Abnormal sound detection by two microphones using virtual microphone technique

Kouei Yamaoka*, Nobutaka Ono†‡, Shoji Makino* and Takeshi Yamada*
* University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577, Japan
E-mail: yamaoka@mmlab.cs.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp
† National Institute of Informatics (NII), 2-1-2 Hitotsubashi, Chiyoda, Tokyo, 101-8430, Japan
E-mail: onono@nii.ac.jp
‡ SOKENDAI (The Graduate University for Advanced Studies), Hayama, Japan

***Abstract*—In this paper, we propose a new method of microphone array signal processing for detecting abnormal sounds that is applicable to monitoring elderly people at home and can be implemented on small equipment. This method consists of noise reduction based on a subspace method and sound activity detection (SAD), which is the same as voice activity detection using the signal power. The performance of noise reduction may degrade for underdetermined conditions (the number of microphones is less than that of sound sources). To resolve this issue, we previously proposed a technique of microphone array signal processing that introduced virtual microphones. In this method, signals of virtual microphones are interpolated with those of real ones. By using both real and virtual microphones, this noise reduction method can be applied for a critical / overdetermined condition. In this paper, we apply this method to the subspace method for the first time. After noise reduction, the abnormal sounds can be detected by the SAD method. We conducted an experiment and confirm that the proposed method is effective for detecting abnormal sounds in noisy environments and is robust to abnormal sound directions.**

I. Introduction

Abnormal sound detection is an important task in surveillance or monitoring. For instance, the detection of a voice calling for help or the sound of broken glass at home will contribute to the safety and security of elderly people living alone. The detection of abnormal sounds is also useful for finding problems with machines in a factory.

Abnormal sound detection can be straightforward in a quiet environment. However, in many cases, the detection of abnormal sounds must be carried out under the existence of normal sound sources. For example, in a home environment, the television may always be on and water may sometimes be running in a kitchen, whereas in a factory, several machines normally make sounds. An abnormal sound detection system must be robust to such normal sound sources.

If the locations or directions of the normal sound sources are fixed, spatial information is effective for distinguishing between normal and abnormal sounds (e.g., [1]). A simple way to detect abnormal sounds in such a case is to suppress all known normal sound sources by null beamforming or adaptive beamforming. Then, when an abnormal sound occurs in a different direction from the normal sound sources, it can be easily detected. However, this is only possible when there is a sufficient number of microphones. If there are N normal sound sources, we need $(N+1)$ microphones to suppress all of them in conventional array processing techniques. Because commonly used small recording devices such as IC recorders have only stereo (two) channels, two-channel processing should be more convenient.

If an abnormal sound is a human voice, it can be detected accurately using the information that the sound is a human voice such as its harmonic structure. Voice activity detection (VAD) techniques have been subjected to extensive studies and can be used to detect abnormal sounds. However, because an abnormal sound is not always a human voice, we also have to detect abnormal sounds such as the breaking of glass.

In this paper, we propose a new technique of signal processing that combines noise reduction and sound activity detection (SAD) with two microphones. This technique can detect various abnormal sounds under the existence of normal sounds. Noise reduction is based on a subspace method using our previously proposed virtual microphone technique [2]–[4]. We expect to be able to effectively suppress the normal sounds with two microphones. For SAD, we employ VAD using signal power as a feature value to detect various abnormal sounds. Hereafter, we refer the VAD as SAD. To evaluate the performance of the proposed method, we conducted an experiment simulating a system used to monitor elderly people. We evaluate the performance and the robustness of abnormal sound detection in normal sound environments.

II. New abnormal sound detection combining virtual microphone technique and subspace-based noise reduction

A. Technical approach

There are four conditions that make the detection of abnormal sounds complicated. 1) There are some interfering normal sounds, 2) the direction of arrival (DOA) of abnormal sounds is unknown, 3) the type of abnormal sound is also unknown and 4) we can use only two microphones in small equipment.

In this paper, we apply a noise reduction technique to resolve condition 1. Because of condition 2, we use the noise reduction technique which does not require prior information about abnormal sounds. After the noise reduction, to detect target sounds (abnormal sounds) under condition 3, we apply an SAD technique that does not use the information that the

target sound is a human voice. Moreover, to satisfy condition 4, we use the noise reduction technique and SAD technique which works effectively by using only two microphones.

