EFFECT OF PCA FILTER IN BLIND SOURCE SEPARATION

Futoshi Asano, Yoichi Motomura, Hideki Asoh and Toshihiro Matsui

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Japan
asano/motomura/asoh/matsui@etl.go.jp

Abstract

A problem of applying blind source separation (BSS) to an acoustically mixed signal in a real environment is room reflections. In array processing, room reflections can be reduced by the subspace method. In this paper, the subspace method is realized as a PCA (principal component analysis) stage in the BSS framework. Experimental results show that room reflection was reduced by the PCA filter by around 10 dB.

1. INTRODUCTION

The purpose of blind source separation (BSS) is to separate a mixture of signals from multiple sources without any prior knowledge of signals and sound field. When applying BSS to an acoustic mixture such as mixed voices from multiple talkers in a room, that is known as a cocktail party problem, BSS must solve a convolved mixture problem. The difficulty in this acoustic problem is the reflections of a room as depicted in Fig. 1. In this case, the filter network **B** that separates the mixed sources becomes a *spatial inverse*. When the mixing system **A** is known in advance, an efficient algorithm for building the inverse system **B** had already been proposed [1]. However, in a blind problem, in which the information of **A** is unknown, the estimation of **B** is extremely difficult, especially for rooms with rich reflections. For example, the meeting room used in the experiment in this paper has a reverberation time of over 0.4 s. This means that the mixing system **A** becomes a FIR filter network of over 6000 filter-taps when the sampling frequency is 16 kHz, and it is considered to be impractical to identify thousands of parameters in a real situation.

The authors have developed a method of reducing reflections and ambient noise using the array processing based on the subspace method [2, 3]. In this method, dominant directional signals and less-directional reflections are classified in the subspace domain based on the spatial extent of the signals. By reducing reflections prior to the separation by the subspace method, the separation problem becomes much simpler and easier. The subspace method is essentially the same as the principal component analysis (PCA,) and can easily be introduced to the BSS framework that includes PCA at the first stage of separation such as [4, 5, 6, 7]. In this paper, an aspect of PCA in BSS which functions as the subspace method is described and its effect on the room reflections are investigated via experiments.

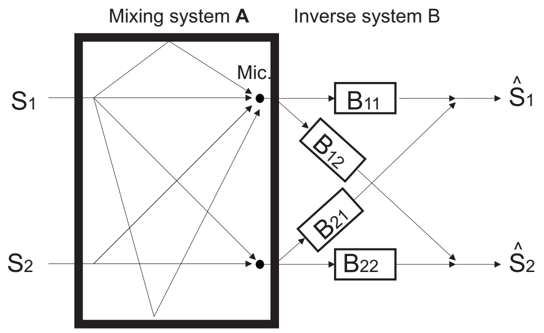


Figure 1: BSS system for an acoustic mixture.

2. SUBSPACE APPROACH

2.1. Model of Acoustic Environment

Let us consider the case when there are D sound sources in the environment. By observing this sound field with M microphones, we obtain the input vector at the tth time frame:

$$\mathbf{x}(t) = [X_1(t), \cdots, X_M(t)] = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t). \quad (1)$$

The symbol $X_m(t)$ is the short-term Fourier transform (STFT) of the input signal at the mth microphone. Matrix $\mathbf{A}$ is termed the mixing matrix, its (m, d) element $A_{m,d}$ being the transfer function from the dth source to the mth microphone. Vector $\mathbf{s}(t)$ consists of the source spectra as $\mathbf{s} = [S_1(t), \cdots, S_D(t)]^T$, where $S_d(t)$ denotes the spectrum of the dth source. The first term, $\mathbf{As}(t)$, expresses the directional components in $\mathbf{x}(t)$. On the other hand, the second term, $\mathbf{n}(t)$, corresponds to a mixture of the other components which spatially spreads and is less directional such as room reflections and ambient noise.

