Multiple Far Noise Suppression in a Real Environment Using Transfer-Function-Gain NMF

Yutaro Matsui*, Shoji Makino*, Nobutaka Ono†, Takeshi Yamada*

*University of Tsukuba, Email: y.matsui@mmlab.cs.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp
†National Institute of Informatics / SOKENDAI, Email: onono@nii.ac.jp

Abstract—Conventional noise suppression methods based on array signal processing use phase information and control the directivity of noises. However, such methods can hardly suppress so-called background noise, whose arrival direction cannot be specified. Thus, multiple far noise suppression based on transfer-function-gain non-negative matrix factorization (NMF) has been proposed as a method that can suppress such background noise. Its effectiveness has been confirmed by an experimental simulation using convolutional mixtures; however, it has not been verified that it is practical in a real environment. Thus, in this paper, we examine the performance of this method by recording a target and multiple far noises with asynchronous microphones in a real environment. We confirm that this method can suppress far noises in a real environment with diverse distances between microphones and interference sources.

I. INTRODUCTION

Many noise suppression methods based on array signal processing have been proposed to improve the quality of recorded speech [1]–[4]. They suppress noises by using phase information and controlling the directivity of noises. However, such methods have the following issues. First, their performance is degraded if they are employed for asynchronous recording. Asynchronous recording has recently attracted considerable interest because it is possible to easily and flexibly construct a multichannel microphone array using familiar portable recording devices such as mobile phones and voice recorders. However, in asynchronous recording, drifts are caused by differences in the recording start time or the sampling frequency mismatch among the channels. The drifts change the time difference of arrival of each source over time and degrade the performance of array signal processing based on phase analysis [5], [6]. Next, they can hardly suppress so-called background noise, whose arrival direction cannot be specified. These methods also use a number of sound sources and the arrival directions of noises. Therefore, they cannot suppress the background noises if their arrival directions and the number of sources are unknown.

Recently, a new method of noise suppression has been proposed [7] that solves the above issues. In this method, phase information is not necessary because array signal processing is employed in the amplitude domain. Also, by assuming that the background noises arrive from far away, such noises are modeled by a single basis. In other words, it is assumed that the observed signal is composed of two sources, a target source and a mixed noise source. Then transfer-function-gain non-negative matrix factorization (NMF) is employed with this mixing model. Note that the amplitude-spectrum beamformer [8] is also a method of noise suppression in the amplitude domain. Although the amplitude-spectrum beamformer requires advance learning for all voices contained in an observed signal using each single sound source, transfer-function-gain NMF does not require such learning. Therefore, this method is expected to be more suitable for background noises than the amplitude-spectrum beamformer.

The effectiveness of noise suppression based on transfer-function-gain NMF has been confirmed by a computer simulation using convolutional mixtures [7]; however, it has not been verified that it is practical in a real environment. Thus, in this paper, we verify the effectiveness of this method in a real environment by recording multiple noises asynchronously and employing this method of noise suppression. Although the computer simulation [7] realized asynchronization by artifically generating a sampling frequency mismatch, we recorded voices with separate IC recorders to naturally realize a sampling frequency mismatch. Our experiment evaluated the practical performance of the proposed method.

II. OBSERVED SIGNAL MODEL OF FAR NOISE

A. Amptitude-based mixing model

In this section, we describe the signal modeling of asynchronous observed signals. Before describing asynchronous observation, we begin with the synchronous observation of a target source and K noise sources by M microphones. The difference between synchronization and asynchronization is whether or not phase drifts are caused. In asynchronous recording, phase drifts are caused among the observed signals by the difference in the recording start time or the sampling frequency mismatch among the recording devices.

