社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
EA2003-45 (2003-06)

[招待論文]独立成分分析に基づくブラインド音源分離

牧野　昭二†　荒木　章子†　向井　良†　澤田　宏†

† NTT コミュニケーション科学基礎研究所　〒 619-0237 京都府相楽郡精華町光台 2-4
E-mail: †{maki,shoko,ryo,sawada}@cslab.kecl.ntt.co.jp

あらまし　私たちが普段それほど意識せずに行っている「聞きたい音を聞き分ける」という能力がコンピュータには欠けている．独立成分分析に基づく手法は，ある人が話している声と別の人の声，背景に流れる音楽，雑音等，それぞれの音は互いに統計的に独立であるという仮定により，複数のマイクで観測した信号を互いに独立な信号に分離すれば，それぞれのもとの音を復元できる，という原理に基づいている．この手法は，音源や混合系の情報を原理的に必要としない，いわゆるブラインドな分離が可能である．招待講演では，独立成分分析とは何か，ブラインド音源分離とは何か，どのようにして分離が達成されるのか，分離のメカニズムはどのようなものか，などについて，できるだけ直感的に分り易く説明する [1].

キーワード　独立成分分析，ブラインド音源分離，適応ビームフォーマ

ICA-Based Audio Source Separation

Shoji MAKINO†, Shoko ARAKI†, Ryo MUKAI†, and Hiroshi SAWADA†

† NTT Communication Science Laboratories　2-4 Hikaridai, Seika-cho Soraku-gun, Kyoto 619-0237, Japan
E-mail: †{maki,shoko,ryo,sawada}@cslab.kecl.ntt.co.jp

Abstract　This paper introduces the blind source separation (BSS) of convolutive mixtures of acoustic signals, especially speech. A statistical and computational technique, called independent component analysis (ICA), is examined. By achieving nonlinear decorrelation, nonstationary decorrelation, or time-delayed decorrelation, we can find source signals only from observed mixed signals. Particular attention is paid to the physical interpretation of BSS from the acoustical signal processing point of view. Frequency-domain BSS is shown to be equivalent to two sets of frequency domain adaptive microphone arrays, i.e., adaptive beamformers (ABFs). Although BSS can reduce reverberant sounds to some extent in the same way as ABF, it mainly removes the sounds from the jammer direction. This is why BSS has difficulties with long reverberation in the real world. If sources are not "independent," the dependence results in bias noise when obtaining the correct unmixing filter coefficients. Therefore, the performance of BSS is limited by that of ABF. Although BSS is upper bounded by ABF, BSS has a strong advantage over ABF. BSS can be regarded as an intelligent version of ABF in the sense that it can adapt without any information on the array manifold or the target direction, and sources can be simultaneously active in BSS.

Key words　independent component analysis, blind source separation, adaptive beamformer

1. Introduction

Speech recognition is a fundamental technology for communication with computers, but with existing computers, the recognition rate drops rapidly when more than one person is speaking or when there is background noise. On the other hand, humans can engage in comprehensible conversations at a noisy cocktail party. This is the well known cocktail-party effect, where the individual speech waveforms are found from the mixtures. The aim of source separation is to provide computers with this cocktail party ability, thus making it possible for computers to understand what a person is saying at a noisy cocktail party.

Blind source separation (BSS) is an emerging technique, which enables the extraction of target speech from observed mixed speeches without the need for source positioning, spectral construction, or a mixing system. To achieve this, attention has focused on a method based on independent component analysis (ICA). ICA extracts independent sounds from among mixed sounds. This paper considers ICA in a wide

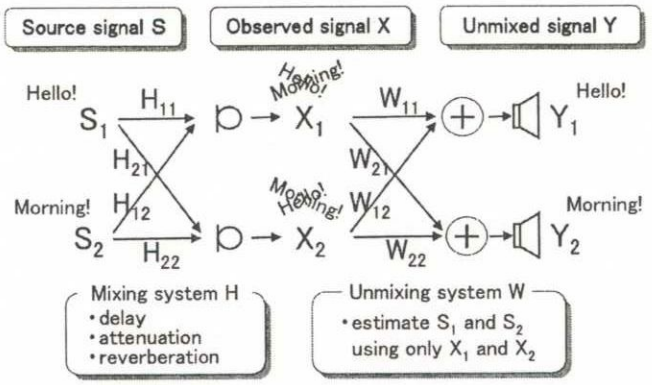

図 1 BSS system configuration.

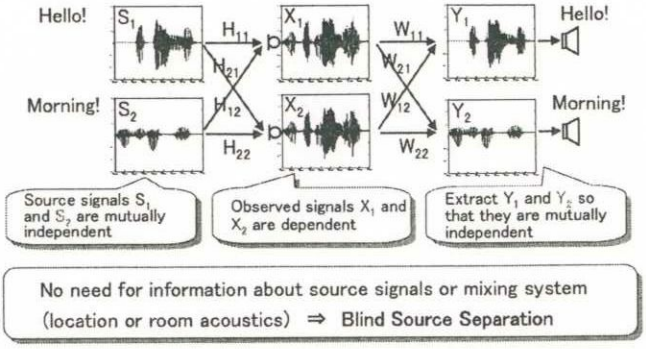

図 2 Task of blind source separation of speech signals.

sense, namely nonlinear decorrelation together with nonstationary decorrelation and time-delayed decorrelation. These three methods are discussed in a unified manner [2]. There are a number of applications for the BSS of mixed speech signals in the real world [3], but the separation performance is still not good enough [4], [5].