2.2. Spatial Correlation Matrix

The spatial correlation matrix is defined using the input vector $\mathbf{x}(t)$ as

$$\mathbf{R} = E[\mathbf{x}(t)\mathbf{x}^H(t)]. \tag{2}$$

where $\cdot^H$ denotes the Hermitian transpose. Assuming that $\mathbf{s}(t)$ and $\mathbf{n}(t)$ are uncorrelated, the spatial correlation matrix can be written using (1) as

$$\mathbf{R} = \mathbf{APA}^H + \mathbf{K}, \tag{3}$$

where $\mathbf{P} = E[\mathbf{s}(t)\mathbf{s}^H(t)]$ and $\mathbf{K} = E[\mathbf{n}(t)\mathbf{n}^H(t)]$. When $\mathbf{n}(t)$ includes the room reflections of $\mathbf{s}$(t), the above assumption does not hold. However, when the window length of STFT is short and the time interval between the direct sound and the reflection exceeds this window length, and, moreover, if the source signal is nonstationary such as speech, reflections behave like an *incoherent* additive noise and the above assumption holds to some extent in a practical sense. A typical example of this is the consonant portion of speech overlapped by the reflections of the preceding vowel.

2.3. Subspace Method

By taking the generalized eigenvalue decomposition of $\mathbf{R}$ as

$$\mathbf{R} = \mathbf{KE\Lambda E}^{-1}, \tag{4}$$

we have the eigenvector matrix $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_M]$ and the eigenvalue matrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_M)$ where $\mathbf{e}_m$ and λ_m are the eigenvector and the eigenvalue, respectively. The eigenvalues and eigenvectors have the following properties [3, 8]:

1. The energy of the D directional signals $\mathbf{s}(t)$ is concentrated on the D dominant eigenvalues.

2. The energy of $\mathbf{n}(t)$ is equally spread over all eigenvalues.

3. $\Re(\mathbf{A}) = \Re(\mathbf{E}_s)$, where $\mathbf{E}_s = [\mathbf{e}_1, \cdots, \mathbf{e}_D]$ denotes the eigenvectors corresponding to the D dominant eigenvalues and $\Re(\mathbf{A})$ denotes the space spanned by the column vectors of $\mathbf{A}$ (the column space).

4. $\Re(\mathbf{A}) = \Re(\mathbf{E}_n)^\perp$, where $\mathbf{E}_n = [\mathbf{e}_{D+1}, \cdots, \mathbf{e}_M]$ denotes the eigenvectors corresponding to the other $M - D$ eigenvalues.

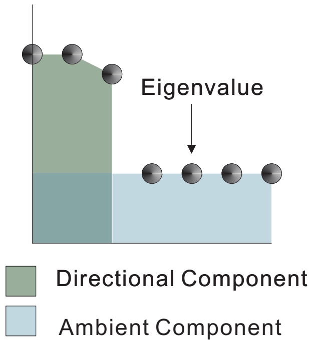


Figure 2: Typical eigenvalue distribution.

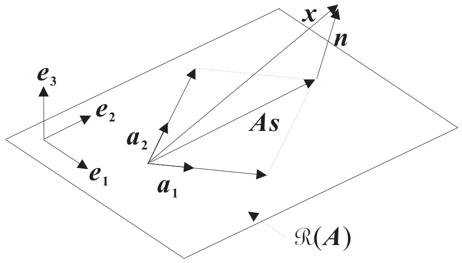


Figure 3: Relation of vectors.

A typical eigenvalue distribution and the corresponding energy distribution, that reflects Properties 1 and 2, is depicted in Fig. 2. Also, a geometric relation of vectors, that reflects Properties 3 and 4, is depicted in Fig. 3. These properties have been utilized in the array processing for the sound source localization (direction-of-arrival estimation) in the radar/sonar application (e.g. [8, 9]) and the ambient noise reduction in the speech enhancement application [3].

2.4. PCA Filter

Let us consider the following filtering, in which the filter consists of the eigenvectors corresponding to the D dominant eigenvalues $\mathbf{E}_s$:

$$\mathbf{y}(t) = \mathbf{E}_s^H \mathbf{x}(t). \tag{5}$$

Due to the orthogonality of Property 4, the component of $\mathbf{n}(t)$ belonging to the subspace $\Re(\mathbf{E}_n)$ (denoted as $\mathbf{n}_n(t)$ hereafter) is canceled by this filtering operation, while $\mathbf{s}(t)$ is perfectly preserved. Since this filter is derived by the eigenvalue decomposition of the correlation matrix, that is equivalent to the principal component analysis (PCA), this filter is termed the PCA filter hereafter.

