ICA-BASED BLIND SOURCE SEPARATION OF SOUNDS

Shoji Makino, Shoko Araki, Ryo Mukai, Hiroshi Sawada, and Hiroshi Saruwatari †

NTT Communication Science Laboratories, NTT Corporation, Kyoto Japan
† Nara Institute of Science and Technology, Nara Japan

ABSTRACT

This paper introduces blind source separation (BSS) of convolutive mixtures of acoustic signals, especially speech. The statistical and computational technique, called independent component analysis (ICA), is examined. By achieving nonlinear decorrelation, time-delayed decorrelation, or nonstationary 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 like ABF, they mainly remove the sounds from the jammer direction. This is the reason for the difficulty of BSS in the real world with long reverberation. If sources are not "independent," the dependence results in bias noise when getting 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 source positions or period of source existence/absence.

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 understand their conversation at a noisy cocktail party. This is the well known cocktail-party effect, where the individual speech waveforms are found from their mixtures. The aim of source separation is to give this cocktail party ability to computers. Then it will be possible to make computers understand what a person is saying at a noisy cocktail party.

Blind source separation (BSS) is an emerging technique, which enables extraction of target speech from observed mixed speech without the need for source positioning, spectral construction, or a mixing system. To achieve this, we focus on a method based on independent component analysis (ICA). ICA extracts independent sounds from among mixed sounds. There are a few applications of BSS to mixed speech signals in the real world [1], but the separation performance is still not good enough [2, 3].

Because 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. We have investigated the ICA method in detail, gradually uncovering its mechanisms by using theoretical analysis from the perspective of acoustic signal processing [4] as well as experimental analysis based on impulse response [5]. The mechanism of BSS based on ICA has been shown to be equivalent to that of an adaptive microphone array system, i.e., two sets of adaptive beamformers (ABFs) with an adaptive null directivity aimed in the direction of unnecessary sounds.

From the equivalence between BSS and ABF, we can make it clear that the physical behavior of BSS reduces the jammer signal by making a spatial null towards the jammer. We also found that BSS can 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.

BSS is applicable to the achievement of noise robust speech recognition and high-quality hands-free telecommunication.

2. What's BSS?

Blind source separation (BSS) is an approach to estimate source signals $s_i(t)$ using only the information of mixed signals $x_j(t)$ observed at each input channel. Typical examples of source signals are mixtures of simultaneous speech signals that have been picked up by several microphones, brain waves recorded by multiple sensors, interfering radio signals arriving at a mobile phone, etc.

The Model of Mixed Signals

In the case of audio source separation, several sensor microphones are put in different positions so that each records a mixture of the original source signals with 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 are affected by reverberation and observed by microphones. Therefore, the N signals recorded by

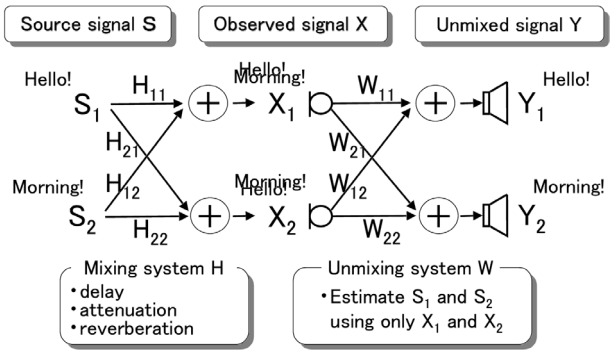

Figure 1: BSS system configuration.

M microphones are modeled as

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

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

The Model of Unmixed Signals

In order to obtain unmixed signals, we estimate unmixing filters $w_{ij}(k)$ of Q-taps, and the unmixed signals are obtained as:

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

The unmixing filters are estimated so that the unmixed signals become mutually independent. In this paper, we consider a two-input, two-output convolutive BSS problem, i.e., $N = M = 2$ (Fig. 1).