Then, the observed signals in the time-frequency domain can be expressed by (1) as the sum of the target signal $\boldsymbol{X}^{\mathbf{S}}(\omega)$ and the noise signal $\boldsymbol{X}^{\mathbf{I}}(\omega)$.

$$\boldsymbol{X}(\omega) = \boldsymbol{X}^{\mathbf{S}}(\omega) + \boldsymbol{X}^{\mathbf{I}}(\omega) \tag{1}$$

$\boldsymbol{X}(\omega), \boldsymbol{X}^{\mathbf{S}}(\omega)$ and $\boldsymbol{X}^{\mathbf{I}}(\omega)$ are matrices of size $M \times N$ and have the complex values $X_{mn}(\omega), X_{mn}^{\mathbf{S}}(\omega)$ and $X_{mn}^{\mathbf{I}}(\omega)$, respectively, for their (m, n) elements. ω and N represent the frequency index and the number of time frames, respectively. We indicate the components of the target signal and the noise signals with the superscripts $\mathbf{S}$ (*signal*) and $\mathbf{I}$ (*interference*), respectively. Then, $\boldsymbol{X}^{\mathbf{S}}(\omega)$ and $\boldsymbol{X}^{\mathbf{I}}(\omega)$ are expressed by

$$\boldsymbol{X}^{\mathbf{S}}(\omega) = \boldsymbol{a}^{\mathbf{S}}(\omega) s^{\mathbf{S}}(\omega), \tag{2}$$

$$\boldsymbol{X}^{\mathbf{I}}(\omega) = \sum_{k=1}^{K} \boldsymbol{a}^{\mathbf{I}_k}(\omega) s^{\mathbf{I}_k}(\omega), \tag{3}$$

where $\boldsymbol{a}^{\mathbf{S}}(\omega)$ and $\boldsymbol{a}^{\mathbf{I}_k}(\omega)$ are $M \times 1$ vectors and the elements $a_m^{\mathbf{S}}(\omega)$ and $a_m^{\mathbf{I}_k}(\omega)$ describe the transfer function from the target source or noise source k to microphone m, respectively. $\boldsymbol{s}^{\mathbf{S}}(\omega)$ and $\boldsymbol{s}^{\mathbf{I}_k}(\omega)$ are $1 \times N$ vectors and the elements $s_n^{\mathbf{S}}(\omega)$ and $s_n^{\mathbf{I}_k}(\omega)$ describe the time-frequency components of the target source or noise source k in time frame n, respectively. The above model is valid for a synchronous microphone array but invalid for an asynchronous microphone array, because $a_m(\omega)$ is affected by phase drifts and can be time-varying.
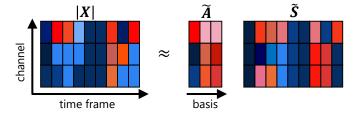
Fig. 1. Time-channel domain representation of observed signals for each frequency bin.

Although an asynchronous recording has phase drifts, we can assume that the amplitudes of the transfer function are time-invariant if the phase drifts are sufficiently less than the short-time Fourier transform (STFT) frame width. Therefore, assuming the additivity of the amplitudes in the frequency domain, the mixing model can be expressed by the product sum of the amplitude spectrum omitting the phase as follows:

$$|\boldsymbol{X}(\omega)| \approx \left|\boldsymbol{X}^{\mathbf{S}}(\omega)\right| + \left|\boldsymbol{X}^{\mathbf{I}}(\omega)\right|. \tag{4}$$

Such a mixing model in the power or amplitude domain is frequently assumed when NMF is employed [9]. Moreover, the target source $\left|\boldsymbol{X}^{\mathbf{S}}(\omega)\right|$ and the noise sources $\left|\boldsymbol{X}^{\mathbf{I}}(\omega)\right|$ in the amplitude spectrum are expressed by

$$\left|\boldsymbol{X}^{\mathbf{S}}(\omega)\right| = \left|\boldsymbol{a}^{\mathbf{S}}(\omega)\right| \left|\boldsymbol{s}^{\mathbf{S}}(\omega)\right|. \tag{5}$$

$$\left|\boldsymbol{X}^{\mathbf{I}}(\omega)\right| \approx \sum_{k=1}^{K} \left|\boldsymbol{a}^{\mathbf{I}_k}(\omega)\right| \left|\boldsymbol{s}^{\mathbf{I}_k}(\omega)\right|. \tag{6}$$