Since ICA is a purely statistical process, the separation mechanism has not been clearly understood in the sense of acoustic signal processing, and it has been difficult to know which components were separated, and to what degree. Recently, the ICA method has been investigated in detail, and its mechanisms have been gradually uncovered by using theoretical analysis from the perspective of acoustic signal processing [6] as well as experimental analysis based on impulse response [7]. The mechanism of BSS based on ICA has been shown to be equivalent to that of an adaptive microphone array system, i.e., N sets of adaptive beamformers (ABFs) with an adaptive null directivity aimed in the direction of unnecessary sounds.

From the equivalence between BSS and ABF, it becomes clear that the physical behavior of BSS reduces the jammer signal by making a spatial null towards the jammer. BSS can further be regarded as an intelligent version of ABF in the sense that it can adapt without any information on the source positions or period of source existence/absence.

2. What Is BSS?

Blind source separation (BSS) is an approach for estimating source signals $s_i(n)$ using only the information of mixed signals $x_j(n)$ observed at each input channel. Typical examples of such source signals include mixtures of simultaneous speech signals that have been picked up by several microphones, brain waves recorded by multiple sensors, and interfering radio signals arriving at a mobile station.

2.1 Mixed Signal Model for Speech Signals in a Room

In the case of audio source separation, several sensor microphones are placed in different positions so that each records a mixture of the original source signals at a slightly different time and level. In the real world where the source signals are speech and the mixing system is a room, the signals that are picked up by the microphones are affected by reverberation [8], [9]. Therefore, the N signals recorded by M microphones are modeled as

$$x_j(n) = \sum_{i=1}^{N} \sum_{p=1}^{P} h_{ji}(p) s_i(n-p+1) \ \ (j = 1, \cdots, M), \quad (1)$$

where s_i is the source signal from a source i, x_j is the signal received by a microphone j, and h_{ji} is the P-taps impulse response from source i to microphone j.

This paper focuses on speech signals as sources that are nongaussian, nonstationary, colored, and that have a zero mean.

2.2 Unmixed Signal Model

To obtain unmixed signals, unmixing filters $w_{ij}(k)$ of Q-taps are estimated, and the unmixed signals are obtained as

$$y_i(n) = \sum_{j=1}^{M} \sum_{q=1}^{Q} w_{ij}(q) x_j(n-q+1) \ \ (i = 1, \cdots, N). \quad (2)$$

The unmixing filters are estimated so that the unmixed signals become mutually independent. This paper considers a two-input, two-output convolutive BSS problem, i.e., $N = M = 2$ (Fig. 1) without a loss of generality.

2.3 Task of Blind Source Separation of Speech Signals

It is assumed that the source signals s_1 and s_2 are mutually independent. This assumption usually holds for sounds in the real world. There are two microphones which pick up the mixed speech. Only the observed signals x_1 and x_2 are available and they are dependent. The goal is to adapt the unmixing systems w_{ij}, and extract y_1 and y_2 so that they are mutually independent. With this operation, we can obtain s_1 and s_2 in the output y_1 and y_2. No information is needed on the source positions or period of source existence/absence.

Nor is any information required on the mixing systems h_{ji}. Thus, this task is called *blind* source separation (Fig. 2).

Note that the unmixing systems w_{ij} can at best be obtained up to a scaling and a permutation, and thus cannot itself solve the dereverberation/deconvolution problem [10].

2.4 Instantaneous Mixtures vs. Convolutive Mixtures

2.4.1 Convolutive Mixtures.

If the sound separation is being undertaken in a room, the mixing systems h_{ji} are of course FIR filters with several thousand taps. This is the very difficult and relatively new problem of *convolutive* mixtures.

2.4.2 Instantaneous Mixtures

By contrast, if the mixing systems h_{ji} are scalars, i.e., there is no delay and no reverberation, such as when we use an audio mixer, this becomes a problem of *instantaneous* mixtures.

In fact, other applications such as the fMRI and EEG signals found in biomedical contexts, and images are almost all instantaneous mixtures problems. Instantaneous mixtures problems have been well studied and there are many good results.

2.5 Time-Domain Approach vs. Frequency-Domain Approach

Several methods have been proposed for achieving the BSS of convolutive mixtures. Some approaches consider the impulse responses of a room h_{ji} as FIR filters, and estimate those filters in the time domain [11], [12], [13]; other approaches transform the problem into the frequency domain so that they can simultaneously solve an instantaneous BSS problem for every frequency [14], [15].