Task of Blind Source Separation

We assume that the source signals s_1 and s_2 are mutually independent. This assumption is usually true for sounds in the real world. There are two microphones which pick up the mixed speech. We only have the observed signals x_1 and x_2 which are dependent. Our goal is to adapt the unmixing system w_{ij}, and extract y_1 and y_2 so that they are mutually independent. With this operation, we can get s_1 and s_2 in the output y_1 and y_2. We do not need any information on the source positions or period of source existence/absence. Nor do we need any information on the mixing system h_{ji}. Thus, this task is called *blind* source separation.

Note that an unmixing system w_{ij} can at best be obtained up to a scaling and a permutation. Note also that BSS algorithm cannot solve the dereverberation/deconvolution problem in itself [6].

3. What's ICA?

Independent component analysis (ICA) is a statistical method and was originally introduced in the context of neural network modeling [7]. Recently, this method has been used for blind source separation (BSS) of sounds, fMRI and EEG signals of biomedical applications, wireless communication signals, images, etc.

In the theory of ICA, we use very general statistical properties: information on statistical independence. In a source separation problem, the source signals are the "independent components" of the data set. The problem of BSS is summarized to find a linear representation in which the components are mutually independent. ICA consists of estimating both the coefficient w_{ij} and sources s_i, when we only have the observed signals x_j.

The coefficient w_{ij} is determined so that the 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.

4. HOW CAN WE SEPARATE SOUNDS?

With the ICA-based BSS framework, how can we separate speech?
The simplified answer is to diagonalize $\boldsymbol{R}_Y$, where $\boldsymbol{R}_Y$ is a (2×2) matrix:

$$\boldsymbol{R}_Y = \begin{bmatrix} \langle \Phi(Y_1) Y_1 \rangle & \langle \Phi(Y_1) Y_2 \rangle \\ \langle \Phi(Y_2) Y_1 \rangle & \langle \Phi(Y_2) Y_2 \rangle \end{bmatrix}. \tag{3}$$

The function $\Phi(\cdot)$ is the activation function. The operation $\langle\cdot\rangle$ is the averaging operation to get statistical information. We want to minimize the off-diagonal components, while we want to constrain the diagonal components to proper constants.

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

$$\langle\Phi(Y_1)Y_2\rangle = 0 \qquad \langle\Phi(Y_2)Y_1\rangle = 0. \tag{4}$$

While the diagonal components, which relate to the average amplitude of Y_1 and Y_2, are constrained to proper constants:

$$\langle\Phi(Y_1)Y_1\rangle = c_1 \qquad \langle\Phi(Y_2)Y_2\rangle = c_2. \tag{5}$$

To achieve this convergence, we use the recursive algorithm

$$\boldsymbol{W}_{i+1} = \boldsymbol{W}_i + \eta\Delta\boldsymbol{W}_i, \tag{6}$$

$$\Delta\boldsymbol{W}_i = \begin{bmatrix} c_1 - \langle\Phi(Y_1)Y_1\rangle & \langle\Phi(Y_1)Y_2\rangle \\ \langle\Phi(Y_2)Y_1\rangle & c_2 - \langle\Phi(Y_2)Y_2\rangle \end{bmatrix}. \tag{7}$$

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

Second Order Statistics Approach

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

$$\langle\Phi(Y_1)Y_2\rangle = \langle Y_1Y_2\rangle = 0. \tag{8}$$

This is not sufficient to achieve independence. However, if we have nonstationary sources, we have this equation for multiple time blocks, thus we can solve the problem. This is the *nonstationary decorrelation* approach [6].

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

$$\langle\Phi(Y_1)Y_2\rangle = \langle Y_1(t)Y_2(t+\tau_i)\rangle = 0, \tag{9}$$

thus we can solve the problem. This is the *time-delayed decorrelation* (TDD) approach [8].

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

Higher Order Statistics Approach

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

$$\langle\Phi(Y_1)Y_2\rangle = \langle\tanh(Y_1)Y_2\rangle = 0. \tag{10}$$

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

$$\langle(Y_1 - \frac{Y_1^3}{3} + \frac{2Y_1^5}{15} - \frac{17Y_1^7}{315}...)\ Y_2\rangle = 0, \tag{11}$$

then we have higher order or nonlinear decorrelation, thus we can solve the problem. This is the approach of *higher order statistics* (HOS) [9].