$\left|\boldsymbol{a}^{\mathbf{S}}(\omega)\right|$ and $\left|\boldsymbol{a}^{\mathbf{I}_k}(\omega)\right|$ are the transfer function gains of the target and the noise source, respectively. $\left|\boldsymbol{s}^{\mathbf{S}}(\omega)\right|$ and $\left|\boldsymbol{s}^{\mathbf{I}_k}(\omega)\right|$ are the absolute values of the amplitude of the target and the noise source, respectively. As described above, we introduce the mixing model in the amplitude domain. $\tilde{\boldsymbol{a}}(\omega)$ and $\tilde{\boldsymbol{s}}(\omega)$ are estimated from $|\boldsymbol{X}(\omega)|$ by employing transfer-function-gain NMF [10] in the time-channel domain as shown in Fig. 1. Note that such a mixing model in the power or amplitude domain has been frequently assumed in the NMF context. It is only valid to apply transfer-function-gain NMF to the model when the numbers of targets and noises are known. However, the number of noise sources is assumed to be unknown and thus we cannot suppress the noises. Therefore, in the next section we describe the mixing model [7] used to suppress such background noises.

B. Mixing model of far noise for suppressing background noises

In this section, we describe the mixing model [7] used to suppress the background noises. It is based on the idea that the background noises are composed of multiple far noise sources. Here, we assume that K background noises arrive from much farther than the target and are scattered. In this case, the average energies of the observed noise sources are typically similar in a diffuse noise field [11]. Therefore, we assume that the transfer function gains of all the noise sources are similar if they are far away, and the transfer function gain vectors of the noise sources can be expressed by a common vector. Thus, the observed signal $\left|\hat{\boldsymbol{X}}^{\mathbf{I}}\right|$, the transfer function

gain $\left|\hat{\boldsymbol{a}}^{\mathbf{I}}\right|$ and the absolute value of the amplitude $\left|\hat{\boldsymbol{s}}^{\mathbf{I}}\right|$ of such far noises can be expressed as

$$\left|\hat{\boldsymbol{X}}^{\mathbf{I}}(\omega)\right| \approx \left|\hat{\boldsymbol{a}}^{\mathbf{I}}(\omega)\right| \left|\hat{\boldsymbol{s}}^{\mathbf{I}}(\omega)\right|, \tag{7}$$

$$\left|\hat{\boldsymbol{a}}^{\mathbf{I}}(\omega)\right| \approx \left|\boldsymbol{a}^{\mathbf{I}_1}(\omega)\right| \approx \cdots \approx \left|\boldsymbol{a}^{\mathbf{I}_K}(\omega)\right|, \tag{8}$$

$$\left|\hat{\boldsymbol{s}}^{\mathbf{I}}(\omega)\right| \approx \sum_{k=1}^{K} \left|\boldsymbol{s}^{\mathbf{I}_k}(\omega)\right|. \tag{9}$$

According to the equations above, the observed signal model in the amplitude domain composed of a target and background far noises can be expressed by

$$|\boldsymbol{X}(\omega)| \approx |\boldsymbol{A}(\omega)| |\boldsymbol{S}(\omega)|, \tag{10}$$

$$|\boldsymbol{A}(\omega)| \approx \left[\left|\boldsymbol{a}^{\mathbf{S}}(\omega)\right| \ \left|\hat{\boldsymbol{a}}^{\mathbf{I}}(\omega)\right|\right], \tag{11}$$

$$|\boldsymbol{S}(\omega)| \approx \left[\begin{array}{c} \left|\boldsymbol{s}^{\mathbf{S}}(\omega)\right| \\ \left|\hat{\boldsymbol{s}}^{\mathbf{I}}(\omega)\right| \end{array} \right]. \tag{12}$$

We conduct background noise suppression on the basis of the above observation. In particular, we assume that the value of $\left|a_1^{\mathbf{S}}(\omega)\right|$ is the highest among $\left|a_j^{\mathbf{S}}(\omega)\right| (j = 1, \cdots, M)$ by placing microphone 1 of the microphone array closest to the target source, utilizing the flexibility of asynchronous recording. This assumption is necessary because this method distinguishes between the target and the noise by considering the difference in the transfer-function gain among the channels. In the following, all the modeling and processing can be carried out in each frequency bin. Therefore, we omit ω for simplicity.

