14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

BLIND SOURCE SEPARATION OF ANECHOIC MIXTURES IN TIME DOMAIN UTILIZING AGGREGATED MICROPHONES

Mitsuharu Matsumoto and Shuji Hashimoto

Department of Applied Physics, Waseda University
55N-4F-10A, 3-4-1 Okubo, Shinjuku-ku, 169-8555, Tokyo, Japan
phone: +81-3-5286-3233 , fax: +81-3-3202-7523 , email: matsu@shalab.phys.waseda.ac.jp, shuji@waseda.jp
web: www.shalab.phys.waseda.ac.jp

ABSTRACT

This paper introduces a blind source separation (BSS) algorithm in the time domain based on the amplitude gain difference of two directional microphones located at the same place, namely aggregated microphones. A feature of our approach is to treat the BSS problem of the anechoic mixtures in the time domain. Sparseness approach is one of the attractive methods to solve the problem of the sound separation. If the signal is sparse in the frequency domain, the sources rarely overlap. Under this condition, it is possible to extract each signal using time-frequency binary masks. In this paper, we treat the non-stationary, partially disjoint signals. In other words, most of the signals overlap in the time domain and the frequency domain though there exist some intervals where the sound is disjoint. We firstly show the source separation problem can be described not as a convolutive model but as an instantaneous model in spite of the anechoic mixing when the aggregated microphones are assumed. We then show the necessary conditions and show the algorithm with the experimental results. In this method, we can treat the problem not in the time-frequency domain but in the time domain due to the characteristics of the aggregated microphones. In other words, we can consider the problem not in the complex space but in the real space. The mixing matrix can be directly identified utilizing the observed signals without estimating the intervals where the signal is disjoint through all the processes.

1. INTRODUCTION

Blind source separation (BSS) has been studied for a long time. The earliest approach is to separate an instantaneous linear even-determined mixture of non-Gaussian independent sources utilizing the recurrent neural network [1]. Independent Component Analysis (ICA) is one of the useful blind source separation algorithm [2, 3]. It can separate the mixed signals if the signals are statistically independent. However, most of these approaches can separate only stationary non-Gaussian signals. There has been reported a BSS algorithm for the non-stationary signals utilizing the higher order statistics of the mixed sounds [4]. However, the computational cost is large due to calculating the higher order statistics.

Some approaches utilize sparseness to solve the problem of the sound separation [5, 6, 7]. Sparseness means that most of the frequency components of a signal are zero. If the signal is sparse, the sources rarely overlap [8, 9]. Under this assumption, it is possible to extract each signal using time-frequency binary masks. However, due to the binary masks, these methods result in too much zero-padding to the separated sounds, and so the separated sounds are severely distorted. These methods require that the frequency components of the mixed sounds rarely overlap in any time.

In this paper, we consider the problem of separating the non-stationary, partially disjoint signals and propose a BSS algorithm utilizing the amplitude of two microphones located at the same place, namely aggregated microphones. Partially disjoint means the signals overlapping in most of the time domain and the frequency domain, while there exist some intervals where the sound is disjoint. In our method, the mixing matrix is identified utilizing the interval where the sound is disjoint by utilizing the characteristics of the aggregated microphones. The aggregated microphones method is another type of microphone array not utilizing the differences in the position of the microphones, but utilizing the differences in the directivity of the microphones. The conventional microphone array, namely the phased microphone array, realizes the delay-sum type microphone array [10], the adaptive microphone array [11, 12] and DOA estimation such as high resolution algorithms [13, 14] by utilizing the phase difference of each microphone. However, it is difficult to miniaturize the microphone array due to utilize the phase difference of each microphone. In the aggregated microphones, all the microphones are located at the same position, and the directional microphones are arranged to differentiate the directivity of the microphones [15, 16, 17]. Hence it is easy to miniaturize the system. This feature is useful when applied to the small robots, the conference systems, and so on.

