EUSIPCO 2013 1569744343

VIRTUALLY INCREASING MICROPHONE ARRAY ELEMENTS BY INTERPOLATION IN COMPLEX-LOGARITHMIC DOMAIN

†*Hiroki Katahira,*‡*Nobutaka Ono,* †*Shigeki Miyabe,* †*Takeshi Yamada,* †*Shoji Makino*

†University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
‡National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda, Tokyo, 101-8430 Japan
katahira@mmlab.cs.tsukuba.ac.jp, onono@nii.ac.jp,
{miyabe, maki}@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp

ABSTRACT

In this paper, we propose a new array signal processing technique for an underdetermined condition by increasing the number of observation channels. We introduce virtual observation as an estimate of the observed signals at positions where real microphones are not placed. Such signals at virtual observation channels are generated by the complex logarithmic interpolation of real observed signals. With the increased number of observation channels, conventional linear array signal processing methods can be applied to underdetermined conditions. As an example of the proposed array signal processing framework, we show experimental results of speech enhancement obtained with maximum SNR beamformers modified using the virtual observation.

Index Terms— Linear array signal processing, underdetermined, speech enhancement, virtual observation, beamformer

1. INTRODUCTION

Microphone array signal processing is a speech processing framework based on signals observed with multiple sensors. Its typical applications are blind source separation (BSS), source localization and speech enhancement. Among various classes of array signal processing, the most established and powerful approach is linear array signal processing, which includes Independent Component Analysis (ICA) [1] for BSS and MUltiple SIgnal Classification (MUSIC) [2] for direction of arrival (DOA) estimation. However, many of the linear array processing methods are effective when the number of microphones equals or exceeds the number of sound sources, and the performance is often greatly degraded with a larger number of sound sources. The problem with these methods is that we need a large-scale microphone array to deal with a large number of sources, or we cannot apply these methods to signals observed with few sensors.

The most typical way of solving the underdetermined array signal processing problem is to cluster the observed signals assuming sparseness among sources [3, 4, 5]. While these methods are based on an idea that is slightly different from linear array signal processing, the straightforward way to realize the underdetermined extension of the linear array signal processing is to increase the number of observation channels by using higher dimensional nonlinear maps. The conventional way of increasing the number of observation channels is to use higher dimensional maps to reproduce higher order statistics, such as higher-order cumulants [6, 7, 8] and higher-order moments [9, 10]. Since higher-order statistics are useful for the classification problem, the higher-dimensional maps are used to improve the DOA estimation performance by MUSIC [6, 8, 10]. However, these extensions based on higher dimensional maps for the analysis of higher-order statistics suffer from distortion in their output signals when these extensions are applied to speech enhancement or BSS because these maps greatly change the nature of the signals.

In this paper, we propose an extension of array signal processing by introducing virtual observation which causes less distortion and residual noise in the output signal of processing. A virtual observation is obtained as an estimation of the signals at a point where no sensor is placed. We apply the pseudo increase of the observation channels with the virtual observation to the maximum SNR beamformer [11, 12], and confirm the effectiveness of the proposed virtual observation.

2. SPEECH ENHANCEMENT WITH LINEAR ARRAY PROCESSING

2.1. Speech enhancement by linear filter in STFT domain

Throughout this paper, all the processing is conducted in the STFT-domain. Let $s_i(\omega, t)$ be the i-th source signal at an angular frequency ω in the t-th frame, and let $x_j(\omega, t)$ be the observed signal at the j-th microphone. Signals can be modeled as

$$\begin{aligned} \mathbf{x}(\omega, t) &= \left[x_1(\omega, t), \cdots, x_M(\omega, t)\right]^{\mathrm{T}} \\ &\approx \sum_{i=1}^{N} \mathbf{a}_i(\omega)\, s_i(\omega, t), \qquad (1) \\ \mathbf{a}_i(\omega) &= \left[a_{1,i}(\omega), \cdots, a_{M,i}(\omega)\right]^{\mathrm{T}}, \qquad (2) \end{aligned}$$

where $a_{j,i}(\omega)$ is the transfer function from the i-th source to the j-th microphone and $\{\cdot\}^{\mathrm{T}}$ stands for the transposition of