2.6 Time-Domain Approach for Convolutive Mixtures

In the time-domain approach for convolutive mixtures, unmixing systems w_{ij} can be FIR filters or IIR filters. However, FIR filters are usually used to realize a non-minimum-phase filter [11].

In the BSS of convolutive mixtures in the time domain, Sun and Douglas clearly distinguished *multichannel blind deconvolution* from *convolutive blind source separation* [13].

Multichannel blind deconvolution tries to make the output both spatially and temporally independent. The sources are assumed to be temporally as well as spatially independent, i.e., they are assumed to be independent from channel to channel and from sample to sample. On the other hand, convolutive BSS tries to make the output spatially (mutually) independent without deconvolution. Since speech is temporally correlated, convolutive BSS is appropriate for the task of speech separation. If we apply multichannel blind deconvolution to speech, it imposes undesirable constraints on the output, causing undesirable spectral equalization, flattening, or whitening. Therefore, we need some pre/post-filtering method that maintains the spectral content of the original speech in the separated output [13], [16].

An advantage of the time-domain approach is that we do not have to think about the heavy permutation problem, i.e., the estimated source signal components are recovered with a different order. Permutation poses a serious problem in relation to frequency-domain BSS, whereas it is a trivial problem in time-domain BSS.

A disadvantage of the time-domain approach is that the performance depends strongly on the initial value [11], [16].

2.7 Frequency-Domain Approach for Convolutive Mixtures

Smaragdis [14] proposed working directly in the frequency domain applying a nonlinear function to signals with complex values.

The frequency domain approach to convolutive mixtures is to transform the problem into an instantaneous BSS problem in the frequency domain [14], [15], [17], [18].

Using a T-point short-time Fourier transformation for (1), we obtain,

$$\mathbf{X}(\omega, m) = \mathbf{H}(\omega)\mathbf{S}(\omega, m), \tag{3}$$

where ω denotes the frequency, m represents the time-dependence of the short-time Fourier transformation, $\mathbf{S}(\omega, m) = [S_1(\omega, m), S_2(\omega, m)]^T$ is the source signal vector, and $\mathbf{X}(\omega, m) = [X_1(\omega, m), X_2(\omega, m)]^T$ is the observed signal vector. We assume that the (2×2) mixing matrix $\mathbf{H}(\omega)$ is invertible, and that $H_{ji}(\omega) \neq 0$. Also, $\mathbf{H}(\omega)$ does not depend on time m.

The unmixing process can be formulated in a frequency bin ω:

$$\mathbf{Y}(\omega, m) = \mathbf{W}(\omega)\mathbf{X}(\omega, m), \tag{4}$$

where $\mathbf{Y}(\omega, m) = [Y_1(\omega, m), Y_2(\omega, m)]^T$ is the estimated source signal vector, and $\mathbf{W}(\omega)$ represents a (2×2) unmixing matrix at frequency bin ω. The unmixing matrix $\mathbf{W}(\omega)$ is determined so that $Y_1(\omega, m)$ and $Y_2(\omega, m)$ become mutually independent. The above calculation is carried out at each frequency independently. This paper considers the DFT frame size T to be equal to the length of unmixing filter Q.

Hereafter, the convolutive BSS problem is considered in the frequency domain unless stated otherwise. Note that digital signal processing in the time and frequency domains are essentially identical, and all discussions here in the frequency domain are also essentially true for the time-domain convolutive BSS problem.

2.8 Scaling and Permutation Problems

Applying the model in the frequency domain introduces a new problem: the frequency bins are treated as being mutually independent. As a result, the estimated source signal

components are recovered with a different order in the different frequency bins. Thus the trivial *permutation* ambiguity associated with time-domain ICA, i.e., the ordering of the sources, now becomes nontrivial.

In frequency-domain BSS, the *scaling* problem also becomes nontrivial, i.e., the estimated source signal components are recovered with a different gain in the different frequency bins. The scaling ambiguity in each frequency bin results in a convolutive ambiguity for each source, this results in an arbitrary filtering. This reflects the fact that filtered versions of independent signals remain independent.

3. What Is ICA?

Independent component analysis (ICA) is a statistical method that was originally introduced in the context of neural network modeling [19], [20], [21], [22], [23], [24], [25], [26]. Recently, this method has been used for the BSS of sounds, fMRI and EEG signals of biomedical applications, wireless communication signals, images, and other applications. ICA thus became an exciting new topic in the fields of signal processing, artificial neural networks, advanced statistics, information theory, and various application fields.

Very general statistical properties are used in ICA theory, namely information on statistical independence. In a source separation problem, the source signals are the "independent components" of the data set. In brief, BSS poses the problem of finding a linear representation in which the components are mutually independent. ICA consists of estimating both the unmixing matrix $\mathbf{W}(\omega)$ and sources s_i, when we only have the observed signals x_j.

The unmixing matrix $\mathbf{W}(\omega)$ is determined so that one output contains as much information on the data as possible. The value of any one of the components gives no information on the values of the other components. If the unmixed signals are mutually independent, then they are equal to the source signals.

