Wavelength Proportional Arrangement of Virtual Microphones Based on Interpolation/Extrapolation for Underdetermined Speech Enhancement

Ryoga Jinzai*, Kouei Yamaoka*, Mitsuo Matsumoto*, Shoji Makino*, and Takeshi Yamada*
*University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
Email: {s1820656@s, yamaoka@mmlab.cs, makilab-research@tara, maki@tara, takeshi@cs}.tsukuba.ac.jp

***Abstract*—We previously proposed the virtual microphone technique to improve the speech enhancement performance in underdetermined situations, in which the number of channels is virtually increased by estimating extra observed signals at arbitrary positions along the straight line formed by real microphones. In our previous work, the effectiveness of the interpolation of virtual microphone signals for speech enhancement was experimentally confirmed. In this study, to examine the effectiveness of the extrapolation of a virtual microphone in improving the speech enhancement performance, we apply this technique to speech enhancement using a maximum signal-to-noise ratio (SNR) beamformer. Next, to improve the speech enhancement performance on the basis of the virtual microphone technique, we propose a new arrangement where a virtual microphone is placed in a position proportional to the wavelength. From the results of an experiment in an underdetermined situation, we confirmed that the proposed method markedly improves speech enhancement performance. Moreover, we present directivity patterns to confirm the behavior of each method of positioning the virtual microphone.**

***Index Terms*—array signal processing, virtual microphone, speech enhancement, underdetermined situation, beamforming**

I. Introduction

Signal processing using a microphone array includes various techniques such as blind source separation [1], direction of arrival estimation [2], and speech enhancement using a beamformer. Basically, the performance of these techniques depends on the number of microphones. Performance degradation may arise in a situation where the number of microphones is smaller than that of sound sources, which is called an underdetermined situation. Although several methods such as time–frequency masking [3] and multichannel Wiener filtering [4] can work well in an underdetermined situation, if more microphones are available, the performance of these techniques will be better.

Therefore, to improve the performance in an underdetermined situation, we previously proposed the virtual microphone technique, in which the number of channels is virtually increased [5]–[8]. In this technique, extra observed signals are estimated at arbitrary positions along the straight line formed by the real microphones. Signal processing using a virtually extended microphone array is possible by using virtual microphone signals in addition to the observed signals of the real microphones. This technique involves the interpolation and extrapolation of a virtual microphone depending on its position of placement. In our previous study, the interpolation was mainly used for speech enhancement [5], [6], and the extrapolation was mainly used for sound image localization [7], [8]. However, speech enhancement performance with the extrapolation of the virtual microphone has not been examined yet.

In general, if the microphone interval is small, the observed phase difference between microphones is small and it thus becomes difficult to construct an optimal spatial filter at low frequencies. On the other hand, with a long microphone interval, the spatial filter steers better nulls. However, a long interval tends to cause spatial aliasing at high frequencies. Thus, there is a trade-off with the array interval length in array signal processing techniques. However, if the microphone interval is half of the wavelength at each frequency, the observed phase differences should be sufficient to construct a spatial filter without spatial aliasing; and thus, this problem can be solved. In an actual microphone array, since the arrangement of microphones is fixed, changing the arrangement of the microphones for each frequency is unrealistic. On the other hand, since signal processing is performed independently for each frequency bin in the virtual microphone technique, it is possible to change the position of the virtual microphone.

Therefore, first, to examine the effectiveness of the extrapolation of a virtual microphone in speech enhancement, we perform experiments using a maximum signal-to-noise ratio (SNR) beamformer [9], [10] and an extrapolated virtual microphone signal. Next, based on the interpolation and extrapolation of the virtual microphone, we propose a new arrangement of the virtual microphone, in which it placed at a position depending on the wavelength for each frequency for speech enhancement using a beamformer.

II. Increasing number of channels by virtual microphone technique for maximum SNR beamformer

A. Interpolation and Extrapolation of Virtual Microphone Signals

In this section, we introduce the virtual microphone technique involving interpolation based on β-divergence [5], [6] and extrapolation of the virtual microphone [8]. In this technique, all microphone signals are processed in the short-time Fourier transform (STFT) domain. A virtual microphone signal, $v(\omega, t)$, is generated from the observed signals from

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

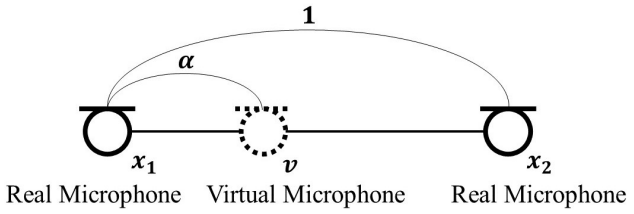


Fig. 1. Arrangement of real and virtual microphones in interpolation technique.