In [18], the source separation problem is categorized as the instantaneous, anechoic and echoic mixings. In this paper, we discuss the anechoic mixing. In the next section, we firstly describe the difference of the phased microphone array and the aggregated microphones and formulate the problem. We also show the source separation problem can be described not as the convolutive model but as the instantaneous model in the case of the anechoic mixing when the aggregated microphones are assumed. In Sec.3, we explain the procedure of the proposed method. This method can treat the problem not in the time-frequency domain but in the time domain due to the characteristics of the aggregated microphones. In other words, we can consider the problem not in the complex space but in the real space. The mixing matrix can directly be identified utilizing the observed signals without estimating the interval where the signal is disjoint. In Sec.4, we show the experimental results of the sound separation.

This research was supported (in part) by the Grant-in-Aid for the WABOT-HOUSE Project by Gifu Prefecture and the 21st Century Center of Excellence Program, "The innovative research on symbiosis technologies for human and robots in the elderly dominated society", Waseda University.

2. PROBLEM FORMULATION

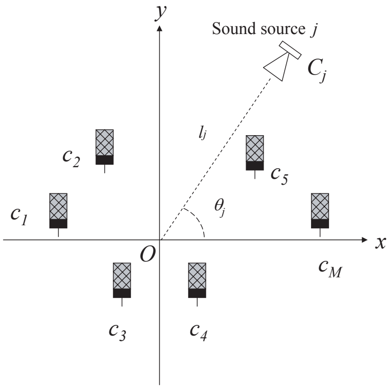


Figure 1: Basic setup of the phased microphone array

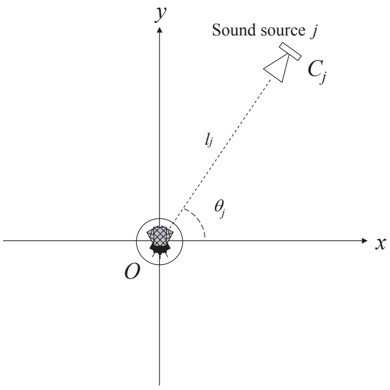


Figure 2: Basic setup of the aggregated microphones

Let us consider N sounds recorded by M microphones in the anechoic room. We firstly describe the difference of the phased microphone array and the aggregated microphones. Figs.1 and 2 illustrate the basic setup of the phased microphone array and aggregated microphones, respectively. In Figs.1 and 2, O represents the origin. c_i and C_j represent the position of the i-th microphone and the j-th sound source, respectively. θ_j and l_j represent the angle and the distance from the sound source j to the origin, respectively. In the phased microphone array, all the microphones are located at the different positions to differentiate the phase. Omni-directional microphones are usually utilized in order to make the amplitude identical. Hence, the received signal of the i-th microphone $x_i(t)$ can be represented as follows:

$$x_i(t) = \sum_{j=1}^{N} s_j(t - \tau_{ij}) \ (i = 1,2,...,M) \tag{1}$$

where $s_j(t)$ represents the sound j at the origin. τ_{ij} is the time delay from the origin to the i-th microphone regarding the sound j. In the phased microphone array, the delay τ_{ij} needs to be considered in the anechoic room. Various applications such as the sound separation, DOA estimation and so on, are realized utilizing the difference of the time delay τ_{ij}. On the other hands, in the aggregated microphones, directional microphones are utilized to differentiate the amplitude gain. All the microphones are located at the same positions to make the phase identical. Suppose all the microphones are located at the origin as shown in Fig.2. The received signal $x_i(t)$ can be represented as the follows:

$$x_i(t) = \sum_{j=1}^{N} d_{ij} s_j(t) \ (i = 1,2,...,M) \tag{2}$$

where d_{ij} represents the amplitude gain of the i-th microphone regarding the sound j. In the aggregated microphones, we can describe the blind decomposition problem of anechoic mixing as the instantaneous mixtures unlike the phased microphone array. In the aggregated microphones, various applications are realized utilizing the difference of the amplitude gain d_{ij}. The phased microphone array requires the information of the positions regarding all the microphones in advance because we utilize the time delay to separate the sounds. In a similar way, the aggregated microphones method requires the information of the amplitude gain d_{ij} regarding all the microphones in advance. One of our research targets is to simplify this process, that is, to separate the sounds without utilizing the knowledge of the amplitude gain d_{ij}.