III. FAR NOISE SUPPRESSION WITH NMF

A. Noise suppression using transfer-function-gain NMF

In this section, we describe noise suppression using the above observed signal model. This noise suppression employs transfer-function-gain NMF in the time-channel domain to estimate the parameters of the model. In this method, the typical decomposition of NMF in audio and acoustic signal processing [12]–[14], such as decomposition into spectral patterns and activations, is not used. The parameterization of the NMF is shown in Fig. 1.

NMF approximates a non-negative matrix as two low-rank non-negative matrices as follows [9]:

$$|\boldsymbol{X}| \approx \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{S}}. \tag{13}$$

The tilde represents the matrices or elements estimated by NMF. In such a low-rank approximation, the solutions will be sparse owing to the non-negative constraint. Therefore, the transfer function gain $\tilde{\boldsymbol{A}}$ and the source activation $\tilde{\boldsymbol{S}}$ are identified along with the estimation of the source amplitudes $|\boldsymbol{A}|$ and $|\boldsymbol{S}|$. Furthermore, NMF minimizes the distance between $|\boldsymbol{X}|$ and $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{S}}$. In this method, I-divergence is employed as the distance regulation. Each parameter is estimated using the following multiplicative update rules:

$$\tilde{a}_m^i \leftarrow \tilde{a}_m^i \frac{\sum_n \dfrac{|X_{mn}| \tilde{s}_n^i}{\tilde{a}_m^{\mathbf{S}} \tilde{s}_m^{\mathbf{S}} + \tilde{a}_m^{\mathbf{I}} \tilde{s}_m^{\mathbf{I}}}}{\sum_n \tilde{s}_n^i} \quad (i = \mathbf{S}, \mathbf{I}), \tag{14}$$

$$\tilde{s}_m^i \leftarrow \tilde{s}_m^i \frac{\sum_n \dfrac{|X_{mn}| \tilde{a}_n^i}{\tilde{a}_m^{\mathbf{S}} \tilde{s}_m^{\mathbf{S}} + \tilde{a}_m^{\mathbf{I}} \tilde{s}_m^{\mathbf{I}}}}{\sum_n \tilde{a}_n^i} \quad (i = \mathbf{S}, \mathbf{I}). \tag{15}$$

Moreover, the initial values of $\tilde{a}^{\mathbf{S}}$ are set as (16)

$$\tilde{a}_m^{\mathbf{S}} = \begin{cases} 1 - (1 - M)\alpha & (m = 1) \\ \alpha & (\text{otherwise}) \end{cases}, \qquad (16)$$

where α is an arbitrary number that satisfies $0 < \alpha < 1/(M-1)$. This is because it is assumed that the value of $\left|a_1^{\mathbf{S}}(\omega)\right|$ is the highest among $\left|a_j^{\mathbf{S}}(\omega)\right|$ $(j = 1, \cdots, M)$. Also the initial values of $\tilde{a}^{\mathbf{I}}$ are given by

$$\tilde{a}_m^{\mathbf{I}} = \frac{1}{M} \quad (m = 1, \cdots, M), \qquad (17)$$

Then, the enhanced signal $\tilde{Y}_n$ is obtained from the observed signal of microphone 1 (X_{1n}) and a Wiener mask as

$$\tilde{Y}_n = X_{1n} \frac{\left(\tilde{a}_1^{\mathbf{S}} \tilde{s}_n^{\mathbf{S}}\right)^2}{\left(\tilde{a}_1^{\mathbf{S}} \tilde{s}_n^{\mathbf{S}}\right)^2 + \left(\tilde{a}_1^{\mathbf{I}} \tilde{s}_n^{\mathbf{I}}\right)^2}. \qquad (18)$$

B. Semi-supervised transfer-function-gain NMF

If the numbers of microphones and sources are similar, high performance cannot be expected for the parameter estimation described in the previous section because the estimation accuracy of NMF is degraded. Some methods have been proposed to solve this problem including semi-supervised transfer-function-gain NMF [7]. In this method, semi-supervised NMF in the time-frequency domain [15] is applied to transfer-function-gain NMF. In particular, the transfer function gain vector of the background noises $\tilde{a}^{\mathbf{I}}$ is obtained by training using NMF. The training requires the observation of the single-source duration of the background noises. After that, the transfer function gain vector of the target $\tilde{a}^{\mathbf{S}}$ and the activation vectors $\tilde{s}^{\mathbf{S}}$ and $\tilde{s}^{\mathbf{I}}$ are obtained by updating (14) and (15) with $\tilde{a}^{\mathbf{I}}$ fixed.