3.1 What Is Independence?

Independence is a stronger concept than "no correlation," since correlation only deals with second-order statistics whereas independence deals with higher-order statistics. Independent components can be found by nonlinear, nonstationary, or time-delayed decorrelation.

In the nonlinear decorrelation approach, if the unmixing matrix $\mathbf{W}(\omega)$ is a true separating matrix and y_1 and y_2 are independent and have a zero mean, and the nonlinear function $\Phi(\cdot)$ is an odd function such that $\Phi(y_1)$ also has a zero mean, then

$$E[\Phi(y_1)y_2] = E[\Phi(y_1)]E[y_2] = 0. \tag{5}$$

We look for such unmixing matrix $\mathbf{W}(\omega)$ that gives (5). The question here concerns, how should the nonlinear function be chosen?

The answers can be found by using several background theories for the ICA. Using any of these theories, we can determine the nonlinear function in a satisfactory way. These are the minimization of mutual information, maximization of nongaussianity, and maximization of likelihood.

For the nonstationary and time-delayed decorrelation approaches, see Sect. 4.

3.2 Minimization of Mutual Information

The first approach for ICA, inspired by information theory, is the minimization of mutual information. Mutual information is a natural information-theoretic measure of statistical independence. It is always nonnegative, and zero if, and only if, the variables are statistically independent. Therefore it is natural to estimate the independent components by minimizing the mutual information of their estimates. Minimization of mutual information can be interpreted as giving the maximally independent component.

3.3 Maximization of Nongaussianity

The second approach is based on the maximization of nongaussianity. The central limit theorem in probability theory says that the distribution of a sum of independent random variables tends toward a Gaussian distribution. Roughly speaking, the sum of independent random variables usually has a distribution that is closer to Gaussian than either of the original random variables. Therefore, the independent components can be found by finding the directions in which the data is maximally nongaussian.

Note that in most classic statistical theories, random variables are assumed to have a Gaussian distribution. By contrast, in the ICA theory, random variables are assumed to have a nongaussian distribution.

Many real-world data sets, including speech, have supergaussian distributions. Supergaussian random variables typically have a spiky probability density function (pdf), i.e., the pdf is relatively large at zero compared with the Gaussian distribution. A speech signal is a typical example (Fig. 3).

3.4 Maximization of Likelihood

The third approach is based on the maximization of likelihood. Maximum likelihood (ML) estimation is a fundamental principle of statistical estimation, and a very popular approach for estimating the ICA. We take the ML estimation parameter values as estimates that give the highest probability for the observations.

ML estimation is closely related to the neural network principle of maximization of information flow (infomax). The infomax principle is based on maximizing the output entropy, or information flow, of a neural network with nonlinear outputs. We maximize the mutual information between the inputs x_i and outputs y_i. Maximization of this mutual information is equivalent to a maximization of the output entropy,

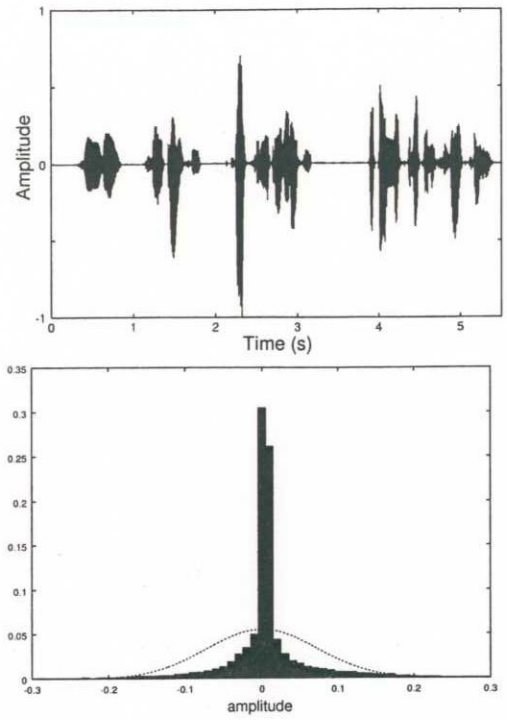

図 3 Speech signal and its probability density function (pdf). Dotted line is the pdf of the Gaussian distribution.

so infomax is equivalent to maximum likelihood estimation.

3.5 Three ICA Theories Are Identical

It is of interest to note that all the above solutions are identical [27]. The mutual information $I(y_1, y_2)$ between the output y_1 and y_2 is expressed as

$$I(y_1, y_2) = \sum_{i=1}^{2} H(y_i) - H(y_1, y_2), \tag{6}$$

where $H(y_i)$ are the marginal entropies and $H(y_1, y_2)$ is the joint entropy of the output. The entropy of y can be calculated by using $p(y)$ (pdf of y) as follows:

$$H(y) = E[\log \frac{1}{p(y)}] = \sum p(y) \log \frac{1}{p(y)}. \tag{7}$$

Mutual information $I(y_1, y_2)$ in Sect. 3.2 is minimized by minimizing the first term, or maximizing the second term of (6). Gaussian signals maximize the first term, namely maximization of nongaussianity in Sect. 3.3 achieves minimization of the first term. On the other hand, maximization of the joint entropy of the output in Sect. 3.4 maximizes the second term. Accordingly, the above mentioned three approaches are identical. For more details of these theories, see [11], [28], [29], [30].

