AUDIO SOURCE SEPARATION
BASED ON INDEPENDENT COMPONENT ANALYSIS

Shoji Makino † *Shoko Araki* † *Ryo Mukai* † *Hiroshi Sawada* †

† NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{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.

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 audio 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 sense, namely nonlinear decorrelation together with nonstationary decorrelation and time-delayed decorrelation. These three methods are discussed in a unified manner [1, 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].

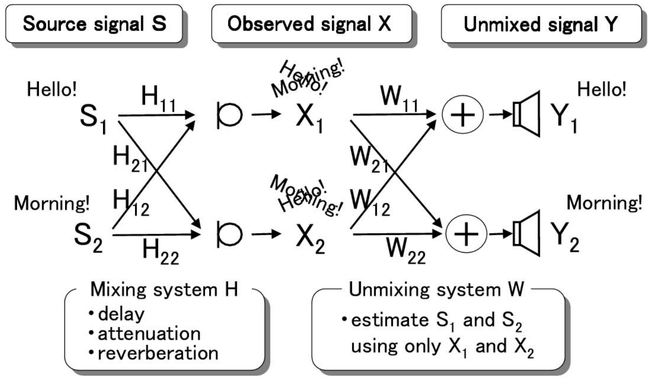


Figure 1: BSS system configuration.

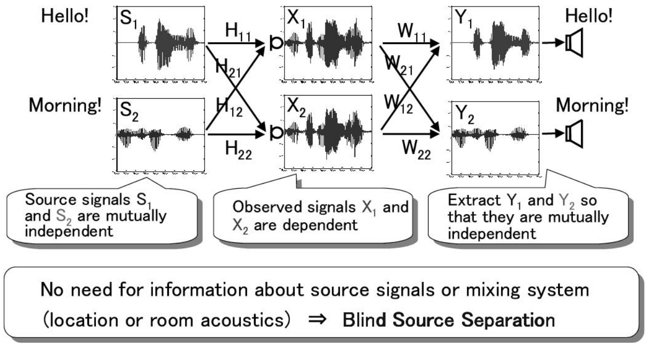
Since ICA is a purely statistical process, the separation mechanism has not been clearly understood in the sense of acoustic signal processing, and it has been difficult to know which components were separated, and to what degree. Recently, the ICA method has been investigated in detail, and its mechanisms 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.

0-7803-8251-X/04/$17.00 ©2004 IEEE



Source signals S_1 and S_2 are mutually independent

Observed signals X_1 and X_2 are dependent

Extract Y_1 and Y_2 so that they are mutually independent

No need for information about source signals or mixing system (location or room acoustics) ⇒ Blind Source Separation

Figure 2: Task of blind source separation of speech signals.

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. 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, nonwhite, 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 [8]. A robust and precise method for solving the permutation problem of frequency-domain BSS was proposed in [9], and a minimal distortion principle for solving the scaling problem was proposed in [10].

3. WHAT IS ICA?

Independent component analysis (ICA) is a statistical method that was originally introduced in the context of neural network modeling [11]. 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.

4. HOW SPEECH SIGNALS CAN BE SEPARATED?

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

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

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

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:

$$\langle \Phi(Y_1)Y_2 \rangle = 0, \quad \langle \Phi(Y_2)Y_1 \rangle = 0. \quad (4)$$

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:

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

To achieve this convergence, we use the recursive learning rule [12, 13].

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

$$\Delta \mathbf{W}_i = \left[\begin{array}{cc} c_1 - \langle \Phi(Y_1)Y_1 \rangle & \langle \Phi(Y_1)Y_2 \rangle \\ \langle \Phi(Y_2)Y_1 \rangle & c_2 - \langle \Phi(Y_2)Y_2 \rangle \end{array} \right] \mathbf{W}_i. \tag{7}$$

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

4.1. Second Order Statistics (SOS) Approach

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

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

This is not sufficient to achieve independence, therefore, we cannot solve the problem. This 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 (4) and (5) 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. As a result, more equations are available and the simultaneous equations can be solved. This is the *nonstationary decorrelation* approach [14].

Similarly, when the sources are nonwhite, we have a delayed correlation for a multiple time delay:

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

The second order statistics is different in each time delay, thus more equations are available and the simultaneous equations can be solved. This is the *time-delayed decorrelation* (TDD) approach [15].