Note that supervised transfer-function-gain NMF [16] is also an effective method, which requires the training of $\tilde{a}^{\mathbf{S}}$ and $\tilde{a}^{\mathbf{I}}$. Although the performance of supervised NMF is higher than that of semi-supervised NMF, it is difficult to obtain the single-source durations of the target in a constantly noisy environment. Therefore, semi-supervised NMF is employed in this paper.

IV. EXPERIMENTAL EVALUATION

A. Experimental conditions

In this experiment, we confirm the effectiveness of the mixing model and transfer-function-gain NMF in a real environment. Table I shows the differences between the conditions in the simulation [7] and this experiment. The other conditions are almost equivalent to those in the simulation [7]. Figure 2 shows the arrangement of the target source, the interference sources and the microphones, and Fig. 3 shows the microphone arrangement. As shown in Fig. 2, the distances between the microphones and the interference sources vary between 2 m and 3.5 m. The microphone placed closest to the target source is microphone 1. The signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) [18] are used as the evaluation scores. We calculated the evaluation scores of the unprocessed observation (*unproc*) and of the observation processed by the method employing semi-supervised transfer-function-gain NMF with 3 channels, 6 channels and 9 channels (*3 ch, 6 ch, 9 ch*) as shown in Fig. 3.

Table I
EXPERIMENTAL CONDITIONS

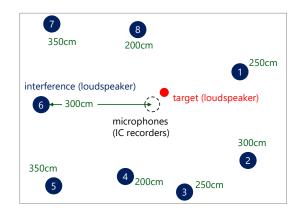|	simulation [7]	this experiment
observed signals	convolutive mixture of RIR [17] and clean speech	speech recorded with IC recorders
reverberation time	0.3 s	0.64 s
background noise except interference	——	35.3–35.8 dB(A)
cause of phase drift	sampling frequency mismatch among channels by artificial resampling	difference in recording start times and sampling frequency mismatch among IC recorders
sampling frequency mismatch	within 3 Hz	within 0.1 Hz

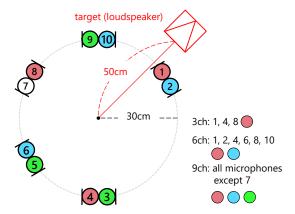Fig. 2. Arrangement of target, interference sources and microphones.

Fig. 3. Arrangement of microphones (IC recorders).

B. Experimental results

Figure 4 shows the experimental results. In transfer-function-gain NMF employing the observed signal model, the SDRs and SIRs are higher than those of the unprocessed observation. From the results for *unproc* and *3 ch*, we confirm that this method can supress the noises by assuming that the observed signal is composed of a target and a mixed noise and that it can avoid underdetermination. Furthermore, according to the results for *3 ch, 6 ch* and *9 ch*, we confirm that the
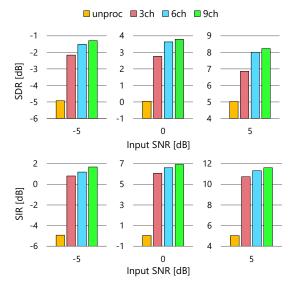
Fig. 4. Experimental results.

greater the number of microphones used to suppress noises, the greater the performance of noise suppression, similarly to in the simulation [7].

The tendency of the results in this experiment and the simulation is similar. However, the improvements in the SDR and SIR in this experiment are less than those in the simulation [7]. This is because the distance between the microphones and the interference sources in this experiment is shorter than that in the simulation. Also, according to Table I, the conditions, such as the reverberation time, the background noise and the sampling frequency mismatch, in this experiment are less favorable than those of the simulation. Thus, we can conclude that the background noises are effectively suppressed in a real environment by the proposed method.