a matrix. Speech enhancement by beamforming is conducted by constructing a multi-channel filter given by

$$\mathbf{w}(\omega) = \left[w_1(\omega), \cdots, w_M(\omega)\right]^{\mathrm{T}}, \tag{3}$$

to reduce the contamination of non-target sources, where $w_n^*(\omega)$ is a filter for the n-th channel and $\{\cdot\}^*$ denotes a complex conjugation. The enhanced signals $y(\omega, t)$ are given as

$$y(\omega, t) = \mathbf{w}^{\mathrm{H}}(\omega)\,\mathbf{x}(\omega, t), \tag{4}$$

where $\{\cdot\}^{\mathrm{H}}$ stands for the conjugate transposition of a matrix.

2.2. Performance degradation in underdetermined case

The ideal speech enhancement is obtained when $\mathbf{w}(\omega)$ is orthogonal to all the transfer functions $\mathbf{a}_i(\omega)$ of the non-target sources as

$$\mathbf{w}^{\mathrm{H}}(\omega)\,\mathbf{a}_i(\omega) = 0, \ \forall i \neq i_{\mathrm{T}}, \tag{5}$$

$$\begin{aligned} y(\omega, t) &= \mathbf{w}^{\mathrm{H}}(\omega) \sum_{i=1}^{N} \mathbf{a}_i(\omega)\, s_i(\omega, t) \\ &= \mathbf{w}^{\mathrm{H}}(\omega)\,\mathbf{a}_{i_{\mathrm{T}}}(\omega)\, s_{i_{\mathrm{T}}}(\omega, t), \end{aligned} \tag{6}$$

where the i_{T}-th source is the target. However, such a filter $\mathbf{w}(\omega)$ does not generally exist with linearly independent transfer function vectors when the dimension M of the observed signal vectors is smaller than the number N of source signals. Therefore, the performance of the beamformer degrades in an underdetermined condition.

3. VIRTUAL OBSERVATION DERIVED FROM COMPLEX LOGARITHMIC INTERPOLATION

3.1. Virtual observation with interpolation in nonlinear domain

In this section, we formulate our proposed virtual observation approach. We generate the virtual observation as the estimates of signals observed by a virtual microphone placed at the point without a real microphone is placed. A virtually-observed signal $v(\omega, t, \alpha)$ is generated as the estimated observation obtained with a virtual microphone placed at the $\alpha : (1-\alpha)$ internally dividing point of the positions of two real microphones (Fig. 1) as an interpolation in the domain of a function f;

$$v(\omega, t, \alpha) = f^{-1}\left((1-\alpha)\, f(x_1(\omega, t)) + \alpha f(x_2(\omega, t))\right), \tag{7}$$

where $x_1(\omega, t)$ and $x_2(\omega, t)$ denote the complex amplitudes of the actually-observed signals of two microphones. The choice of the nonlinear function f is the most important factor in the proposed method for generating a virtual observation as a good estimate.

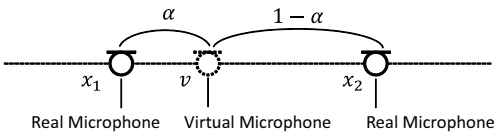

Fig. 1. Arrangement of real and virtual microphones.

3.2. Derivation of appropriate interpolation domain based on plane-wave model

We derive the function f where the observation is interpolated to generate the virtual observation from the plane wave model, which is the simplest model of wave propagation. In this model, the time difference of arrivals is in a linear relationship with the inter-microphone distance for any direction of the source. Thus in the STFT domain, the phase $\angle v(\omega, t, \alpha)$ of the virtually-observed signal should satisfy the following condition.