In our proposed method, a microphone signal is modeled in the short-time Fourier transform (STFT) domain. Here, let $x_i(\omega, t)$ be the ith real microphone signal $(i = 1, 2)$ at angular frequency ω in the tth frame. The amplitude of $x_i(\omega, t)$ is denoted as $A_i = |x_i(\omega, t)|$ and the phase is denoted as $\phi_i = \angle x_i(\omega, t)$.

B. Noise reduction using subspace method

In this paper, we use a subspace method as a noise reduction technique that satisfies condition 2. Speech enhancement based on a subspace method was proposed in [5], for example, and the process is as follows. Although we focus on stereo recording in this paper, we virtually increase the number of channels as described later. Therefore, we here consider that we have M channels as observation. When we have M channel input signal $\boldsymbol{x}$ (STFT domain), eigenvalue decomposition is performed for the spatial correlation matrix $\mathbf{R}$ of the normal sound period as prior training,

$$\mathbf{R} = E[\boldsymbol{x}\boldsymbol{x}^H], \quad (1)$$

$$\mathbf{R}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i, \quad (2)$$

where $\boldsymbol{e}_i = (e_{1i}, \cdots, e_{Mi})^T$ $(i = 1, \cdots, M)$ denotes the eigenvectors, λ_i denotes the eigenvalues, $\{\cdot\}^T$ denotes the transpose and $\{\cdot\}^H$ is the Hermitian transpose. In the formulation in this paper, we suppress the normal sounds to detect abnormal sounds. Therefore, we obtain the spatial correlation matrix of the normal sound period as prior information.

If we have more microphones than the number of sound sources N $(M > N)$, we obtain N dominant eigenvalues and $(M - N)$ non-dominant eigenvalues. The eigenvectors corresponding to the dominant eigenvalues are the basis of the signal subspace of the sound source contained in $\mathbf{R}$, and the other eigenvectors are orthogonal to the signal subspace. Therefore, we project the observed signals to the non-dominant eigenvectors $\boldsymbol{e}_{nd}$ as

$$y = \boldsymbol{e}_{nd}^H \boldsymbol{x}, \quad (3)$$

By projecting the observed signals to $\boldsymbol{e}_{nd}$, the normal sounds are suppressed. Then, the abnormal sound included in the observed signals is also projected to $\boldsymbol{e}_{nd}$. Although the abnormal sound may contain some distortion, it is not suppressed because $\boldsymbol{e}_{nd}$ is not orthogonal to the signal subspace of the abnormal sound.

This technique satisfies condition 2 because it requires no prior information about the abnormal sounds. Under condition 4 $(M = 2)$, however, the spatial correlation matrix of the normal sounds period becomes a two-by-two matrix. Thus, this technique can suppress only one normal sound. Generally, $(M - 1)$ sources can be suppressed using M microphones. Therefore, we expand this technique to be applicable to N sources using two microphones.

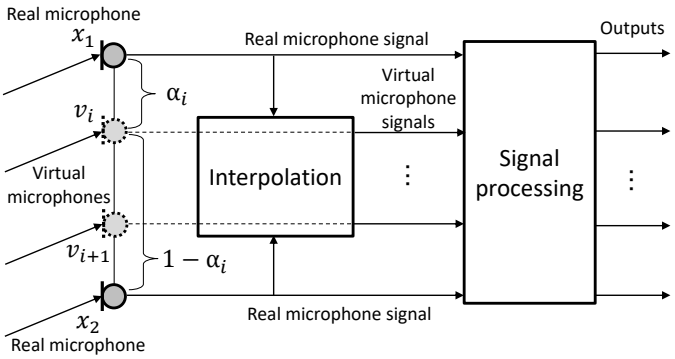


Fig. 1: Microphone array signal processing with virtual increase of channels

C. Application of virtual microphone technique to subspace method

As a technique for realizing high speech enhancement performance using only two microphones, we previously proposed a virtual increase of channels based on *virtual microphone* signals [2]–[4]. In this technique, we create arbitrary channels of virtual microphone signals by using two channels of real microphones. We perform microphone array signal processing using microphone signals consisting of both real and virtual microphone signals (Fig. 1). This technique is applicable to various types of microphone array signal processing since we generate virtual signals in the audio signal domain, which is different from techniques in which signals are generated in the power domain [6]–[8] or a higher-order statistical domain [9], [10].