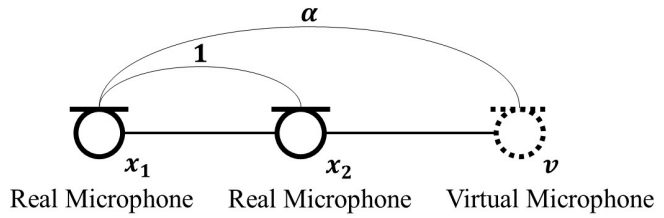


Fig. 2. Arrangement of real and virtual microphones in extrapolation technique.

two real microphones $x_i(\omega, t)$ in the time–frequency domain, where $x_i(\omega, t)$ is the ith microphone signal ($i = 1, 2$) at the angular frequency ω in the tth time frame. The number of channels of the microphone array is virtually increased by using real and virtual microphones. The arrangement of the real and virtual microphones is shown in Figs. 1 and 2, where α is a coefficient that determines the position of the virtual microphone.

In an environment where there are multiple sounds arriving from different directions, the relationship between the microphone position and waveform is generally complicated. In this method, by assuming the W-disjoint orthogonality (W-DO) [3] of the observed signals, we can simplify the modeling of the relationship. W-DO indicates the strong sparsity of a signal in the time–frequency domain, assuming that the component from a sound source dominates one time–frequency bin. By assuming W-DO, when multiple sounds arrive, we can regard them as a single sound in each time–frequency bin.

In this technique, the phase and amplitude of a virtual microphone signal are estimated individually. Herewith, different models are applied for the phase and amplitude estimation, and each formulation will be simplified. Additionally, the interpolation and extrapolation satisfy nonlinearity, which is requisite for applying linear signal processing. The phase and amplitude of the observed signals $x_i(\omega, t)$ are respectively defined as

$$\phi_i = \angle x_i(\omega, t) = \tan^{-1} \frac{\mathrm{Im}(x_i(\omega, t))}{\mathrm{Re}(x_i(\omega, t))}, \tag{1}$$

$$A_i = |x_i(\omega, t)|. \tag{2}$$

In both the interpolation and extrapolation, we can estimate the phase ϕ_v of the virtual microphone signal using the linear equation

$$\begin{aligned} \phi_\mathrm{v} &= \phi_1 + \alpha(\phi_2 - \phi_1) \\ &= (1 - \alpha)\phi_1 + \alpha\phi_2. \end{aligned} \tag{3}$$

The values of the phase is arbitrary for a natural number n in $\phi_i \pm 2\pi n$. Thus, the phase of the virtual microphone signal is estimated with the assumption that

$$|\phi_1 - \phi_2| \leq \pi. \tag{4}$$

For the estimation of the amplitude of the virtual microphone signal, since the expressions are different for the interpolation and extrapolation, we explain them separately. The appropriate interpolation of the amplitude of the virtual microphone depends on many conditions such as the direction of arrival and reverberation. Therefore, it is difficult to faithfully model the actual amplitude attenuation. Instead of using a physical model, amplitude interpolation based on β-divergence, which have a ease of processing and parameter adjustment, was proposed and its effectiveness was confirmed [5], [6]. Therefore, this method also uses β-divergence for amplitude interpolation. The amplitude of the virtual microphone is interpolated as

$$A_\mathrm{v} = \begin{cases} \exp\left((1 - \alpha)\log A_1 + \alpha \log A_2\right) & (\beta = 1) \\ \left((1 - \alpha)A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (\text{otherwise}). \end{cases} \tag{5}$$

For the extrapolation, the conceivable amplitude of the virtual microphone is more complex than that for the interpolation. When (5) is diverted to extrapolation, it may output unrealistic amplitudes such as a complex amplitude except when $\beta = 1$, a negative amplitude, or an amplitude diverging positive infinity. Therefore, in this study, as the simplest way to avoid these problems, we use the amplitude of the real microphone closest to the virtual microphone position as the amplitude of the extrapolated virtual microphone.

$$A_\mathrm{v} = \begin{cases} A_1 & (\alpha < 0) \\ A_2 & (\alpha > 1). \end{cases} \tag{6}$$

From the above, the virtual microphone signal $v(\omega, t, \alpha)$ is represented as

$$v(\omega, t, \alpha) = A_\mathrm{v} \exp(j\phi_\mathrm{v}), \tag{7}$$

where we can use an arbitrary number of α values to generate virtual microphones.