$$\angle v(\omega, t, \alpha) = (1-\alpha)\angle x_1(\omega, t) + \alpha \angle x_2(\omega, t), \tag{8}$$

where $\angle\{\cdot\}$ denotes the phase angle. Note that here we assume that the distance between the real microphones is smaller than half of the wavelength, and the following condition is satisfied without spatial aliasing;

$$|\angle x_1(\omega, t) - \angle x_2(\omega, t)| \leq \pi. \tag{9}$$

The phase of the complex number is expressed as the imaginary part of the complex logarithmic function as follows

$$\mathrm{Log}x = \log|x| + j\angle x, \tag{10}$$

where Log stands for the complex logarithmic function. Thus the linear phase interpolation is operated in Eq. (7) using the following logarithmic function;

$$f(x) = \mathrm{Log}x, \tag{11}$$

$$f^{-1}(x) = \exp(x), \tag{12}$$

and the virtually-observed signal $v(\omega, t, \alpha)$ is estimated as follows

$$\begin{aligned} v(\omega, t, \alpha) &= \exp\left((1-\alpha)\,\mathrm{Log}\left(x_1(\omega, t)\right) + \alpha \mathrm{Log}\left(x_2(\omega, t)\right)\right) \\ &= \exp\left((1-\alpha)\log|x_1(\omega, t)| + \alpha \log|x_2(\omega, t)|\right. \\ &\quad \left. + j\left((1-\alpha)\angle x_1(\omega, t) + \alpha \angle x_2(\omega, t)\right)\right). \end{aligned} \tag{13}$$

Note that the interpolation of the amplitude in the logarithmic domain is valid when a single plane wave arrives because plane waves do not decay with distance, and $|x_1(\omega, t)| = |x_2(\omega, t)|$.

We discuss the propagation of waves other than the plane wave. With a single spherical wave, the interpolation depends on the source direction and it is difficult to interpolate the complex amplitude only with information about the actually-observed signals $x_1(\omega, t)$ and $x_2(\omega, t)$. However, the spherical wave can be approximated as a plane wave when the microphones are closely spaced and the inter-microphone

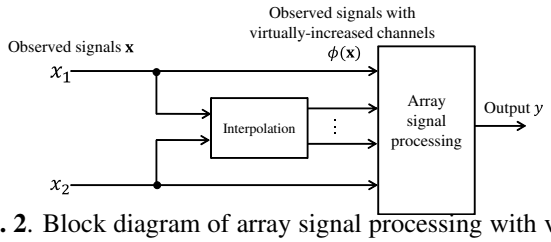

Fig. 2. Block diagram of array signal processing with virtual observation.

distance is much smaller than the distances between the source and the microphones. When multiple waves are propagating, it is also difficult to estimate the virtual observation with $x_1(\omega, t)$ and $x_2(\omega, t)$ because the observation is the sum of multiple wave fronts. However, by assuming that the source signals are sparse enough and each time-frequency slot is dominated by a single source, the virtual observation is effectively estimated by using the proposed interpolation.

We can estimate an arbitrary number of virtual observation channels. Assume we have N_{v} virtual microphones between two real microphones at regular intervals. We denote the observation obtained by these N_{v} virtual microphones as

$$v_n(\omega, t) = v\left(\omega, t, \frac{n}{N_{\mathrm{v}}+1}\right). \tag{14}$$

Let $\phi(\mathbf{x})$ be the observed signal vector composed of actual and virtual observations:

$$\phi(\mathbf{x}(\omega, t)) = \left[x_1(\omega, t), \quad v_1(\omega, t), \quad \cdots, \quad v_{N_{\mathrm{v}}}(\omega, t), \quad x_2(\omega, t)\right]^{\mathrm{T}}. \tag{15}$$

Since the proposed method brings no other modification than adding the virtual observation in the signal processing as Fig. 2, it has a potential to be applied to other kinds of array signal processing techniques such as source localization, blind source separation, etc. We will investigate them in future work.

4. APPLICATION OF VIRTUAL OBSERVATION APPROACH TO MAXIMUM SNR BEAMFORMER

4.1. Maximum SNR beamformer

In this section, we apply the proposed virtual observation technique to a maximum SNR beamformer. The modification is realized simply by substituting the actual observation vector $\mathbf{x}(\omega, t)$ with the observation vector $\phi(\mathbf{x}(\omega, t))$ of the actual and virtual channels. We construct a multi-channel filter $\widehat{\mathbf{w}}(\omega)$ for actually and virtually observed signals given by

$$\widehat{\mathbf{w}}(\omega) = \left[\widehat{w}_1(\omega), \cdots, \widehat{w}_{N_{\mathrm{v}}+2}(\omega)\right]^{\mathrm{T}}. \tag{16}$$