3.6 Learning Rules

To achieve separation, we vary the unmixing matrix $\mathbf{W}(\omega)$ in (4), and see how the distribution of the output changes. We search for the unmixing matrix $\mathbf{W}(\omega)$ that minimizes the mutual information, maximize the nongaussianity, or maximize the likelihood of the output. This can be accomplished by the gradient method.

Bell and Sejnowski derived a very simple gradient algorithm [31]. Amari proposed the natural gradient version, and increased the stability and convergence speed [32]. This is a nonlinear extension of the ordinary requirement of uncorrelatedness, and in fact, this algorithm is a special case of the nonlinear decorrelation algorithm. The theory makes it clear that the nonlinear function must correspond to the derivative of the logarithm of the pdf of the sources.

Hereafter, we assume that the pdf of the (speech) sources is known, that is, the supergaussian distribution of the speech sources is known. It also assumes that the nonlinear function is set in a satisfactory way that corresponds to the derivative of the logarithm of the pdf, namely the nonlinear function is properly set at $\tanh(\cdot)$.

4. How Speech Signals Can Be Separated?

This paper attempts a simple and comprehensive (rather than accurate) exploration from the acoustical signal processing perspective. With the ICA-based BSS framework, how can we separate speech signals?

The simple answer is to diagonalize $\mathbf{R}_Y$, where $\mathbf{R}_Y$ is a (2×2) matrix:

$$\mathbf{R}_Y = \begin{bmatrix} \langle \Phi(Y_1) Y_1 \rangle & \langle \Phi(Y_1) Y_2 \rangle \\ \langle \Phi(Y_2) Y_1 \rangle & \langle \Phi(Y_2) Y_2 \rangle \end{bmatrix}. \tag{8}$$

The function $\Phi(\cdot)$ is a nonlinear function. The operation $\langle \cdot \rangle$ is the averaging operation used to obtain statistical information. We want to minimize the off-diagonal components, while at the same time, constraining the diagonal components to proper constants.

The components of the matrix $\mathbf{R}_Y$ correspond to the mutual information between Y_i and Y_j. At the convergence point, the off-diagonal components, which are the mutual information between Y_1 and Y_2, become zero:

$$\langle \Phi(Y_1) Y_2 \rangle = 0, \quad \langle \Phi(Y_2) Y_1 \rangle = 0. \tag{9}$$

While at the same time, the diagonal components, which only control the amplitude scaling of the output Y_1 and Y_2, are constrained to proper constants:

$$\langle \Phi(Y_1) Y_1 \rangle = c_1, \quad \langle \Phi(Y_2) Y_2 \rangle = c_2. \tag{10}$$

To achieve this convergence, we use the recursive learning rule

$$\mathbf{W}_{i+1} = \mathbf{W}_i + \eta \Delta \mathbf{W}_i, \tag{11}$$

$$\Delta \mathbf{W}_i = \begin{bmatrix} c_1 - \langle \Phi(Y_1) Y_1 \rangle & \langle \Phi(Y_1) Y_2 \rangle \\ \langle \Phi(Y_2) Y_1 \rangle & c_2 - \langle \Phi(Y_2) Y_2 \rangle \end{bmatrix} \mathbf{W}_i. \tag{12}$$

When $\mathbf{R}_Y$ is diagonalized, $\Delta \mathbf{W}$ converges to zero.

If $c_1 = c_2 = 1$, the algorithm is called *holonomic*. If $c_1 = \langle \Phi(Y_1)Y_1 \rangle$ and $c_2 = \langle \Phi(Y_2)Y_2 \rangle$, the algorithm is called *nonholonomic*.

4.1 Second Order Statistics vs. Higher Order Statistics

If $\Phi(Y_1) = Y_1$, we have the simple decorrelation:

$$\langle \Phi(Y_1)Y_2 \rangle = \langle Y_1 Y_2 \rangle = 0. \tag{13}$$

This is not sufficient to achieve independence. However, if we have nonstationary sources, we have this equation for multiple time blocks, and thus can solve the problem. This is the *nonstationary decorrelation* approach [33].

Or, if we have colored sources, we have a delayed correlation for a multiple time delay:

$$\langle \Phi(Y_1)Y_2 \rangle = \langle Y_1(m)Y_2(m+\tau_i) \rangle = 0, \tag{14}$$

thus we can solve the problem. This is the *time-delayed decorrelation* (TDD) approach [34], [35].

These are the approaches of *second order statistics* (SOS).

On the other hand if, for example, $\Phi(Y_1) = \tanh(Y_1)$, we have:

$$\langle \Phi(Y_1)Y_2 \rangle = \langle \tanh(Y_1)Y_2 \rangle = 0. \tag{15}$$

With a Tailor expansion of $\tanh(\cdot)$, (15) can be expressed as

$$\langle (Y_1 - \frac{Y_1^3}{3} + \frac{2Y_1^5}{15} - \frac{17Y_1^7}{315} \ldots)\, Y_2 \rangle = 0, \tag{16}$$

thus we have higher order or *nonlinear decorrelation*, then we can solve the problem. This is the approach of *higher order statistics* (HOS).