B. Wavelength-Proportional Arrangement of Virtual Microphones

In this paper, we propose setting the position of the virtual microphone in proportion to the wavelength for each frequency bin. In this method, the coefficient of the position of the virtual microphone (α) is denoted as

$$\alpha(\omega) = \frac{2\pi \mathrm{c} k}{\omega \mathrm{d}}, \tag{8}$$

where c is the velocity of sound, d is the distance between the real microphones, k is the scaling of the interval between reference microphone and the virtual microphone relative to wavelength. This equation implies that the virtual microphone is placed at a position k times the wavelength corresponding to the processing frequency; thus, the total length of the

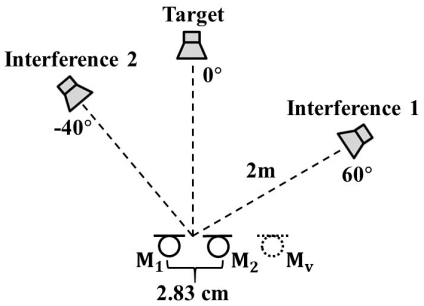


Fig. 3. Layout of sound sources and microphones in experiment.

microphone array including the virtual microphone is larger at low frequencies and smaller at higher frequencies. Spatial aliasing does not occur and a sufficient phase difference between microphones can be obtained at all frequencies by setting an appropriate parameter k. In this paper, we call this proposed technique the wavelength-proportional arrangement of virtual microphones.

C. Maximum SNR Beamformer

In this study, to evaluate the performance of the virtual microphone technique, we apply increasing the number of channels using the virtual microphone technique to a maximum SNR beamformer [9], [10]. The maximum SNR beamformer requires prior information on the covariance matrices of the target-only period and interference-only period. From the prior information of the target and interference, the maximum SNR beamformer constructs a filter so that the power ratio of the target to the interference becomes maximum. The advantage of using the maximum SNR beamformer is that it does not explicitly require the direction of sound sources. In principle, the virtual microphone technique can be similarly applied to other microphone array signal processing techniques as well as maximum SNR beamformer.

III. EXPERIMENTS

In these experiments, we examined the enhancement performance of the maximum SNR beamformer using the extrapolated virtual microphone signal. In addition, we evaluated the speech enhancement performance of the maximum SNR beamformer using the wavelength-proportional arrangement of virtual microphones.

A. Experimental Conditions

We prepared a target sound source, two interference sources, and two real microphones. In this environment, we conducted speech enhancement using the maximum SNR beamformer with the virtual microphone technique.

The layout of the sound sources and microphones is shown in Fig. 3, and other experimental conditions are listed in Table I. M_1 and M_2 are the real microphones, and M_v is the virtual microphone. The target sound source was eight speeches, and each of the interference sound sources was two speeches. In total, 16 combinations of target and interference sound samples were used for the experiments. The types of sound were Japanese and English speeches. The observed signals of each real microphone were generated by convolving the measured impulse responses into speech signals. In this work, we used the impulse responses in the RWCP Sound Scene Database [11].

TABLE I
EXPERIMENTAL CONDITIONS

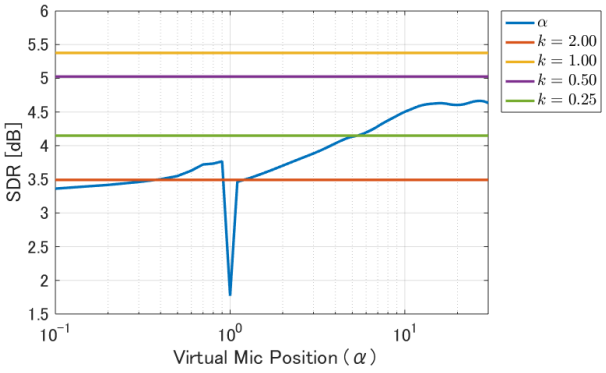
Number of real microphones	2
Number of virtual microphones	1
Input SNR	0 dB
Sampling rate	8 kHz
Interval between real microphones	2.83 cm
Reverberation time	300 ms
FFT frame length / shift	1024 / 256 samples
Number of target speech types	8
Number of interference speech types	2