The filter $\widehat{\mathbf{w}}(\omega)$ is designed to maximize the power ratio $\lambda(\omega)$ between the target-only period Θ_{T}, and the interference-only period Θ_{I}:

$$\lambda(\omega) = \frac{\widehat{\mathbf{w}}^{\mathrm{H}}(\omega)\widehat{\mathbf{R}}_{\mathrm{T}}(\omega)\widehat{\mathbf{w}}(\omega)}{\widehat{\mathbf{w}}^{\mathrm{H}}(\omega)\widehat{\mathbf{R}}_{\mathrm{I}}(\omega)\widehat{\mathbf{w}}(\omega)}, \tag{17}$$

where $\widehat{\mathbf{R}}_{\mathrm{T}}(\omega)$ and $\widehat{\mathbf{R}}_{\mathrm{I}}(\omega)$ are the covariance matrices of the target-only period Θ_{T} and the interference-only period Θ_{I}. The covariance matrices are estimated with both the actually and virtually observed signals as,

$$\widehat{\mathbf{R}}_{\mathrm{T}}(\omega) = \frac{1}{|\Theta_{\mathrm{T}}|}\sum_{t\in\Theta_{\mathrm{T}}}\phi(\mathbf{x}_{\mathrm{T}}(\omega, t))\phi(\mathbf{x}_{\mathrm{T}}(\omega, t))^{\mathrm{H}}, \tag{18}$$

$$\widehat{\mathbf{R}}_{\mathrm{I}}(\omega) = \frac{1}{|\Theta_{\mathrm{I}}|}\sum_{t\in\Theta_{\mathrm{I}}}\phi(\mathbf{x}_{\mathrm{I}}(\omega, t))\phi(\mathbf{x}_{\mathrm{I}}(\omega, t))^{\mathrm{H}}. \tag{19}$$

These covariance matrices include the information of higher order statistics because virtually-observed signals are obtained by non-linear mapping of observed signals. The filter $\widehat{\mathbf{w}}(\omega)$ used to maximize power ratio $\lambda(\omega)$ is given as an eigenvector corresponding to the maximum eigenvalue of the following generalized eigenvalue problem;

$$\widehat{\mathbf{R}}_{\mathrm{T}}(\omega)\widehat{\mathbf{w}}(\omega) = \lambda(\omega)\widehat{\mathbf{R}}_{\mathrm{I}}(\omega)\widehat{\mathbf{w}}(\omega). \tag{20}$$

Since the maximum SNR beamformer $\widehat{\mathbf{w}}(\omega)$ has a scaling ambiguity [11], we revise the beamformer as:

$$\widehat{\mathbf{w}}(\omega) \leftarrow b_k(\omega)\widehat{\mathbf{w}}(\omega), \tag{21}$$

where $b_k(\omega)$ is the k-th component of $\mathbf{b}(\omega)$ given by

$$\mathbf{b}(\omega) = \frac{\widehat{\mathbf{R}}_{\mathbf{x}}(\omega)\widehat{\mathbf{w}}(\omega)}{\widehat{\mathbf{w}}^{\mathrm{H}}(\omega)\widehat{\mathbf{R}}_{\mathbf{x}}(\omega)\widehat{\mathbf{w}}(\omega)}, \tag{22}$$

$$\widehat{\mathbf{R}}_{\mathbf{x}}(\omega) = \frac{1}{T}\sum_{t=1}^{T}\phi(\mathbf{x}(\omega, t))\,\phi(\mathbf{x}(\omega, t))^{\mathrm{H}}. \tag{23}$$

Then enhanced signal $\widehat{y}(\omega, t)$ is obtained as

$$\widehat{y}(\omega, t) = \widehat{\mathbf{w}}^{\mathrm{H}}(\omega)\,\phi(\mathbf{x}(\omega, t)). \tag{24}$$

4.2. Regularization of covariance matrix

The generalized eigenvalue problem of Eq. (20) tends to generate errors caused by the rank deficiency of covariance matrix $\widehat{\mathbf{R}}_{\mathrm{I}}(\omega)$ when the number N_{v} of virtual observation channels is increased. Thus we revise the covariance matrix by adding a regularizer

$$\widehat{\mathbf{R}}_{\mathrm{I}}(\omega) \leftarrow \widehat{\mathbf{R}}_{\mathrm{I}}(\omega) + \varepsilon\mathbf{E}, \tag{25}$$

where $\mathbf{E}$ is the unit matrix and ε is an imperceptible non-negative number.