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

4.2. Higher Order Statistics (HOS) Approach

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

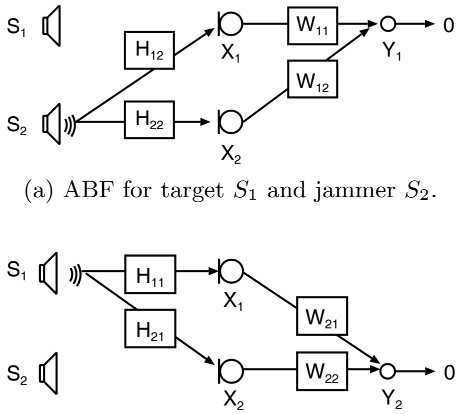
$$\langle \Phi(Y_1)Y_2 \rangle = \langle \tanh(Y_1)Y_2 \rangle = 0. \tag{10}$$

With a Tailor expansion of $\tanh(\cdot)$, (10) can be expressed as

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

thus we have higher order or *nonlinear decorrelation*, then we can solve the problem. Or more simply, we could say that we have four equations in (4) and (5) for four unknown parameters W_{ij} in each frequency bin. Accordingly the simultaneous equations can be solved.

This is the approach of *higher order statistics* (HOS) [16].



(a) ABF for target S_1 and jammer S_2.

(b) ABF for target S_2 and jammer S_1.

Figure 3: Two sets of ABF-system configurations.

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 [6]. An ABF can create only one null towards the jammer when two microphones are used. BSS and ABFs form an adaptive spatial null in the jammer direction, and extract the target.

5.1. Frequency-Domain Adaptive Beamformer (ABF)

Here, we consider the frequency-domain adaptive beamformer (ABF), that can adaptively remove a jammer signal. Since the aim is to separate two signals S_1 and S_2 with two microphones, two sets of ABFs are used (see Fig. 3). That is, an ABF that forms a null directivity pattern towards source S_2 by using filter coefficients W_{11} and W_{12}, and an ABF that forms a null directivity pattern towards source S_1 by using filter coefficients W_{21} and W_{22}. Note that the direction of the target or the impulse responses from the target to the microphones should be known, and that the ABF can adapt only when a jammer is active but a target is silent.

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. DISCUSSIONS

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) [6]. Convolutive BSS can be understood as multiple ABFs that generate statistically inde-

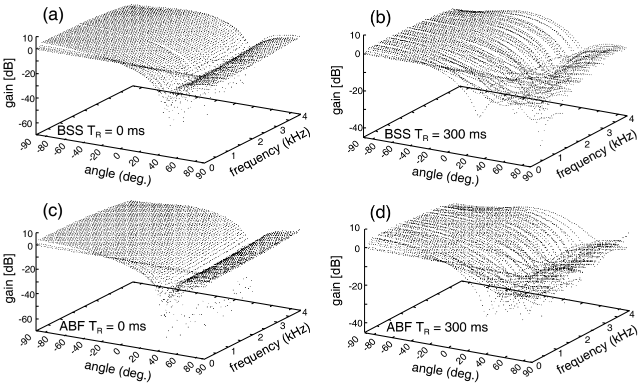


Figure 4: Directivity patterns (a) obtained by BSS (T_{R}=0 ms), (b) obtained by BSS (T_{R}=300 ms), (c) obtained by ABF (T_{R}=0 ms), and (d) obtained by ABF (T_{R}=300 ms).

pendent 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 [5]. 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.

7. 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.

ACKNOWLEDGMENTS

We thank Dr. Hiroshi Saruwatari for detailed discussion.

REFERENCES

[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 component analysis; contrasts and geometry," in *Proc. ICA*, 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. ICASSP2000*, June 2000, pp. 1041–1044.

[5] 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.

[6] 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.

[7] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Separation and dereverberation performance evaluation of frequency domain blind source separation," *Journal on Acoust. Sci. & Tech.*, accepted.

[8] 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.

[9] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust approach to the permutation problem of frequency-domain blind source separation," in *Proc. ICASSP*, Apr. 2003, pp. 381–384.

[10] K. Matsuoka and S. Nakashima, " Minimal distortion principle for blind source separation," in *Proc. ICA*, Dec. 2001, pp. 722–727.

[11] 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.

[12] 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.

[13] A. Cichocki, R. Unbehauen, and E. Rummert, "Robust learning algorithm for blind separation of signals," *Electronics Letters*, vol. 30, no. 17, pp. 1386–1387, 1994.

[14] 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.

[15] 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.

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