5. SEPARATION MECHANISM OF BSS

We can understand the behavior of BSS as two sets of ABFs [10]. An adaptive beamformer can create only one null towards the jammer signal in the case of two microphones. BSS and ABF form an adaptive spatial null to the jammer direction, and extract the target.

We compared the separation performance of BSS with that of ABF. Figure 2 shows the directivity patterns obtained by BSS and ABF. In Fig. 2, (a) and (b) show directivity patterns by $\boldsymbol{W}$ obtained by BSS, and (c) and (d) show directivity patterns by $\boldsymbol{W}$ obtained by ABF. When $T_{\mathrm{R}} = 0$, a sharp spatial null is obtained by both BSS and ABF [see Figs. 2(a) and (c)]. When $T_{\mathrm{R}} = 300$ ms, the directivity pattern becomes duller for both BSS and ABF [see Figs. 2(b) and (d)].

6. CONCLUSIONS

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

We gave an interpretation of BSS from the physical point of view 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, output with minimal crosstalk.

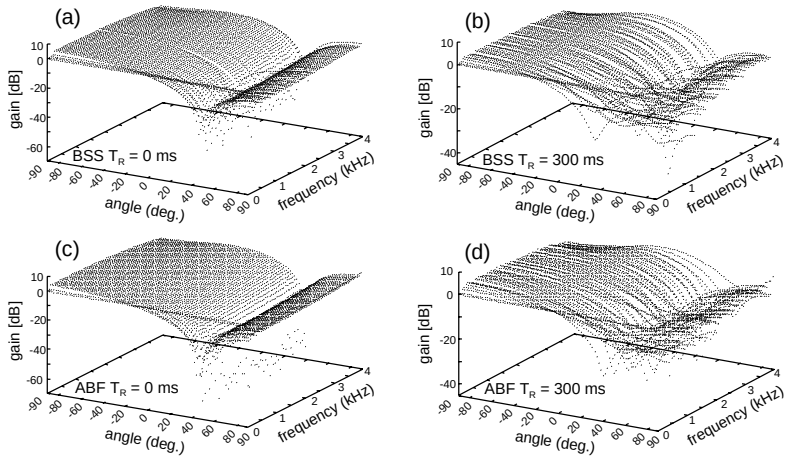

Figure 2: Directivity patterns (a) obtained by BSS (T_{R}=0 ms), (b) obtained by BSS (T_{R}=300 ms), (c) obtained by ABF (T_{R}=0 ms), and (d) obtained by ABF (T_{R}=300 ms).

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. 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 get the correct unmixing filter coefficients. Therefore, the performance of BSS is upper bounded by that of ABF.

However, as opposed to ABF, no assumptions on 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 criteria that adapt the jammer signal out of the target signal, we adopt cross-power minimization criteria that decorrelate the jammer signal from the target signal. We showed 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 with a zero search in the crosscorrelation.

Although BSS is upper bounded by ABF, BSS has a strong advantage over ABF. Strict one-channel power criteria have 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.

The fusion of acoustic signal processing technologies and speech recognition technologies is playing a major role in the development of user-friendly communication with computers, conversation robots, and other advanced audio media processing technologies.

ACKNOWLEDGMENTS

We would like to thank Dr. Shigeru Katagiri for his continuous encouragement.

References

[1] T. W. Lee, *Independent component analysis -Theory and applications*, Kluwer, 1998.

[2] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proc. ICASSP2000*, June 2000, pp. 1041–1044.

[3] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. ICASSP2000*, June 2000, pp. 3140–3143.

[4] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," in *Proc. Eurospeech2001*, Sept. 2001, pp. 2595–2598.

[5] 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. Eurospeech2001*, Sept. 2001, pp. 2599–2602.

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

[7] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural network models," in *Neural networks for computing: AIP conference proceedings 151*, New York J. S. Denker, ed., American Institute of Physics, Ed., 1986.

[8] 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.

[9] A. Hyvarinen, H. Karhunen, and E. Oja, *Independent component analysis*, John Wiley & Sons, 2001.

[10] S. Araki, S. Makino, R. Mukai, Y. Hinamoto, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming," in *Proc. ICASSP2002*, May 2002, vol. 2, pp. 1785–1788.