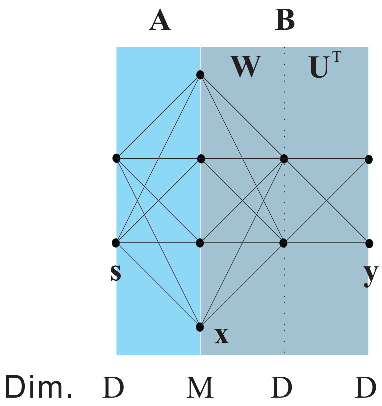


Figure 4: Filter network of BSS with the extended PCA." Dim" indicates the dimension of vectors at each node.

3. BLIND SOURCE SEPARATION USING PCA FILTER

As a BSS algorithm, the time-delayed decorrelation (TDD, [4, 5, 6]), that had been extended to the convolved mixture problem [7], is employed. In this section, the TDD algorithm is briefly explained. Then, how the PCA filter is realized in TDD is described.

3.1. Time-delayed Decorrelation

The separation system is written as

$$\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t). \tag{6}$$

The separation filter $\mathbf{B}$ can be decomposed into

$$\mathbf{B} = \mathbf{U}^H\mathbf{W}. \tag{7}$$

The matrix $\mathbf{W}$ is determined by the following equation [4]:

$$\mathbf{W} = \mathbf{\Lambda}_s^{-\frac{1}{2}}\mathbf{E}_s^H. \tag{8}$$

where $\mathbf{\Lambda}_s$ is defined as $\mathbf{\Lambda}_s = \mathrm{diag}(\lambda_1, \cdots, \lambda_{\mathrm{D}})$. The matrix $\mathbf{U}$ is determined by the joint-diagonalization problem[5, 6]:

$$\mathbf{U}\bar{\boldsymbol{R}}(\tau)\mathbf{U}^H = \mathrm{diag}(\sigma_1, \cdots, \sigma_D), \qquad \forall\tau \neq 0 \tag{9}$$

where

$$\bar{\boldsymbol{R}}(\tau) = \mathbf{W}\mathbf{R}(\tau)\mathbf{W}^H, \tag{10}$$

and

$$\mathbf{R}(\tau) = E[\mathbf{x}(t)\mathbf{x}^H(t-\tau)]. \tag{11}$$

The matrix $\mathbf{R}(\tau)$ is termed the time-delayed correlation (TDC) and $\mathbf{R}(0) = \mathbf{R}$. In the joint-diagonalization, the off-diagonal elements in (9) are reduced to 0 while the diagonal elements, $\sigma_1, \cdots, \sigma_D$, may have arbitrary values.

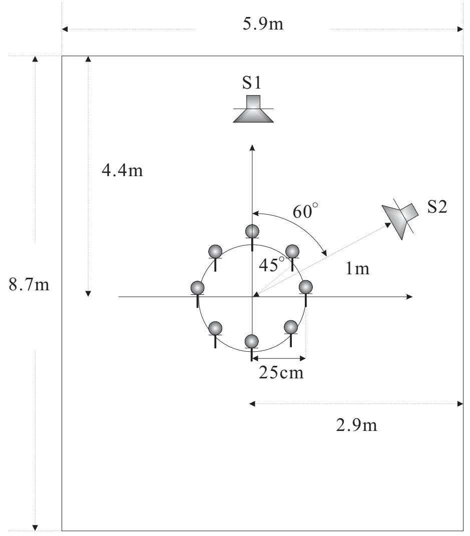


Figure 5: Configuration of sound sources and microphones.

3.2. Realization of PCA Filter

As indicated in (8), the PCA filter, which is normalized by the corresponding eigenvalues, has already been introduced as the filter $\mathbf{W}$ in TDD. For taking advantage of the cancellation of $\mathbf{n}_n(t)$ by the PCA filter, the number of microphones, M, must be greater than that of the sources, D. When $M > D$, (8) is termed the extended PCA in this paper for the sake of convenience. Moreover, D must be known to select $\mathbf{\Lambda}_s$ and $\mathbf{E}_s$ from $\mathbf{\Lambda}$ and $\mathbf{E}$ (subspace selection). When D is unknown, D can be estimated from the eigenvalue distribution using the AIC or MDL criterion [10].