4.2 Second Order Statistics (SOS) Approach

The second order statistics (SOS) approach exploits the second order nonstationary/colored structure of the sources, namely crosstalk minimization with additional nonstationary/colored information on the sources. Weinstein et al. [10] pointed out that nonstationary signals provide enough additional information to estimate the unmixing matrix $\mathbf{W}(\omega)$ and proposed a method based on nonstationary decorrelation. Some authors have used the SOS approach for mixed speech signals [4], [36].

This approach can be understood in a comprehensive way in that we have four unknown parameters W_{ij} in each frequency bin, but only three equations in (9) and (10) since $Y_1Y_2 = Y_2Y_1$ when $\Phi(Y_i) = Y_i$, that is, the simultaneous equations become underdetermined. Accordingly the simultaneous equations cannot be solved.

However, when the sources are nonstationary, the second order statistics is different in each time block. Similarly, when the sources are colored, the second order statistics is different in each time delay. As a result, more equations are available and the simultaneous equations can be solved.

In the nonstationary decorrelation approach, the source signals $S_1(\omega, m)$ and $S_2(\omega, m)$ are assumed to have a zero mean and be mutually uncorrelated. To determine the unmixing matrix $\mathbf{W}(\omega)$ so that $Y_1(\omega, m)$ and $Y_2(\omega, m)$ become mutually uncorrelated, we seek a $\mathbf{W}(\omega)$ that diagonalizes the covariance matrices $\mathbf{R}_Y(\omega, k)$ simultaneously for all time blocks k,

$$\begin{aligned} \mathbf{R}_Y(\omega, k) &= \mathbf{W}(\omega)\mathbf{R}_X(\omega, k)\mathbf{W}^H(\omega) \\ &= \mathbf{W}(\omega)\mathbf{H}(\omega)\mathbf{\Lambda}_s(\omega, k)\mathbf{H}^H(\omega)\mathbf{W}^H(\omega) \\ &= \mathbf{\Lambda}_c(\omega, k), \end{aligned} \tag{17}$$

where ^{H} denotes the conjugate transpose, $\mathbf{R}_X$ is the covariance matrix of $\mathbf{X}(\omega)$ as follows,

$$\mathbf{R}_X(\omega, k) = \frac{1}{M} \sum_{m=0}^{M-1} \mathbf{X}(\omega, Mk+m)\mathbf{X}^H(\omega, Mk+m), \tag{18}$$

$\mathbf{\Lambda}_s(\omega, k)$ is a covariant matrix of source signals that is a different diagonal matrix for each time block k, and $\mathbf{\Lambda}_c(\omega, k)$ is an arbitrary diagonal matrix.

The diagonalization of $\mathbf{R}_Y(\omega, k)$ can be written as an overdetermined least squares problem,

$$\begin{aligned} \arg \min_{W(\omega)} \sum_k & ||\mathrm{diag}\{\mathbf{W}(\omega)\mathbf{R}_X(\omega, k)\mathbf{W}^H(\omega)\} \\ & -\mathbf{W}(\omega)\mathbf{R}_X(\omega, k)\mathbf{W}^H(\omega)||^2 \\ s.t., \sum_k & ||\mathrm{diag}\{\mathbf{W}(\omega)\mathbf{R}_X(\omega, k)\mathbf{W}^H(\omega)\}||^2 \neq 0, \end{aligned} \tag{19}$$

where $||\mathbf{x}||^2$ is the squared Frobenius norm and $\mathrm{diag}\mathbf{A}$ is the diagonal components of the matrix $\mathbf{A}$. The solution can be found by the gradient method.

In the time-delayed decorrelation approach, $\mathbf{R}_X$ is defined as follows,

$$\mathbf{R}_X(\omega, \tau_i) = \frac{1}{M} \sum_{m=0}^{M-1} \mathbf{X}(\omega, m)\mathbf{X}^H(\omega, m+\tau_i), \tag{20}$$

and we seek a $\mathbf{W}(\omega)$ that diagonalizes the covariance matrices $\mathbf{R}_Y(\omega, \tau_i)$ simultaneously for all time delays τ_i.