We consider a two-input, two-output sound separation problem utilizing the aggregated microphones, that is, $N = M = 2$. The received signal $x_1(t)$ and $x_2(t)$ can be described as follows:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix} \tag{3}$$

To simplify the explanation, we rewrite $x_1(t)$ and $x_2(t)$ as follows:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ D_1 & D_2 \end{bmatrix} \begin{bmatrix} s'_1(t) \\ s'_2(t) \end{bmatrix} \tag{4}$$

where

$$s'_1(t) = d_{11} s_1(t) \tag{5}$$

$$s'_2(t) = d_{12} s_2(t) \tag{6}$$

D_j represents the amplitude gain ratio of two microphones as follows:

$$D_j = \frac{d_{2j}}{d_{1j}} \tag{7}$$

As d_{11} and d_{12} are constant, $s_i(t)$ and $s'_i(t)$ is homothetic in the time domain. In this paper, we regard the sound separation problem as obtaining $s'_1(t)$ and $s'_2(t)$ utilizing $x_1(t)$ and $x_2(t)$. We assume $D_1 > D_2$ for illustrative purposes.

3. TIME DOMAIN BSS METHOD

3.1 Assumption regarding the sounds

We assume the following conditions regarding the sound sources.

- **Assumption 1.** The signal is partially disjoint.

Define $\Phi_{11}(k,t)$ and $\Phi_{22}(k,t)$, the short-time auto-correlation function of $s_1(t)$ and $s_2(t)$ at time t as follows:

$$\Phi_{11}(k,t) = \frac{1}{L} \sum_{l=0}^{L-1} s_1(l+t) s_1(k+l+t) \tag{8}$$

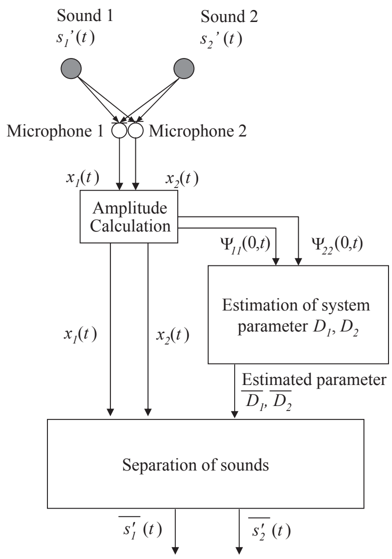


Figure 3: Block diagram of the proposed method

$$\Phi_{22}(k,t) = \frac{1}{L}\sum_{l=0}^{L-1} s_2(l+t)s_2(k+l+t) \tag{9}$$

where k and L represent the time deviation and the number of the data used for averages, respectively. This condition is described as follows:
$\exists t_\alpha$, $\exists t_\beta$ such that

$$\begin{cases} \Phi_{11}(0,t_\alpha) \neq 0 \\ \Phi_{22}(0,t_\alpha) = 0 \end{cases} \tag{10}$$

and

$$\begin{cases} \Phi_{11}(0,t_\beta) = 0 \\ \Phi_{22}(0,t_\beta) \neq 0 \end{cases} \tag{11}$$

where $\Phi_{11}(0,t)$ and $\Phi_{22}(0,t)$ represent the short-time mean power of $s_1(t)$ and $s_2(t)$ at time t, respectively. This assumption is similar to the one in the spectral subtraction method. The spectral subtraction requires that the sound is silent when the noise signal is estimated. In binary mask, it should be assumed that the frequency components of the mixed sounds hardly overlap in any time. On the other hands, in our method, we only assume that the sounds do not overlap in a short time L. The mixed sounds may overlap in most of the time domain. The frequency components of mixed sounds may also overlap in most of the frequency domain. It should be noted that we do not need to know t_α and t_β through all the procedure to separate the sounds.