In the experiment to compare interpolation and extrapolation, the coefficient of the position of the virtual microphone (α) was varied from 0.1 to 30, where $0 < \alpha < 1$ indicates the interpolation and $\alpha > 1$ extrapolation. $\alpha = 1$ indicates that no virtual microphone was used (i.e., only the two real microphones were used). In the interpolation, since it has been experimentally shown that $\beta = 20$ provided the highest performance in a previous study [5], we set β to 20 in this study. In the experiment to evaluate the wavelength-proportional arrangement of virtual microphones, the parameter k was set to 2, 1, 0.5, and 0.25. The SNR of the target signal to interference signals was set to 0 dB. To evaluate the speech enhancement performance, we used the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) as the objective evaluation criteria [12].

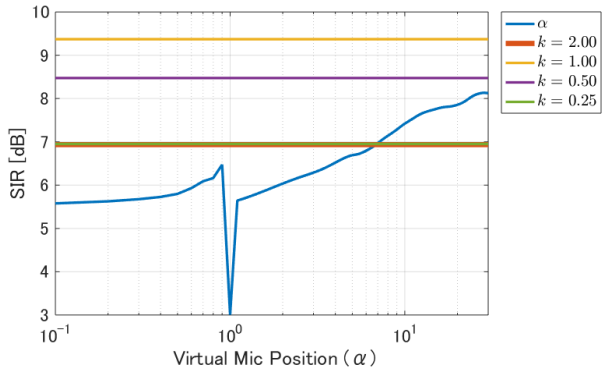
B. Results and Discussion

We evaluated the speech enhancement performance using the average of the results of 16 target and interference speech combinations. Figure 4 shows the relationship between the position of the virtual microphone and the speech enhancement performance. The blue line indicates the speech enhancement performance using the interpolation and extrapolation of the virtual microphone, and the other lines indicate the speech enhancement performance using the wavelength-proportional arrangement of virtual microphones. Note that the horizontal axis is a logarithmic scale.

The blue line shows the enhancement performance of each value of α. By comparing the enhancement performance between the case of interpolation ($\alpha < 1$) and the extrapolation ($\alpha > 1$), we found that SDR was improved by up to 2 dB compared with that without the virtual microphone ($\alpha = 1$) by using interpolation, whereas it was improved by up to about 3 dB by using extrapolation. Thus, the enhancement performance using the extrapolation is 1 dB better than that using interpolation in terms of SDR. Similarly, SIR was improved by 3.5 dB and 5 dB by using interpolation and extrapolation, respectively. From these results, it can be seen that the extrapolation of the virtual microphone is more effective than interpolation in this situation.

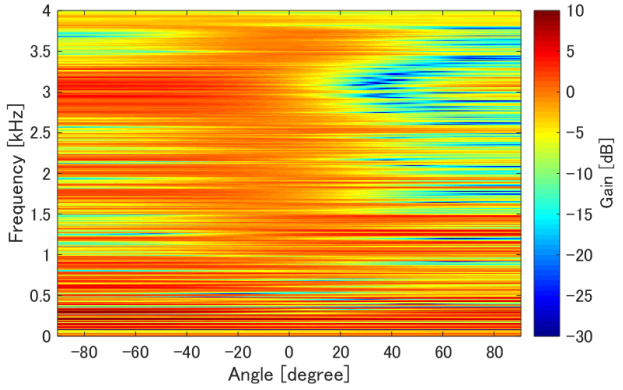


(a) SDR

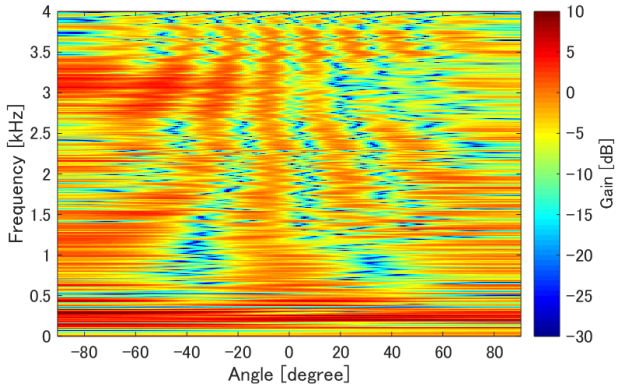


(b) SIR

Fig. 4. Relationship between position of virtual microphone and speech enhancement performance.