The BSS filter network including the extended PCA is depicted in Fig. 4. In summary of the above discussion, the first-stage filter $\mathbf{W}$ has the following two roles:

- subspace selection
- orthogonalization

In the subspace selection, the D-dimensional subspace $\Re(\mathbf{A})$, in which D dominant directional sources are included, is selected from the M-dimensional space. During this process, a portion of the reflections and the ambient noise $\mathbf{n}_n(t)$ belonging to the orthogonal compliment $\Re(\mathbf{A})^\perp$ is canceled. In the second role, the output of the filter $\mathbf{W}$ is orthogonalized as a preparation of the second stage. The second-stage filter $\mathbf{U}$ finally separates the independent sources from the selected D-dimensional subspace $\Re(\mathbf{A})$.

Figure 6: A scene of experiment.

4. EXPERIMENT

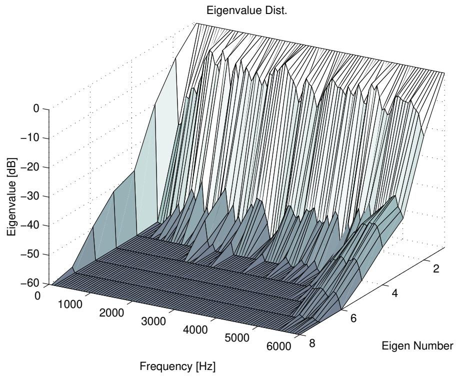
4.1. Condition

A signal separation experiment was conducted in usual meeting room with a reverberation time of 0.4 s. The configuration of the sound sources (loudspeakers) and the microphones is depicted in Fig. 5. A microphone array with $M = 8$, mounted on the top of a mobile robot Nomad XR-4000 (see Fig. 6), was used. The microphone array was circular in shape with a diameter of 50 cm. The impulse responses from the sound sources to the microphones were measured and then convolved with the source signal (speech) to generate the input signal $\mathbf{x}(t)$.

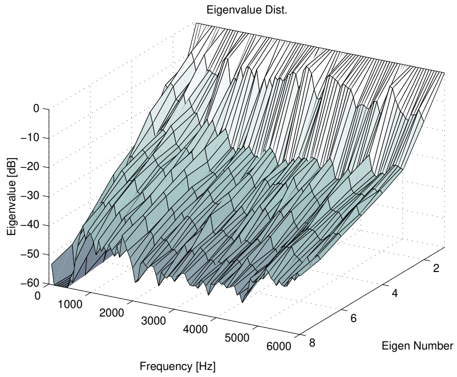
For training the filter network $\mathbf{W}$ and $\mathbf{U}$, the correlation matrix $\mathbf{R}(\tau)$ was calculated from the input signal with a duration of 1.5-3.0 s, and then the filter matrices, $\mathbf{W}$ and $\mathbf{U}$, were obtained by using the algorithm described in Section 3.1 for each frequency. The window length in STFT was 256 points (16 ms) and the frame shift was 32 points. The sampling frequency was 16 kHz. The spatial correlation matrix of $\mathbf{n}(t)$, $\mathbf{K}$, was unknown and was assumed to be $\mathbf{K} = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. When the reverberation time of the room is long as in the case of this experiment, the cross-correlation of reflections between the microphones is reduced, and this assumption holds to some extent in a practical sense. For solving the permutation problem in the extended TDD, the method using the inter-frequency correlation [7] was used.

4.2. Results

Figure 7 shows the eigenvalue distribution of $\mathbf{R}$. For the sake of comparison, the eigenvalue distribution without reflection is also shown. By comparing these, it can be seen that the assumption in Section 2.2 holds to some extent, since, if $\mathbf{As}(t)$ and $\mathbf{n}(t)$ is perfectly correlated, the effective rank of $\mathbf{R}$ with reflection (Fig. 7(b))



(a) Without reflection



(b) With reflection

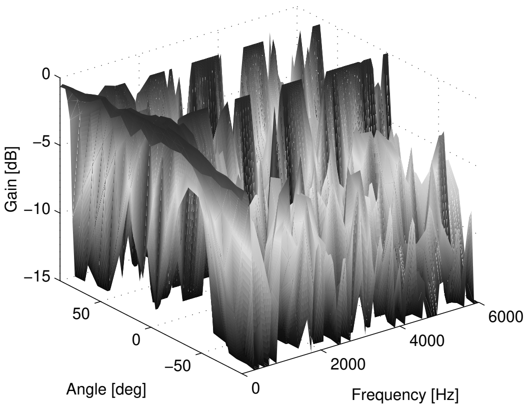
Figure 7: Eigenvalue distribution.

is also reduced to 2 as that of $\mathbf{R}$ without reflection (Fig. 7(a)).