5. EXPERIMENTS

To confirm the effectiveness of the proposed virtual observation, we conducted speech enhancement experiments and evaluated the performance.

Table 1. Experimental conditions

Number of real microphones	2
Number of sources	3
Number N_{v} of virtual observation channels	0–9
Source directions	$-40°, 20°, 50°$
Real microphone spacing	2.15 cm
Reverberation time	130 ms
Sampling rate	8 kHz
Frame length	1024 samples
Frame shift	256 samples
Signal length	20 sec
Target only period Θ_{T} length	5 sec
Interference only period Θ_{I} length	5 sec
Coefficient ε of regularizer	$10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}, 10^{-14}, 0$

5.1. Experimental conditions

We used the 20 sec samples of Japanese and English speech for the sources. Observed signals were formed by the convolution of measured impulse responses and speech signals. We used three sound sources and two microphones, which was an underdetermined condition, then it is difficult to achieve good speech enhancement with conventional linear microphone array technique.

The other experimental conditions are shown in Table 1. To evaluate the performance of the beamformers, we used objective evaluation criteria, namely the Signal-to-Distortion Ratio (SDR) and Signal-to-Interference Ratio (SIR) [13]. Higher SDR and SIR values mean better speech enhancement performance.

5.2. Experimental result and consideration

Figure 3 shows spectrograms of the source signal and the observed signal, and Fig. 4 shows spectrograms of enhanced signals. The spectrograms show that the signal is better enhanced with the virtual observation than without it. Figure 5 shows the SDR and SIR when there were of 0 to 9 virtual observation channels. Note that the zero in the number of virtual signals in the figure indicates that the beamformer was processed with only the actually-observed signals. The SDR and SIR values are greatly improved when one virtual observation channel is introduced. We consider this to be because the underdetermined condition is resolved when one virtual observation channel. Although the SDR value decreases as the number of virtual observation channels increases when there are no regularization of covariance matrices, it does not decrease when there is a regularizer with a larger coefficient. Thus we can say that the distortion caused by the rank deficiency of non-target covariance matrices has been reduced. However, we can observe higher SIR values in when ε is set at a low value. It is considered that regularization caused errors with the beamformers and degraded there ability to reduce interference. We have also compared speech enhancement performance of the virtual microphone array (two real microphones and one virtual microphone) with the real microphone array (three real microphones) in the same layout shown in Fig. 6. The comparative result in Fig. 7 shows that the virtual observation improves the SIR close to that of real microphone array. Our proposed virtual observation approach has improved both SDR and SIR, especially the latter. Thus we have confirmed the effectiveness of our proposed virtual observation.