In this paper, the virtual microphone technique is applied to subspace-based noise reduction for the first time. Note that since this technique assumes W-disjoint orthogonality [11], [12], mixture signals must be sparse. If they are not sparse, this technique may fail to interpolate virtual microphones correctly at many time-frequency bins, decreasing the performance of signal processing.

A virtual microphone signal $v(\omega, t, \alpha)$ is defined as the observation estimated at the point obtained by internally dividing the line joining two real microphones in the ratio $\alpha : (1 - \alpha)$. Hereafter, when there is no need to distinguish ω, t and α, the signal is simply denoted as v. The virtual microphone signal v is obtained by a nonlinear interpolation in each time-frequency bin as follows. We derive the amplitude A_{v} that minimizes the sum $\sigma_{\boldsymbol{D}_\beta}$ of the β-divergence between the amplitudes of a real microphone signal and a virtual microphone signal weighted by the virtual microphone interpolation parameter α,

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

$$A_{\mathrm{v}\beta} = \mathrm{argmin}_{A_{\mathrm{v}}} \sigma_{\boldsymbol{D}_\beta}, \quad (5)$$

where $\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_i)$ is defined as

$$\boldsymbol{D}_\beta(A_\mathrm{v}, A_i) = \begin{cases} A_\mathrm{v}(\log A_\mathrm{v} - \log A_i) + (A_i - A_\mathrm{v}) & (\beta = 1), \\ \dfrac{A_\mathrm{v}}{A_i} - \log \dfrac{A_\mathrm{v}}{A_i} - 1 & (\beta = 0), \\ \dfrac{A_\mathrm{v}^\beta}{\beta(\beta-1)} + \dfrac{A_i^\beta}{\beta} - \dfrac{A_\mathrm{v} A_i^{\beta-1}}{\beta - 1} & (\text{otherwise}). \end{cases} \quad (6)$$

By differentiating $\sigma_{\boldsymbol{D}_\beta}$ with respect to A_v and setting it to 0, the interpolated amplitude extended using the β-divergence is obtained as

$$A_{\mathrm{v}\beta} = \begin{cases} \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right) & (\beta = 1), \\ \left((1-\alpha)\, A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (\text{otherwise}). \end{cases} \quad (7)$$

Note that $A_{\mathrm{v}\beta}$ is continuous at $\beta = 1$ because

$$\begin{aligned} A_{\mathrm{v}1} &= \lim_{\beta \to 1} \left((1-\alpha)\, A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} \\ &= \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right), \end{aligned} \quad (8)$$

and this interpolation is equivalent to *complex logarithmic interpolation* [2].

The phase ϕ_v of a virtual microphone signal v is interpolated linearly in each time-frequency bin as

$$\phi_\mathrm{v} = (1-\alpha)\,\phi_1 + \alpha\phi_2, \quad (9)$$

and this interpolation is valid when there is no spatial aliasing. From (7) and (9), the virtual microphone signal v is represented as

$$v = A_{\mathrm{v}\beta} \exp(j\phi_\mathrm{v}). \quad (10)$$

Using both real and virtual microphone signals as the inputs for noise reduction based on the subspace method, high-performance noise reduction is expected. What is important in the subspace method in the proposed method is that the eigenvector corresponding to the minimum eigenvalue is orthogonal to the signal subspace. Even if the interpolation by the virtual microphone technique is somewhat incorrect, orthogonality is normally satisfied, this method is expected to be effective.

D. Sound activity detection independent of the type of target source

We apply an SAD technique to the output signal of noise reduction based on the subspace method to detect abnormal sounds. To satisfy condition 3, this SAD technique must be applicable for various abnormal sounds. Therefore, we use a signal power as a feature value. Using this feature value, if a signal has power greater than a threshold, the signal is regarded as an abnormal sound. Thus, it is possible to detect all sounds in directions that are not suppressed by noise reduction as abnormal sounds.