- **Assumption 2.** The amplitude gain ratio of two directional microphones is different for two sounds.

This condition is represented as follows:

$$D_1 \neq D_2 \tag{12}$$

It should be noted that we do not need to know the directivity of the microphones in advance to separate the sounds. We can estimate the directivity of the microphones by utilizing the algorithm described in Sec.3.2. This assumption is the same condition as that of the sparseness approach utilizing aggregated microphones [19].

- **Assumption 3.** Two sounds are noncorrelated.

The condition is represented as follows:

$$\Phi_{12}(k,t) = \frac{1}{L}\sum_{l=0}^{L-1} s_1(l+t)s_2(k+l+t) = 0 \tag{13}$$

As is well known, if two sounds are statistically independent as the assumption in the independent component analysis, this assumption is also satisfied.

3.2 Sound separation algorithm

The processing steps of our method are as follows:

- **i)** Estimation of D_1 and D_2

To obtain D_1 and D_2, we firstly define $\Psi_{11}(k,t)$ and $\Psi_{22}(k,t)$ as the short-time auto-correlation function of the received sounds $x_1(t)$ and $x_2(t)$ at time t as follows:

$$\Psi_{11}(k,t) = \frac{1}{L}\sum_{l=0}^{L-1} x_1(l+t)x_1(k+l+t) \tag{14}$$

$$\Psi_{22}(k,t) = \frac{1}{L}\sum_{l=0}^{L-1} x_2(l+t)x_2(k+l+t) \tag{15}$$

According to the assumption 3, $\Psi_{11}(k,t)$ and $\Psi_{22}(k,t)$ can be expressed as follows:

$$\begin{aligned} \Psi_{11}(k,t) &= {d_{11}}^2\Phi_{11}(k,t) + {d_{12}}^2\Phi_{22}(k,t) \\ &= \Phi'_{11}(k,t) + \Phi'_{22}(k,t) \end{aligned} \tag{16}$$

$$\begin{aligned} \Psi_{22}(k,t) &= {d_{21}}^2\Phi_{11}(k,t) + {d_{22}}^2\Phi_{22}(k,t) \\ &= {D_1}^2\Phi'_{11}(k,t) + {D_2}^2\Phi'_{22}(k,t) \end{aligned} \tag{17}$$

where

$$\Phi'_{11}(k,t) = \frac{1}{L}\sum_{l=0}^{L-1} s'_1(l+t)s'_1(k+l+t) \tag{18}$$

$$\Phi'_{22}(k,t) = \frac{1}{L}\sum_{l=0}^{L-1} s'_2(l+t)s'_2(k+l+t) \tag{19}$$

We define the ratio of the inter-channel power difference $\Delta A(t)$ as follows:

$$\Delta A(t) = \frac{\Psi_{22}(0,t)}{\Psi_{11}(0,t)} \tag{20}$$

$\Delta A(t)$ can be represented as follows:

$$\begin{aligned} \Delta A(t) &= \frac{{D_1}^2\Phi'_{11}(0,t) + {D_2}^2\Phi'_{22}(0,t)}{\Phi'_{11}(0,t) + \Phi'_{22}(0,t)} \\ &= {D_1}^2 + \frac{({D_2}^2 - {D_1}^2)\Phi'_{22}(0,t)}{\Phi'_{11}(0,t) + \Phi'_{22}(0,t)} \\ &= {D_2}^2 + \frac{({D_1}^2 - {D_2}^2)\Phi'_{11}(0,t)}{\Phi'_{11}(0,t) + \Phi'_{22}(0,t)} \end{aligned} \tag{21}$$

As $\Phi'_{11}(0,t)$ and $\Phi'_{22}(0,t)$ are nonnegative, $\Delta A(t)$ can be expressed as follows under the assumption 1.

$$\Delta A(t) \begin{cases} < {D_1}^2 & (t \neq t_\alpha) \\ = {D_1}^2 & (t = t_\alpha) \end{cases} \tag{22}$$