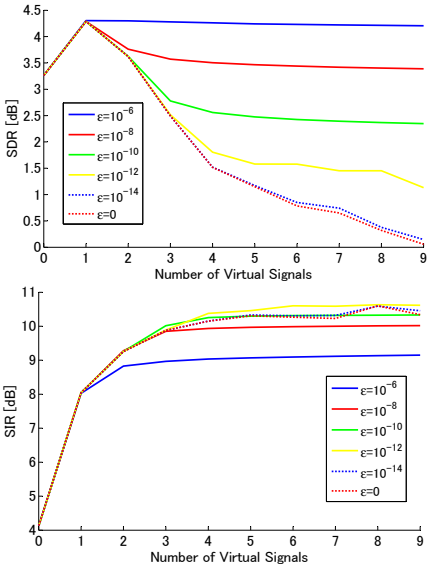

Fig. 5. Evaluation values for maximum SNR beamformers with virtual observation

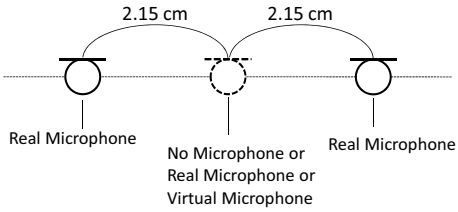

Fig. 6. Microphone layout for comparing with overdetermined case

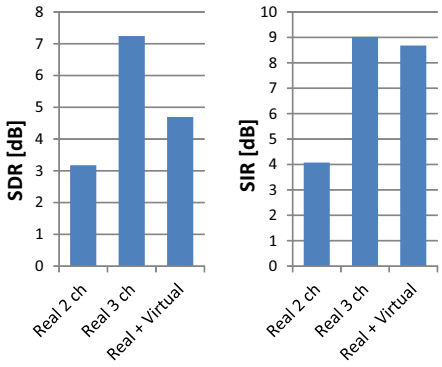

Fig. 7. Evaluation values for maximum SNR beamformers with real and virtual microphone arrays

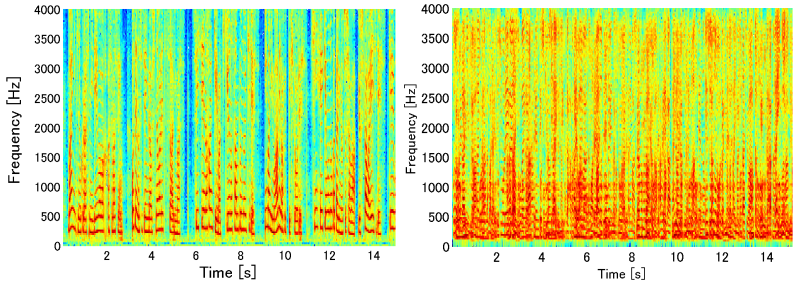

Fig. 3. Spectrograms of source signal (left), observed signal (right)

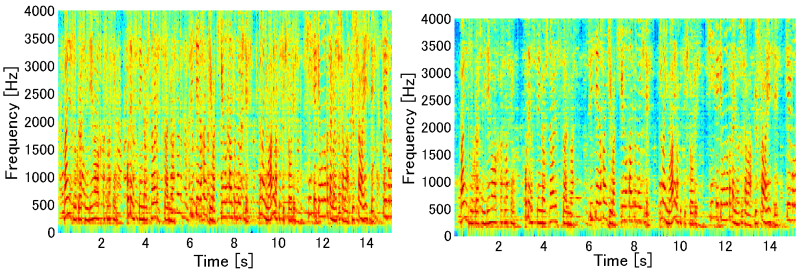

Fig. 4. Spectrograms of enhanced signal: without virtual observation (left), with virtual observation (right)

6. CONCLUSION

In this paper, we proposed a new array signal processing technique for an underdetermined condition. We introduced a virtual observation as an estimate of the observed signals at positions lacking real microphones are not placed. The virtually-observed signals were generated by interpolation in the complex logarithmic domain. We applied our proposed virtual observation approach to a maximum SNR beamformer and that were not equipped with real microphones. The performance improved in the underdetermined condition, and we confirmed the effectiveness of our proposed virtual observation technique.

7. REFERENCES

[1] S. Makino *et al.*, *Blind Speech Separation*, Springer, 2007.

[2] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans.*, vol. 34, no. 3, 1986.

[3] O. Yilmaz, S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] Y. Izumi *et al.*, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," *WASPAA*, pp. 147–150, 2007.

[5] H. Sawada *et al.*, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[6] B. Porat, B. Friedlander, "Direction finding algorithms based on higher order statistics," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 2016–2024, 1991.

[7] P. Chevalier *et al.*, "On the virtual array concept for higher order array processing," *IEEE Trans. Signal Processing*, vol. 53, no. 4, pp. 1254–1271, 2005.

[8] P. Chevalier *et al.*, "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.

[9] S. Miyabe *et al.*, "Kernel-based nonlinear independent component analysis for underdetermined blind source separation," *Proc. ICASSP*, pp. 1641–1644, 2008.

[10] Y. Sugimoto *et al.*, "Underdetermined DOA estimation by the non-linear music exploiting higher-order moments," *Proc IWAENC2012*, 2012.

[11] S. Araki *et al.*, "Blind speech separation in a meeting situation with maximum SNR beamformers," *Proc. ICASSP*, vol. I, pp. 41–45, 2007.

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

[13] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.