In addition, as a technique to improve the VAD performance, the hang-over technique has been proposed [13], [14]. This method is based on the assumption that a voice continues for a long time. Employing this assumption, a non-voice period with a short duration detected between voice periods is regarded as a voice period. In this paper, we also assume that abnormal sounds will last for some time and we apply the hang-over technique to SAD. Moreover, we assume that an abnormal sound detected for a short period is a false detection, and such a sound is regarded as a normal sound.

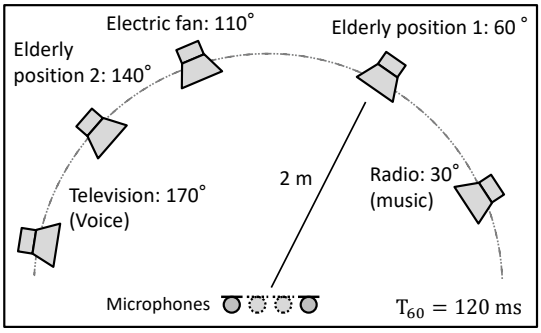


Fig. 2: Layout of the simulated room

III. Experiment on detecting abnormal sounds

In this paper, to evaluate the effectiveness of our proposed method, we conducted an experiment using observed signals that are convolutive mixtures of impulse responses simulated by the Room Impulse Response (RIR) generator [15]. We simulated an environment where an elderly person lives alone and applied the proposed method to detect the voice of the elderly person. This environment has three normal sounds, which are classical music from a radio, electric fan noise and a voice from a television (Fig. 2) with DOAs of 30°, 110° and 170°, respectively. As an abnormal sound, we simulated the voice of the elderly person at two positions, position 1 and 2 with DOAs of 60° and 140°, respectively. We used one speech as the voice of the elderly person and "Winter" from "The Four Seasons" by Vivaldi as the classical music and used the electric fan noise from the JEIDA Noise Database [16].

Performance evaluation was carried out by comparing the results of the proposed method and the following two methods: applying SAD to the unprocessed signal and to the output signal of a comparative noise reduction method. As the comparative method, we used a maximum signal-to-noise ratio (SNR) beamformer (maxSNRbf) [17], [18] with the virtual microphone technique [2]–[4]. This method requires the target-active period and target-inactive period as prior information for speech enhancement. In the abnormal sound detection, the target signal is an abnormal sound. Thus, we need the abnormal sound period as the target-active period and normal sounds period as target-inactive period. This experiment was performed for two directions of an abnormal sound.

For position 1 of the elderly person, we considered the case where the DOA of the abnormal sound is known, that is, the target-active period of the elderly person can be obtained as prior information. In this case, we can expect both the maxSNRbf and the proposed method to work well. For position

TABLE I: Experimental conditions

Number of real microphones	2
Number of virtual microphones	2 ($\alpha = 0.33, 0.67$)
Distance between real microphones	4 cm
Reverberation time	120 ms
Input SNR	-5 dB
Sampling rate	8 kHz
FFT frame length	512 samples
FFT frame shift	128 samples
Training period	10 s
Test period	20 s
The number of abnormal sound period	4 times
Length of abnormal sound period	8.1 seconds

2, we consider the case where the elderly person moves in a different direction from that for position 1. In this case, the maxSNRbf cannot obtain the target-active period because the target DOA is unknown. Thus, the maxSNRbf has no choice but to use the spatial filter produced for position 1, which decreases the performance of noise reduction. On the other hand, we can expect that our proposed method would work as well as it does for position 1 because it does not require the DOA of the abnormal sound.

A. Experimental conditions

The experimental conditions are shown in Table I. We used two real microphones and interpolated two virtual microphone signals at equal intervals. Thus, the microphone array we used was composed of four microphones, two real and two virtual microphones.

For the evaluation criteria, we used the false acceptance rate (FAR) and false rejection rate (FRR), which are denoted as

$$\mathrm{FAR} = \frac{N_{FA}}{N_n} \times 100\ [\%], \tag{11}$$

$$\mathrm{FRR} = \frac{N_{FR}}{N_s} \times 100\ [\%], \tag{12}$$

where N_{FA}, N_{FR}, N_s and N_n are the number of normal sound frames detected as abnormal sounds (false acceptances), the number of abnormal sound frames detected as normal sounds (false rejections), the total number of abnormal sound frames (signals) and the total number of normal sound frames (noise), respectively.