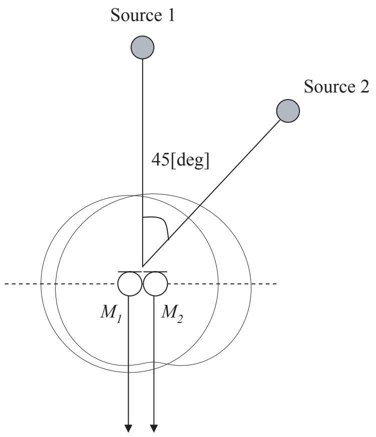


Figure 4: Arrangement for the experiment

$$\Delta A(t) \begin{cases} > {D_2}^2 & (t \neq t_\beta) \\ = {D_2}^2 & (t = t_\beta) \end{cases} \tag{23}$$

Note that we assumed $D_1 > D_2$ in Sec.2. Hence, D_1 and D_2 can be obtained as follows:

$$\overline{D_1} = \sqrt{\max[\Delta A(t)]} = \sqrt{\max[\frac{\Psi_{22}(0,t)}{\Psi_{11}(0,t)}]} \tag{24}$$

$$\overline{D_2} = \sqrt{\min[\Delta A(t)]} = \sqrt{\min[\frac{\Psi_{22}(0,t)}{\Psi_{x1}(0,t)}]} \tag{25}$$

where $\overline{D_1}$ and $\overline{D_2}$ represent the estimated value of D_1 and D_2, respectively. In this procedure, we do not need to estimate even the interval where the sound is disjoint in order to identify D_1 and D_2. What we have to do is only checking the value of $\Delta A(t)$ for a certain time, which includes the intervals where the signal is disjoint. This method works out due to the characteristics of the aggregated microphones. Because autocorrelation function has complete information about only the spectral amplitude and not the phase. Note that we do not apply this operation when $\Psi_{11}(0,t)$ and $\Psi_{22}(0,t)$ are 0.

- **ii)** Sound separation

We can reconstruct the signal $\overline{s_i'}(t)$ in the time domain as follows:

$$\overline{s_1'}(t) = \frac{x_2(t) - \overline{D_2}x_1(t)}{\overline{D_1} - \overline{D_2}} \tag{26}$$

$$\overline{s_2'}(t) = \frac{x_2(t) - \overline{D_1}x_1(t)}{\overline{D_2} - \overline{D_1}} \tag{27}$$

4. EXPERIMENT

To evaluate the proposed method, we conducted the sound separation experiments. The input signals were mixed to simulate the free-field situation as shown in Fig.4. As two microphones, we assume a pair of a omni-directional microphone as M_1 and a uni-directional microphone as M_2 located at the same place. It is known that the general characteristic of the unidirectional microphones can be described as follows:

$$d(\theta) = a + b\cos(\theta) \tag{28}$$

$$0 < b < a \tag{29}$$

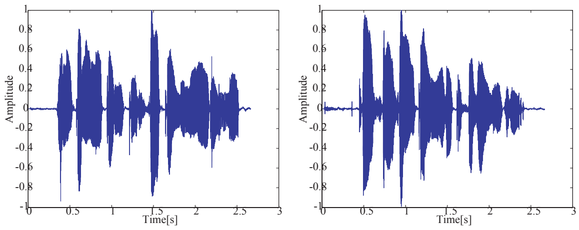
Table 1: The true values and estimated values of D_1 and D_2