C. Estimation of $\tilde{A}$

We investigate the estimation of the transfer-function gain $\tilde{A}$ ($\tilde{a}^{\mathrm{S}}$ and $\tilde{a}^{\mathrm{I}}$) to confirm the effectiveness of processing multiple noises as a single basis. We employ the transfer-function-gain NMF to a single sound comprising all sources (1 target and 8 noises), assuming that we know the numbers of sources and transfer-function gain bases we need. Note that the sensitivity is almost equal among all the microphones, and that the values do not express the true observation because they are normalized in the range of each source. Figure 5 shows the observation of $\tilde{A}$ with 3, 6 and 9 channels. The transfer-function gain of the target has a greater difference among the channels than that of the noises, and all the transfer-function gains of the noises are similar among the channels. The entry "ch1" is the transfer-function gain of microphone 1, which is placed closest to the target. Then "ch1" is by far the highest transfer-function gain of the target, but regarding the noises, an exceptionally high value is not found. Therefore, we consider that the assumptions given by (10) – (12) clearly hold. Furthermore, the dispersion of the transfer-function gain of the noises is different between "Estimation by 3 channels" and "Estimation by (6 or 9) channels" in Fig. 5. This is because the minimum distance between the IC recorders is different (3 channels: 52[cm], 6 and 9 channels: 30[cm]). It is clear that the longer the distances among the microphones, the longer the distances necessary between the microphone array and the noise source.
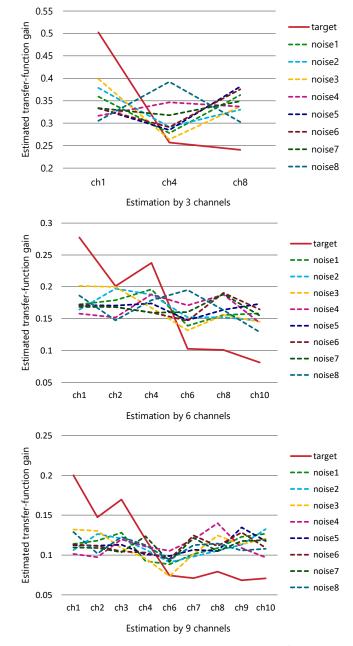

Fig. 5. Estimated transfer function gains (the components of $\tilde{A}(\tilde{a}^{\mathrm{S}}, \tilde{a}^{\mathrm{I}})$) in 3ch, 6ch and 9ch case from top to bottom.

V. CONCLUSION

In this paper, we examined the performance of multiple far noise suppression using transfer-function-gain NMF by recording a target and far noises with asynchronous microphones in a real environment. We confirmed that this method can suppress background far noises effectively in a real environment, regardless of the diversity of distances between the microphones and interference sources.

REFERENCES

[1] J. L. Flanagan, D. A. Berkley, G. W. Elko, and M. M. Sondhi, "Autodirective microphone systems," *Acustica*, pp. 58–71, Feb. 1991.

[2] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust. Speech Signal Process.*, pp. 1391–1400, Dec. 1986.

[3] D. H. Dugeon and D. E. Johnson, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Jan. 1993.

[4] H. L. Van Trees, Ed., *Optimum Array Processing*, Wiley, May 2002.

[5] 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, Oct. 2007.

[6] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," *Proc. IWAENC*, pp. 1–4, Sept. 2008.

[7] 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. APSIPA*, pp. 599–603, Dec. 2015.

[8] T. Kako, K. Kobayashi, and H. Ohmuro, "Speech enhancement using amplitude-spectrum beamformer for an asynchronous distributed microphone array," *IEICE Technical Report*, pp. 51–56, Jan. 2014.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Proc. NIPS*, pp. 556–562, Dec. 2001.

[10] 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, Dec. 2010.

[11] N. Ito, H. Shimizu, N. Ono, and S. Sagayama, "Diffuse noise suppression using crystal-shaped microphone arrays," *IEEE TASLP*, vol. 19, no. 7, pp. 2101–2110, Sept. 2011.

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

[13] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE TASLP*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

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

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

[16] 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.

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

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, July 2006.