figure 2 shows the arrangement of the microphones and sources. Here, we regard the noise sources as diffuse noise by keeping the noise sources at a distance from the microphones. The signal-to-distortion ratio (SDR) is used as the evaluation score [18]. The SDR evaluates the distortion of a noise suppressed signal. Higher SDR values indicate better noise suppression. We calculated the evaluation scores of the unprocessed observation (Unprocessed) and of three methods that employed our proposed observed signal model, 1) the supervised transfer-function-gain NMF (SNMF), 2) the semi-supervised transfer-function-gain NMF (SSNMF) and 3) the penalized transfer-function-gain NMF (PNMF). Here, the supervised transfer-function-gain NMF trained the transfer function gain vectors of the target and diffuse noise source assuming that we could obtain single-source sections containing only one source. In particular, the scores of supervised transfer-function-gain NMF showed the marginal performance of transfer-function-gain NMF. In addition, penalized transfer-function-gain NMF provided the optimum value for all values of parameter λ under each condition.

B. Results of evaluation experiment

Figure 3 shows the experimental results. (a) and (b) show the diffuse noise suppression performance (SNR = 0 dB) recorded with synchronous and asynchronous microphone arrays, respectively. (c) and (d) show the diffuse noise suppression performance (SNR = 5 dB) recorded by synchronous and asynchronous microphone arrays, respectively. In all the noise suppression methods employing our proposed observed signal model, the SDRs were higher than the unprocessed values. This result shows that these methods can suppress diffuse noise. The performance of semi-supervised transfer-function-gain NMF is better than that of penalized transfer-function-gain NMF under all conditions. When SNR = 0 dB, the performance of semi-supervised transfer-function-gain

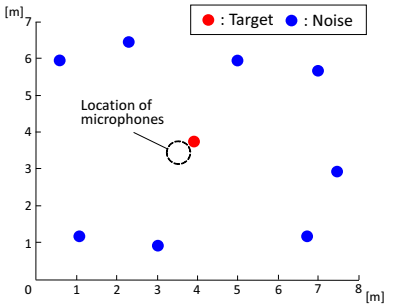

Fig. 2. Arrangement of speakers and microphones

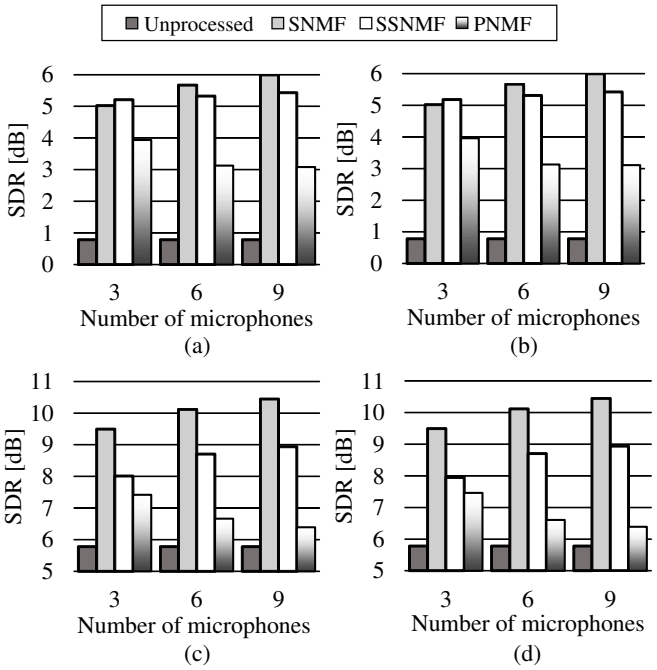

Fig. 3. SDRs with (a) synchronous recording SNR = 0 dB, (b) asynchronous recording SNR = 0 dB, (c) synchronous recording SNR = 5 dB and (d) asynchronous recording SNR = 5 dB

NMF was as good as that of supervised transfer-function-gain NMF. Moreover, a comparison of (a) with (b) and (c) with (d) shows there is little difference between the performance obtained with the synchronous and asynchronous arrays. Thus, these methods were robust against the drift caused by sampling frequency mismatches.

V. CONCLUSION

In this research, we proposed an observed signal model for suppressing diffuse noise. Moreover, we proposed semi-supervised transfer-function-gain NMF, which can suppress the diffuse noise in a constantly noisy environment. As a result, all the methods that employed the proposed observed model could suppress diffuse noise. Furthermore, semi-supervised transfer-function-gain NMF outperformed performance than the penalized transfer-function-gain NMF.

VI. ACKNOWLEDGMENT

This work was supported by Grant-in-Aid for Scientific Research (B) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 25280069).

REFERENCES

[1] M. Souden, K. Kinoshita, M. Delcroix and T. Nakatani, "Distributed microphone array processing for speech source separation with classifier fusion," *Proc. MLSP*, 2012.

[2] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," *Proc. SCVT*, 2011.

[3] 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. WASPAA*, pp. 34–37, 2007.

[4] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," *Proc. IWAENC*, 2008.

[5] S. Miyabe, N. Ono and S. Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain, " *Proc. ICASSP*, pp. 674-678, 2013.

[6] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada and S. Makino, "Speech enhancement with ad-hoc microphone array using single source activity, " *Proc. APSIPA*, pp. 1-6, 2013.

[7] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization, " *Proc. NIPS*, pp. 556-562, 2000.

[8] M. Togami, Y. Kawaguchi, H. Kokubo and Y. Obuchi, "Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization, " *Proc. APSIPA*, pp. 522-525, 2010.

[9] 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. IWAENC*, pp. 204-208, Sept. 2014.

[10] S. Miyabe, N. Ono and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation, " *Signal Processing*, vol. 107, pp. 185-196, Feb. 2015.

[11] N. Ito, H. Shimizu, N. Ono and S. Sagayama, "Diffuse noise suppression using crystal-shaped microphone arrays, " *IEEE Trans. on Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2101-2110, 2011.

[12] P. Smaragdis, and J. C. Brown, "Non-negative matrix factorization for music transcription, " *Proc. WASPAA*, pp. 177-180, 2003.

[13] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, " *IEEE Trans. on Audio, Speech & Language Processing*, vol. 15, no. 3, pp. 1066-1074, 2007.

[14] C. Févotte, N. Bertin and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis, " *Neural Computation*, vol. 21, no. 3, pp. 793-830, 2009.

[15] I. A. McCown and H. Bourlard, "Microphone array post-filter based on noise field coherence, " *IEEE Trans. on Audio, Speech & Language Processing*, vol. 11, no. 6, pp. 709-716, 2003.

[16] P. Smaragdis, B. Raj and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures, " *Proc. ICA*, pp. 414-421, 2007.

[17] E. A. P. Habets, "Room impulse response (RIR) generator," Available at: http:/www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator Oct. 2008.

[18] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation, " *IEEE Trans. on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462-1469, 2006.