(a) Interpolation of virtual microphone ($\alpha = 0.9$)



(b) Extrapolation of virtual microphone ($\alpha = 13$)

Fig. 5. Directivity patterns of beamformer with interpolation and extrapolation of virtual microphone.

In the evaluation of the wavelength-proportional arrangement of virtual microphones, for both evaluation criteria, it is confirmed that the performance is the highest for $k = 1$. Although the enhancement performance for $k = 0.5$ is inferior to that for $k = 1$, it is better than the enhancement performance of extrapolation. On the other hand, the result for $k = 2$ and $k = 0.25$ may be inferior to the conventional method depending on the value of α. This means that the wavelength-proportional method with $k = 2$ and $k = 0.25$ are not always better than the extrapolation.

To clarify the reason underlying these results, we confirm the directivity patterns of the beamformer for one pattern out of 16 combinations. The directivity patterns of the beamformer obtained using interpolation, extrapolation, and the wavelength-proportional arrangement of the virtual microphone are shown in Figs. 5 and 6, where the best values of α and k were selected. In the interpolation of the virtual microphone (Fig. 5(a)), nulls are generated in the frequency range from 1 to 4 kHz and no nulls exist at frequencies below 1 kHz. This means that sounds below 1 kHz cannot be sufficiently suppressed. In the extrapolation of the virtual microphone (Fig. 5(b)), many nulls with a narrow angle range are generated, implying the occurrence of spatial aliasing. As a result, in addition to the interference sound direction, sounds from various directions such as those in the vicinity of the target sound are suppressed. However, unlike in interpolation, it can be seen that nulls exist even at frequencies below 1 kHz, and this means that sounds below 1 kHz can be appropriately suppressed. Since the sounds are speech, it is considered that the enhancement performance of extrapolation, which can suppress sounds of frequency below 1 kHz, is better than that of the interpolation.

In the directivity patterns of the beamformer with the wavelength-proportional arrangement of virtual microphones (Fig. 6), many nulls with a narrow angle range are generated for $k = 2$ (Fig. 6(a)). This indicates the occurrence of spatial aliasing at all frequencies. On the other hand, two fuzzy nulls are generated for $k = 0.25$ (Fig. 6(c)). For $k = 1$ (Fig. 6(b)), which shows the best speech enhancement performance, two belt-shaped nulls are clearly generated and it seems that no spatial aliasing occurs. As the reason for the improvement of the speech enhancement performance, by setting $k = 1$, it is possible to obtain the maximum phase difference within a range where spatial aliasing does not occur, thereby making it possible to generate sharp nulls in the correct direction.

From the results, the left-side nulls tend to be in the -40 degrees direction, which is the same direction as that of the interference sound source, and the right-side nulls tend to slightly deviate from the direction of the interference sound source. We attribute this to the effect of reverberation, which