True value	D_1	D_2
	1.25	0.89
Estimated value	$\overline{D_1}$	$\overline{D_2}$
Exp.1	1.1395	0.8171
Exp.2	1.2755	0.8005
Exp.3	1.2245	0.7763
Exp.4	1.2313	0.8764
Exp.5	1.3294	0.907
Exp.6	1.2479	0.8383
Exp.7	1.2428	0.7831
Exp.8	1.2324	0.7984
Exp.9	1.2559	0.9696
Exp.10	1.2558	0.7925

where θ represents the angle from the frontal direction of the microphone. $d(\theta)$ represents the amplitude gain of the unidirectional microphone for θ. We set a and b as 0.6 and 0.4, respectively. The values of a and b are based on the study of Okada.et.al [20]. The amplitude gain of the omni-directional microphone was set as 0.8. The frontal direction of the uni-directional microphone is set to 0[deg]. As the sound sources, we utilized ”Japanese Newspaper Article Sentences” edited by the Acoustical Society of Japan. We conducted ten experiments utilizing the different kinds of sounds. The speech signals are arrived from two directions, 0[deg] and 45[deg]. Table.1 shows the true values and the estimated values of D_1 and D_2, respectively. The theoretical values of D_1 and D_2 are 1.25 and 0.89, respectively. We also show the wave form of the original and the separated sounds in Exp.1. Fig.5(a)-(1) and (2) represent the wave form of $s_1(t)$ and $s_2(t)$ in the time domain, respectively. Fig.5(b)-(1) and (2) represent the wave form of the mixed sounds $x_1(t)$ and $x_2(t)$ in the time domain, respectively. Fig.5(c)-(1) and (2) represent the wave form of the separated sounds, $\overline{s_1'}(t)$ and $\overline{s_2'}(t)$, respectively. In this experiment, two voices were spoken by the same person. Therefore, the cross correlation of two sounds may not be 0. However, two sounds are separated satisfactorily as shown in Fig.5(c).

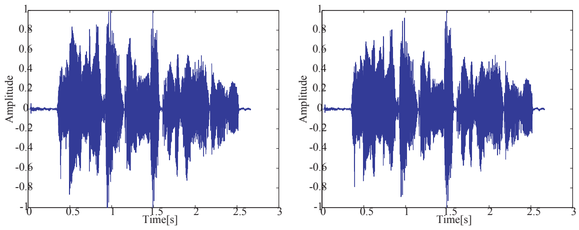
5. CONCLUSION

In this paper, we proposed a novel method to separate two sounds from different directions based on the gain difference between two microphones located in the same place. The system can separate the sounds which overlap in most of the time domain and the frequency domain. We need to know neither the directivity of two microphones nor the interval where the sound is disjoint in advance. We can also estimate the directivity of two microphones in the process of the sound separation. The proposed method inspires us to combine this method and the conventional aggregated microphones which utilize the knowledge about the directivity of the microphones. The proposed aggregated microphones system makes the system size small to be used in various application fields such as robotics and sound environment analyses. However, in this method, we assume two sounds. Hence, it is impossible to separate three or more sounds in this method. As future works, we are considering to separate $N(> 2)$ sounds using N microphones by extending the proposed method which can be applied in echoic environment.



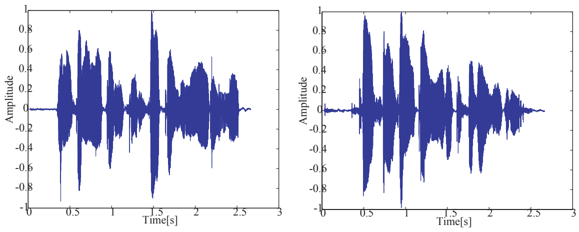
(1) Waveform of the original sound 1 (2) Waveform of the original sound 2

(a) Waveforms of the original sounds.



(1) Waveform of the mixed sound 1 (2) Waveform of the mixed sound 2

(b) Waveforms of the mixed sounds.



(1) Waveform of the separated sound 1 (2) Waveform of the separated sound 2

(c) Waveforms of the separated sounds. $D_1 = 1.2500$, $D_2 = 0.8900$, $\overline{D_1} = 1.1395$, $\overline{D_2} = 0.8171$

Figure 5: The experimental results of the sound separation: The sound sources are Japanese speaking utilizing Japanese Newspaper Article Sentences

REFERENCES

[1] J. Herault and C. Jutten, ”Space or time adaptive signal processing by neural models,” *Proc. AIP Conference: Neural Networks for Computing.*, pp.206-211, 1986.