As the threshold used in SAD, we used the value satisfying $\mathrm{FAR} \simeq \mathrm{FRR}$ as an optimum value. In addition, we assumed that abnormal sounds continue for at least 500 ms and applied the hang-over technique. Also, when the period of the detected abnormal sound was shorter than 250 ms, we considered it to be a false detection and regarded it as a normal sound.

B. Results and discussion

Figure 3 shows the results of the experiment. In Figs. 3(a) and (b), the first column shows the signal and the second column shows the signal power. Also, the orange lines in the second column represent the SAD results and the rising parts of these lines are the periods detected as abnormal sounds. The first row of Figs. 3(a) and (b) shows true abnormal sounds, that is, the ground truth. The second row shows the unprocessed signal, the third row shows the results for the maxSNRbf and the last row shows the results for the proposed method. Additionally, the values of FAR and FRR and the thresholds are listed in Table II.

TABLE II: FAR, FRR and thresholds in the experiment

Elderly position 1	FAR [%]	FRR [%]	Threshold
Unprocessed	21.0	27.4	0.185
maxSNRbf	6.1	1.7	0.1
Proposed	6.7	7.0	0.06
Elderly position 2	**FAR [%]**	**FRR [%]**	**Threshold**
Unprocessed	17.6	25.5	0.175
maxSNRbf	47.0	28.9	0.19
Proposed	6.4	8.9	0.035

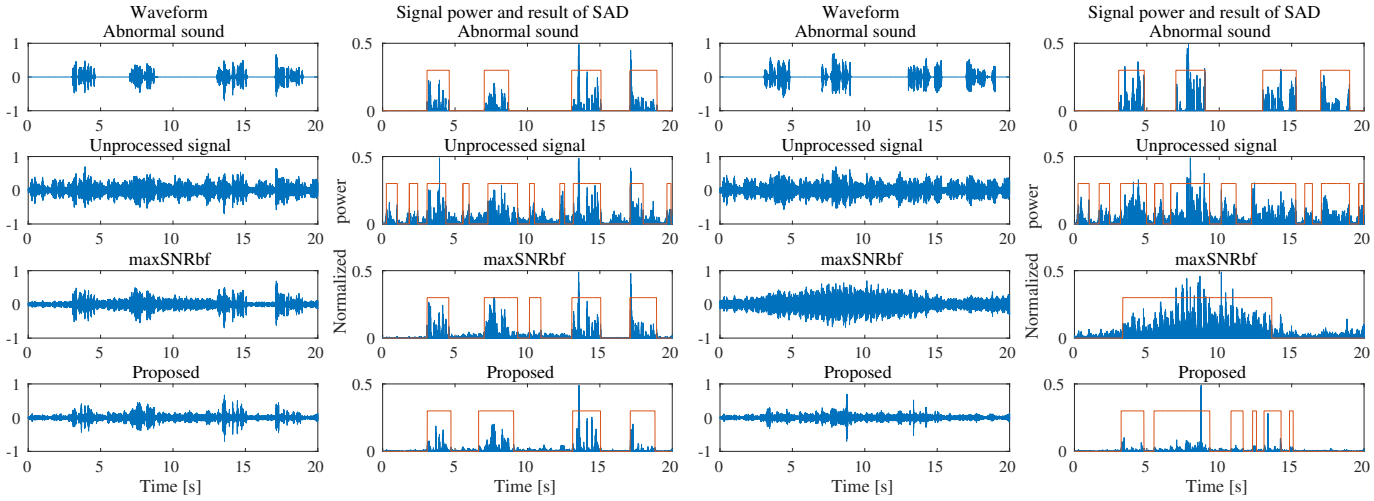
According to Fig. 3(a), both the maxSNRbf and the proposed method work well. The accuracy of the SAD using the maxSNRbf is very high; FAR is 6.1% and FRR is 1.7%, which are superior to those of the proposed method of 6.7% and 7.0%, respectively. Since the maxSNRbf uses prior information on the target, it can enhance the target sound more clearly than the proposed method, which does not use prior information. In contrast, the proposed method does not constrain the gain of the target sound, therefore less noise remains than in the output signal of the maxSNRbf. As a result, the SAD using the maxSNRbf better performance than the proposed method.