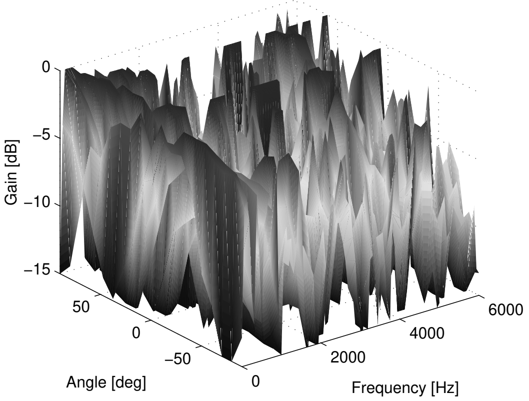
Figure 8 shows the directivity pattern of the PCA filter. As can be seen in this figure, the sensitivity in the directions of the sources, i.e., 0° and 60° , is high while that in the other directions is low. This means that the PCA filter works as a beamformer focused in the direction of the target sound sources.

Figure 9 shows the input/output spectra. For evaluating the effect of the PCA filter, the input signal for the filtering was decomposed into the direct sound and the reflection. This was done by separating the impulse response into its direct and reflective parts in the time domain and by separately convolving their sources. The figure shows the spectra relative to the direct sound. For the sake of comparison, the case of $M = D = 2$, which corresponds to the conventional BSS, is also shown as (b) in Fig. 9. In this case, two of the 8 microphones closest to the sources were used. The output (c) is the case with $M = 8$. By comparing (b) and (c), it can be seen that the reflection for both source #1 and #2 were reduced in (c). This is the effect of the PCA filter.

Table 1 shows the results of the automatic speech



(a) Channel 1



(b) Channel 2

Figure 8: Directivity pattern of PCA filter.

recognition (ASR) applied to the output of BSS. In this test, 492 Japanese words were recognized by an HMM recognizer (HTK). As a speaker distance (the distance from the center of the array to each speaker), not only $r = 1.0$ m as depicted in Fig. 5 but also $r = 0.6$ m was examined. As can be seen from this table, the recognition rate was improved by around 10-15% by the PCA filter.

5. CONCLUSIONS

The effect of the PCA filter in blind source separation was investigated. When the number of microphones is greater than that of the sound sources, the PCA process has an aspect of the subspace method in the array processing, and reduces incoherent reflections in rooms. In this case, PCA filter works as an adaptive beamformer with unsupervised learning, that focuses on the target sources. The essential difference of the PCA filter and the conventional beamformer is that the PCA filter does not use an array response database (prior knowledge of the array), an important feature in the BSS framework. As experimental results, reduction of reflection by around 10 dB was obtained in a real room environment. In the evaluation test using automatic speech recognition, improvement of recognition rate by around 10-15 % was obtained. However, the recognition rate was still low in the highly reflective room used in the experiment and further improvement is required.

Table 1: Speech recognition rate [%].

Room	Source	(a) $M = 2$	(b) $M = 8$
Anechoic	Ch.1	58.7	58.1
Chamber	Ch.2	58.1	59.3
Meeting	Ch.1	26.0	38.8
Rm. (r=0.6)	Ch.2	20.3	30.3
Meeting	Ch.1	18.3	28.5
Rm. (r=1.0)	Ch.2	14.2	29.5

6. REFERENCES

[1] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustics,” *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 36, pp. 145–152, 1988.

[2] F. Asano and S. Hayamizu, “Speech enhancement using array signal processing based on the coherent-subspace method,” *IEICE Trans. Fundamentals*, vol. E80-A(11), pp. 2276–2285, November 1997.