[2] A. J. Bell and T. J. Sejnowski, ”An information maximization approach to blind separation and blind deconvolution,” *Neural Comput.*, vol.7 , pp.1129-1159, 1995.

[3] J. F. Cardoso, ”Blind signal separation: statistical principles,” *Proc. of IEEE.* vol.86, no.10, pp. 2009-2025, 1998.

[4] D. T. Pham and J. F. Cardoso, “ Blind separation of instantaneous mixtures of non-stationary sources, ”*IEEE Trans. on Signal Processing*, Vol.49 No.9 pp.1837-1848, 2001.

[5] T. Ihara, M. Handa, T. Nagai and A. Kurematsu, ”Multi-channel speech separation and localization by frequency assingment,” *IEICE trans on Fundamentals.*, Vol.J86-A, No.10, pp.998-1009, 2003.

[6] S. Rickard and O. Yilmaz, ”On the approximate w-disjoint orthogonality of speech,” *Proc. ICASSP 2002.*, Orlando, Florida, 2002, pp.529-532.

[7] O. Yilmaz and S. Rickard, ”Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. on Signal Processing*, Vol. 52, No. 7, 2004.

[8] M. Aoki, M. Okamoto, S. Aoki, H. Matsi, T. Sakurai, and Y. Kaneda, ”Sound source segregation based on estimating incident angle of each frequency component of input sources acquired by multiple microphones,” *Acoust. Sci and Tech.*, Vol.22, No.2, pp.149-157, 2001.

[9] M. Aoki, Y. Yamaguchi, K. Furuya and A. Kataoka, ”Modified SAFIA: Separation of target signal close to the microphones and noise signal far from the microphones,” *IEICE trans on Fundamentals.*, Vol.E88-A, No.4, pp.468-479, 2005.

[10] K. Kiyohara, Y. Kaneda, S. Takahashi, H. Nomura and J. Kojima, ”A microphone array system for speech recognition,” *Proc. ICASSP97.*, Munich, Germany, 1997, pp.215-218.

[11] Y. Kaneda, and J. Ohga, ”Adaptive microphone array system for noise reduction,” *IEEE Trans. Acoust. Speech Source Process.*, Vol.ASSP-34, No.6, pp.1391-1400, 1986.

[12] K. Takao, M. Fujita, and T. Nishi, ”An adaptive antenna array under directional constraint,” *IEEE Trans. Antennas Propagat.*, Vol.24, pp.662-669, 1976.

[13] J. Capon, ”High resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE*, Vol.57, pp.2408-2418, 1969.

[14] R.O. Schmidt, ”Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propagat.*, vol.AP-34, pp.276-280, 1986.

[15] M. Matsumoto, and S. Hashimoto, ”Multiple signal classification by aggregated microphones,” *IEICE trans on Fundamentals.*, Vol.E88-A, No.7, pp.1701-1707, 2005.

[16] M. Matsumoto, and S. Hashimoto, ”Minimum variance method by aggregated microphones,” *Proc. ISSPA 2005.*, Sydney, Australia, 2005, pp.879-882.

[17] M. Matsumoto, and S. Hashimoto, ”A miniaturized adaptive microphone array under directional constraint utilizing aggregated microphones,” *J. Acoustic Society America*, Vol.119, No.1, pp.352-359, 2006.

[18] P. D. O’Grady, B. A. Pearlmutter, and S. T. Rickard. ”Survey of Sparse and Non-Sparse Methods in Source Separation,” *International Journal of Imaging Systems and Technology*, Vol.15, pp.18-33, 2005.

[19] M. Matsumoto and S. Hashimoto, ”Modified SAFIA utilizing aggregated microphones,” *Proc. SPPRA 2006.*, Innsbruck, Austria, pp.222-227, 2006.

[20] T. Okada, H. Sato and K. Mitomi ”Direction measurement of a sound source in 3-D space utilizing unidirectional characteristics of microphone’s sensitivity,” *Trans of SICE.*, Vol.37, No.1, pp.13-20, 2001.