4.3 Higher Order Statistics (HOS) Approach

The higher order statistics (HOS) approach exploits the nongaussian structure of the sources. Or more simply, we could say that we have four equations in (9) and (10) for four unknown parameters W_{ij} in each frequency bin. Accordingly the simultaneous equations can be solved. To calculate the unmixing matrix $\mathbf{W}(\omega)$, an algorithm has been proposed based on the minimization of the Kullback-Leibler divergence [14], [15]. For stable and faster convergence, Amari [37] proposed an algorithm based on the natural gradient. Using the natural gradient, the optimal unmixing matrix $\mathbf{W}(\omega)$ is obtained by using the following gradient iterative equation:

$$\mathbf{W}_{i+1}(\omega) = \mathbf{W}_i(\omega)$$

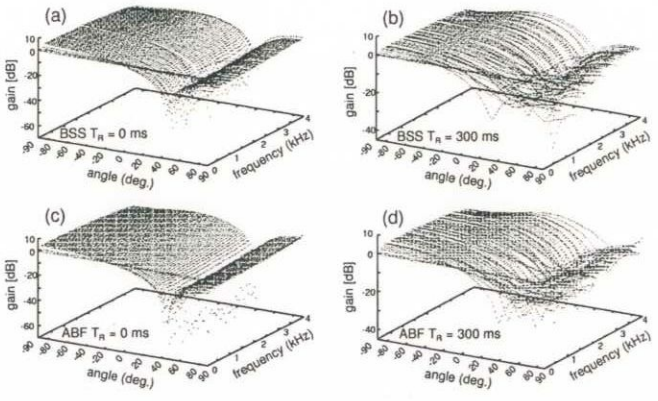

図 4 Directivity patterns (a) obtained by BSS (T_{R}=0 ms), (b) obtained by BSS (T_{R}=300 ms), (c) obtained by ABF (T_{R}=0 ms), and (d) obtained by ABF (T_{R}=300 ms).

$$+\eta\left[\mathrm{diag}\left(\langle\Phi(\mathbf{Y})\mathbf{Y}^{\mathrm{H}}\rangle\right)-\langle\Phi(\mathbf{Y})\mathbf{Y}^{\mathrm{H}}\rangle\right]\mathbf{W}_i(\omega), \quad (21)$$

where $\mathbf{Y}=\mathbf{Y}(\omega,m)$, $\langle\cdot\rangle$ denotes the averaging operator, i is used to express the value of the i-th step in the iterations, and η is the step size parameter. In addition, we define the nonlinear function $\Phi(\cdot)$ to signals with complex values as

$$\Phi(\mathbf{Y}) = \tanh(\mathbf{Y}^{(\mathrm{R})}) + j\ \tanh(\mathbf{Y}^{(\mathrm{I})}), \quad (22)$$

where $\mathbf{Y}^{(\mathrm{R})}$ and $\mathbf{Y}^{(\mathrm{I})}$ are the real part and the imaginary part of $\mathbf{Y}$, respectively [14].

For the complex numbered nonlinear function, the polar coordinate version

$$\Phi(\mathbf{Y}) = \tanh(\mathrm{abs}(\mathbf{Y}))\ e^{j\arg(\mathbf{Y})} \quad (23)$$

was shown to outperform the Cartesian coordinate version (22) both theoretically and experimentally [38].

5. Separation Mechanism of BSS

BSS is a statistical, or mathematical method, so the physical behavior of BSS is not obvious. We are simply attempting to make the two output signals Y_1 and Y_2 independent. Then, what is the physical interpretation of BSS?

We can understand the behavior of BSS as two sets of ABFs [39]. An ABF can create only one null towards the jammer signal when two microphones are used. BSS and ABFs form an adaptive spatial null in the jammer direction, and extract the target.

The separation performance of BSS is compared with that of ABF. Figure 4 shows the directivity patterns obtained by BSS and ABF. In Fig. 4, (a) and (b) show directivity patterns by $\mathbf{W}$ obtained by BSS, and (c) and (d) show directivity patterns by $\mathbf{W}$ obtained by ABF. When $T_{\mathrm{R}}=0$, a sharp spatial null is obtained by both BSS and ABF [see Figs. 4(a) and (c)]. When $T_{\mathrm{R}}=300$ ms, the directivity pattern becomes duller for both BSS and ABF [see Figs. 4(b) and (d)].

6. Conclusions

The blind source separation (BSS) of convolved mixtures of acoustic signals, especially speech, was examined. Source signals can be extracted only from observed mixed signals, by achieving nonlinear, nonstationary, or time-delayed decorrelation. The statistical technique of independent component analysis (ICA) was studied from the acoustic signal processing point of view.

BSS was interpreted from the physical standpoint showing the equivalence between frequency-domain BSS and two sets of microphone array systems, i.e., two sets of adaptive beamformers (ABFs). Convolutive BSS can be understood as multiple ABFs that generate statistically independent output, or more simply, an output with minimal crosstalk.

Because ABF and BSS mainly deal with sound from the jammer direction by making a null towards the jammer, the separation performance is fundamentally limited [40]. This understanding clearly explains the poor performance of BSS in the real world with long reverberation. If the sources are not "independent," their dependency results in bias noise to obtain the correct unmixing filter coefficients. Therefore, the BSS performance is upper bounded by that of the ABF.

However, in contrast to the ABF, no assumptions regarding array geometry or source location need to be made in BSS. BSS can adapt without any information on the source positions or period of source existence/absence. This is because, instead of adopting power minimization criterion that adapt the jammer signal out of the target signal in ABF, a cross-power minimization criterion is adopted that decorrelates the jammer signal from the target signal in BSS. It was shown that the least squares criterion of ABF is equivalent to the decorrelation criterion of the output in BSS. The error minimization was shown to be completely equivalent to a zero search in the cross-correlation.