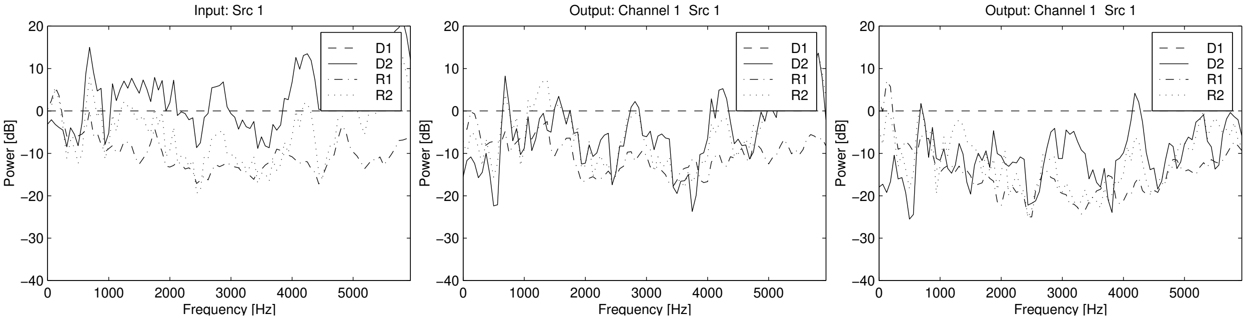
[3] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, “Speech enhancement based on the subspace method,” *IEEE Trans. Speech, Audio Processing*, in printing.

[4] L. Tong, R.-W. Liu, V. C. Soon, and Y.-F. Huang, “Indeterminacy and identifiability of blind identification,” *IEEE Trans. Circuits and Systems*, vol. 38, pp. 499–509, May 1991.

[5] L. Molgedey and H. G. Schuster, “Separation of a mixture of independent signals using time delayed correlations,” *Physical Review Letters*, vol. 72(23), pp. 3634–3637, 1994.

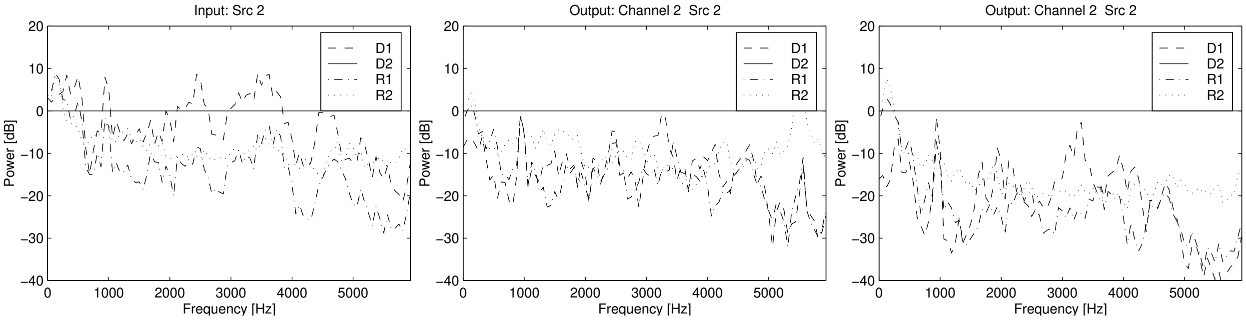
[6] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, “A blind source separation technique using second-order statistics,” *IEEE Trans. Signal Process*, vol. 45(2), pp. 434–443, February 1997.

Ch.1



(a) input (b) output (M=D=2) (c) output (M=8,D=2)

Ch.2



(a) input (b) output (M=D=2) (c) output (M=8,D=2)

Figure 9: Input/output of the system. D1,D2:Direct sound, R1,R2:Reflection.

[7] S. Ikeda and N. Murata, "A method of blind separation based on temporal structure of signals," In *Proceedings of The Fifth International Conference on Neural Information Processing (ICONIP'98 Kitakyushu)*, pp. 737–742, 1998.

[8] D. H. Johnson and D. E. Dudgeon, *Array signal processing*, Prentice Hall, Englewood Cliffs NJ, 1993.

[9] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag*, vol. AP-34(3), pp. 276–280, March 1986.

[10] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-33, pp. 387–392, Apr. 1985.