According to Fig. 3(b), the maxSNRbf failed in noise reduction because it used an incorrect spatial filter. In particular, both FAR and FRR are higher than those for the unprocessed signal. In contrast, the proposed method works well regardless of the target DOA. We confirmed that the proposed method is effective for detecting abnormal sounds under a normal sound environment.

IV. Conclusions

In this paper, we have proposed a new method of microphone array signal processing for detecting abnormal sounds that is applicable to monitoring elderly people at home and can be implemented with two microphones. This method consists of two parts: noise reduction based on a subspace method with virtual microphones and SAD using the signal power. This noise reduction method does not require the DOA of the abnormal sounds. However, it cannot suppress normal sounds effectively in an underdetermined condition. To resolve this issue, we used our previously proposed virtual microphone technique. By applying SAD after noise reduction, our proposed method was shown to be robust to noise and the target DOA. To evaluate this technique, we conducted an experiment simulating a system for monitoring an elderly person living alone.

As a result, we confirmed that our proposed method is effective for detecting abnormal sounds. Moreover, the result showed the robustness of our proposed method to the target DOA in noisy environments.



(a) The DOA of the abnormal sound is 60°

(b) The DOA of the abnormal sound is 140°

Fig. 3: Input signal at left microphone and the result of SAD for unprocessed signal and output signals of maximum SNR beamformer and proposed method.

ACKNOWLEDGEMENT

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

REFERENCES

[1] M. Kawamoto, F. Asano, K. Kurumatani, and Y. Hua, "A system for detecting unusual sounds from sound environment observed by microphone arrays," *Proc. ICIAS*, pp. 729–732, 2009.

[2] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Virtually increasing microphone array elements by interpolation in complex-logarithmic domain," *Proc. EUSIPCO*, pp. 1–5, Sept. 2013.

[3] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Generalized amplitude interpolation by β-divergence for virtual microphone array," *Proc. IWAENC*, pp. 150–154, Sept. 2014.

[4] 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.

[5] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 5, pp. 497–507, Sept. 2000.

[6] H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Speech enhancement using nonlinear microphone array based on complementary beamforming," *IEICE Trans. on Fundamentals*, vol. E82-A(8), pp. 1501–1510, 1999.

[7] S. Miyabe, B. H. (Fred) Juang, H. Saruwatari, and K. Shikano, "Analytical solution of nonlinear microphone array based on complementary beamforming," *Proc. IWAENC*, pp. 1–4, 2008.

[8] Y. Hioka and T. Betlehem, "Under-determined source separation based on power spectral density estimated using cylindrical mode beamforming," *Proc. WASPAA*, pp. 1–4, 2013.

[9] P. Chevalier, A. Ferréol, and L. Albera, "High-resolution direction finding from higher order statistics: The 2q-MUSIC algorithm," *IEEE Trans. on Signal Processing*, vol. 53, no. 4, pp. 2986–2997, 2006.

[10] Y. Sugimoto, S. Miyabe, T. Yamada, S. Makino, and B. H. (Fred) Juang, "Employing moments of multiple high orders for high-resolution under-determined DOA estimation based on music," *Proc. WASPAA*, pp. 1–4, 2013.

[11] O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, pp. 1830–1847, 2004.

[12] A. Jourjine, S. Rickard, and O. Yılmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," *Proc. ICASSP*, pp. 2985–2988, 2000.

[13] ETSI ES 202 050 v. 1.1.4, "Speech processing, transmission and quality aspects (STQ), advanced distributed speech recognition (ADSR); front-end feature extraction algorithm; compression algorithms," Nov. 2006.

[14] D. Vlaj, M. Kos, and Z. Kačič, "Quick and efficient definition of hangbefore and hangover criteria for voice activity detection," *Proc. IWSSIP*, 2016.

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

[16] "JEIDA Noise Database," Available at: http://research.nii.ac.jp/src/en/JEIDA-NOISE.html (in Japanese).

[17] H. L. Van Trees, *Optinum Array Processing*, John Wiley & Sons, 2002.

[18] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," *Proc. ICASSP*, vol. I, pp. 41–45, 2007.