Although the performance of the BSS is limited by that of the ABF, BSS has a major advantage over ABF. A strict one-channel power criterion has a serious crosstalk or leakage problem in ABF, whereas sources can be simultaneously active in BSS. Also, ABF needs to know the array manifold and the target direction. Thus, BSS can be regarded as an intelligent version of ABF.

Acknowledgment

We thank Shigeru Katagiri for his continuous encouragement.

文　献

[1] S. Makino, "Blind Source Separation of Convolutive Mixtures of Speech," in *Adaptive Signal Processing: Applications to Real-World Problems*, J. Benesty and Y. Huang, Eds., Springer, Berlin, Jan. 2003.

[2] J. F. Cardoso, "The three easy routes to independent com-

ponent analysis; contrasts and geometry," in *Proc. Conference Indep. Compon. Anal. Signal. Sep.*, Dec. 2001, pp. 1–6.

[3] T. W. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real world signals," *Neural Networks*, vol. 4, pp. 2129–2134, 1997.

[4] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proc. ICASSP*, June 2000, pp. 1041–1044.

[5] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," in *Proc. ICASSP*, May 2001, vol. 5, pp. 2737–2740.

[6] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," in *Proc. Eurospeech*, Sept. 2001, pp. 2595–2598.

[7] R. Mukai, S. Araki, and S. Makino, "Separation and dereverberation performance of frequency domain blind source separation for speech in a reverberant environment," in *Proc. Eurospeech*, Sept. 2001, pp. 2599–2602.

[8] S. C. Douglas, "Blind separation of acoustic signals," in *Microphone Arrays: Techniques and Applications*, M. Brandstein and D. B. Ward, Eds., pp. 355–380, Springer, Berlin, 2001.

[9] K. Torkkola, "Blind separation of delayed and convolved sources," in *Unsupervised Adaptive Filtering, Vol. I*, S. Haykin, Ed., pp. 321–375, John Wiley & Sons, 2000.

[10] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multichannel signal separation by decorrelation," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 405–413, Oct. 1993.

[11] T. W. Lee, *Independent Component Analysis -Theory and Applications*, Kluwer, 1998.

[12] M. Kawamoto, A. K. Barros, A. Mansour, K. Matsuoka, and N. Ohnishi, "Real world blind separation of convolved non-stationary signals," in *Proc. Workshop Indep. Compon. Anal. Signal. Sep.*, Jan. 1999, pp. 347–352.

[13] X. Sun and S. Douglas, "A natural gradient convolutive blind source separation algorithm for speech mixtures," in *Proc. Conference Indep. Compon. Anal. Signal. Sep.*, Dec. 2001, pp. 59–64.

[14] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[15] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proc. Workshop Indep. Compon. Anal. Signal. Sep.*, Jan. 1999, pp. 365–370.

[16] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Time domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming," in *Proc. NNSP*, Sept. 2002.

[17] J. Anemüeller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. Workshop Indep. Compon. Anal. Signal. Sep.*, 2000, pp. 215–220.

[18] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "A combined approach of array processing and independent component analysis for blind separation of acoustic signals," in *Proc. ICASSP*, May 2001, vol. 5, pp. 2729–2732.

[19] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural network models," in *Neural Networks for Computing: AIP Conference Proceedings 151*, J. S. Denker, Ed., American Institute of Physics, New York, 1986.

[20] C. Jutten and J. Herault, "Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[21] P. Comon, C. Jutten, and J. Herault, "Blind separation of sources, part II: problems statement," *Signal Processing*, vol. 24, pp. 11–20, 1991.

[22] E. Sorouchyari, "Blind separation of sources, part III: stability analysis," *Signal Processing*, vol. 24, pp. 21–29, 1991.

[23] A. Cichocki and L. Moszczynski, "A new learning algorithm for blind separation of sources," *Electronics Letters*, vol. 28, no. 21, pp. 1986–1987, 1992.

[24] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.

[25] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[26] A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. Circuits and Systems*, vol. 43, no. 11, pp. 894–906, 1996.

[27] T. W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski, "A unifying information-theoretic framework for independent component analysis," *Computers and Mathematics with Applications*, vol. 31, no. 11, pp. 1–12, Mar. 2000.

[28] A. Hyvärinen, H. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[29] S. Haykin, *Unsupervised Adaptive Filtering*, John Wiley & Sons, 2000.

[30] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley & Sons, 2002.

[31] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[32] S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind source separation," in *Advances in Neural Information Processing Systems 8*, pp. 757–763, MIT Press, 1996.

[33] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[34] L. Molgedey and H. G. Schuster, "Separation of a mixure of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3636, 1994.

[35] A. Belouchrani, K. A. Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation technique based on second order statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[36] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[37] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, 1998.

[38] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," in *Proc. ICASSP*, May 2002, vol. 1, pp. 1001–1004.

[39] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures," *EURASIP Journal on Applied Signal Processing*, accepted.

[40] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 2, pp. 109–116, Mar. 2003.