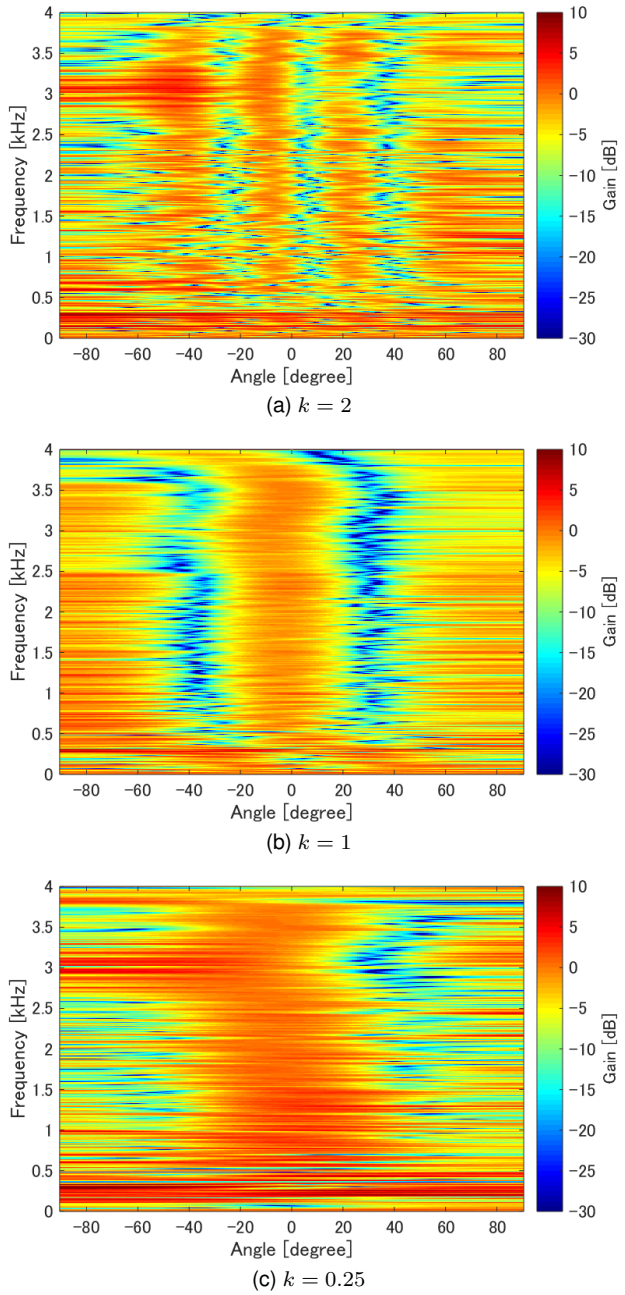


Fig. 6. Directivity pattern of beamformer with the wavelength-proportional arrangement of virtual microphone.

is known to introduce bias.

Taken together, it is considered that when the microphone interval is small, an insufficient phase difference between microphones exists at low frequencies, resulting in nulls not being properly generated, whereas when the microphone interval is large, spatial aliasing occurs at high frequencies. The wavelength-proportional arrangement of virtual microphones using an appropriate parameter can overcome these two problems; thus, this method shows the highest performance.

IV. Conclusion

In this study, we applied the maximum SNR beamformer with an extrapolated virtual microphone signal to speech enhancement in an underdetermined situation. In addition, we proposed a new arrangement where a virtual microphone is placed in a position proportional to the wavelength. The advantage of this method is that no spatial aliasing occurs and the phase difference between microphones is always sufficient to construct a spatial filter at all frequencies by setting an appropriate parameter.

In the experiment, we evaluated the enhancement performance on the basis of SDR and SIR in an underdetermined situation. By comparing the proposed method with the conventional method, we found that the SDR was improved by about 1.5 dB and the SIR by about 3 dB. From these results, the proposed method is effective for speech enhancement using the maximum SNR beamformer in an underdetermined situation.

Acknowledgment

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI through a Grant-in-Aid for Scientific Research under Grants 16H01735 and 19H04131, and the SECOM Science and Technology Foundation.

References

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.
[2] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans*, vol. 34, no. 3, pp. 276–280, 1986.
[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time–frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.
[4] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.
[5] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–8, Jan. 2016.
[6] K. Yamaoka, N. Ono, T. Yamada, and S. Makino, "Performance evaluation of nonlinear speech enhancement based on virtual increase of channels in reverberant environments," in *Proc. EUSIPCO*, pp. 2388–2392, Aug. 2017.
[7] N. Mae, K. Yamaoka, Y. Mitsui, M. Matsumoto, S. Makino, D. Kitamura, N. Ono, T. Yamada, and H. Saruwatari, "Ego noise reduction and sound localization adapted to human ears using hose-shaped rescue robot," in *Proc. International Workshop on Nonlinear Circuits, Communications and Signal Processing*, pp. 371–374, March 2018.
[8] R. Jinzai, K. Yamaoka, M. Matsumoto, T. Yamada, and S. Makino, "Microphone position realignment by extrapolation of virtual microphone," in *Proc. APSIPA ASC 2018*, pp. 367–372, Nov. 2018.
[9] H. L. Van Trees, *Optinum Array Processing*, John Wiley & Sons, 2002.
[10] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in meeting situation with maximum SNR beamformers," *Proc. ICASSP*, vol. I, pp. 41–45, 2007.
[11] S. Nakamura, K. Hiyane, F. Asano, Y. Kaneda, T. Yamada, T. Nishiura, T. Kobayashi, S. Ise, and H. Saruwatari, "Design and collection of acoustic sound data for hands-free speech recognition and sound scene understanding," in *Proc. ICME2002*, Vol. 2, pp. 161–164, Aug. 2002.
[12] 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.