International Workshop on Acoustic Echo and Noise Control (IWAENC2003), Sept. 2003, Kyoto, Japan

Blind Separation for Convolutive Mixture of Many Voices

K. Matsuoka, Y. Ohba, Y. Toyota, and S. Nakashima

Kyushu Institute of Technology

Hibikino, Wakamatsu-ku, Kitakyushu, Japan

matsuoka@brain.kyutech.ac.jp

Abstract: An algorithm of blind source separation is applied to mixture of eight voices taken in an ordinary room. As far as we know, blind separation of such many sounds is the first attempt in the world. The algorithm, which was proposed by some of the authors, has proved to work very well for such a difficult task.
Keywords: blind source separation, independent component analysis, convolutive mixture, voice.

1. Introduction

Many attempts have been made to perform blind separation (BSS) of mixed voice signals, using independent component analysis (ICA). As far as we know, however, every experiment reported until now deals only with data taken in a very limited situation.

Most of them treat artificially mixed sounds on a computer and assume rather simple mixing processes. In a real situation, however, since echo effect cannot be neglected, the mixing process has a very long time lag; the reverberation time is as many as a hundred milliseconds. It implies that, if we implement the separator with an FIR filter, we need around one thousand taps when the sampling rate is 10 kHz.

Even in the reports dealing with 'real' data, the number of sound sources is usually two. It is very doubtful that the algorithms employed there work as well for a larger number of sound sources. This paper reports a challenge to a much more difficult task; blind separation of eight sounds acquired in an ordinary office room.

The convolutive ICA algorithm used here is proposed by some of the authors [3]. It employs a special principle (Minimal Distortion Principle) to eliminate filtering indeterminacy inherent in the problem in ICA.

2. The BSS Algorithm

BSS is a method for recovering a set of source signals from the observation of their mixtures without any prior knowledge about the mixing process. In view of the level of complexity, the mixing process can be classified into two types: instantaneous mixture and convolutive mixture. For separation of sounds the mixing process must be considered to be convolutive, of course.

Inherently BSS has two kinds of indeterminacy. One is the indeterminacy in the numbering of the sources and the other is that in the scaling or filtering. The latter indeterminacy is more essential and will be focused on in this section. In the case of instantaneous mixture the indeterminacy is usually considered unsubstantial, but in the case of convolutive mixture it cannot be overlooked in view of actual implementations and applications of BSS. This section addresses an idea for normalizing the separator, which is an optimal separator in a particular sense.

2.1 The mixing process and the demixing process

Let us consider a situation where statistically independent random signals $s_i(t)$ (i = 1,…, N) are generated by N sources and their mixtures are observed by N sensors. It is assumed that every source signal $s_i(t)$ is a stationary random process with zero mean, and the sensors' outputs $x_i(t)$ (i = 1,…, N) are given by a linear mixing process:

$$\mathbf{x}(t) = \sum_{\tau=0}^{\infty} \mathbf{A}_\tau \mathbf{s}(t-\tau) = \mathbf{A}(z)\mathbf{s}(t)\,. \qquad (1)$$

It is known that, in order to realize BSS, at most one source signal is allowed to be Gaussian.

To recover the source signals from the sensor signals, we consider a demixing process or a separator of the form

$$\mathbf{y}(t) = \sum_{\tau=-\infty}^{\infty} \mathbf{W}_\tau \mathbf{x}(t-\tau) = \mathbf{W}(z)\mathbf{x}(t)\,. \qquad (2)$$

If the mixing process $\mathbf{A}(z)$ is known beforehand, the source signals can be recovered by setting as $\mathbf{W}(z) = \mathbf{A}^{-1}(z)$, of course. Essential difficulty in BSS is that $\mathbf{A}(z)$ or $\mathbf{A}^{-1}(z)$ must be estimated from the observed data $\mathbf{x}(t)$ only. Besides, the impulse response $\{\mathbf{W}_\tau\}$ might need to take a noncausal form in general, i.e., $\mathbf{W}_\tau \neq \mathbf{O}$ ($\tau < 0$).

2.2 Minimal distortion principle

In BSS the definition of the source signals has an indeterminacy. Namely, if $s_1(t)$,…, $s_N(t)$ are source signals, their arbitrarily linear-filtered signals $e_1(z)s_1(t)$, …, $e_N(z)s_N(t)$ can also be considered source signals because they are also mutually independent. The mixing process is then $\mathbf{A}(z)\mathrm{diag}\{e_1^{-1}(z),\ldots,e_N^{-1}(z)\}$.

We call a separator of the following form a valid separator:

$$\mathbf{W}(z) = \mathbf{D}(z)\mathbf{A}^{-1}(z), \qquad (3)$$

where $\mathbf{D}(z)$ is an arbitrary nonsingular diagonal matrix; $\mathbf{D}(z) = \mathrm{diag}\{d_i(z)\}$. If the separator is valid, each of the source signals appears at an output terminal of the separator, though it is subjected to a linear transformation $d_i(z)$. [More generally we should define a valid separator as $\mathbf{W}(z) = \mathbf{P}\mathbf{D}(z)\mathbf{A}^{-1}(z)$, where **P** is a permutation matrix, but we consider only the case of **P** = **I** to make the description below simple.]

In BSS, all the valid separators are usually considered essentially equivalent. However the following separator has a special meaning:

$$\mathbf{W}^*(z) \triangleq \mathrm{diag}\,\mathbf{A}(z)\cdot\mathbf{A}^{-1}(z) \qquad (4)$$

We call this separator the optimal (valid) separator. The optimal separator $\mathbf{W}^*(z)$ can be characterized by either of the following two propositions.

Proposition 1: The optimal separator $\mathbf{W}^*(z)$ is a valid separator that minimizes $\|\mathbf{W}(z)\mathbf{A}(z) - \mathbf{A}(z)\|^2$.

Here, the norm of transfer function matrix $\mathbf{X}(z)$ is defined as $\|\mathbf{X}(z)\| \triangleq \left(\sum_{k=-\infty}^{\infty}\|\mathbf{X}_k\|^2\right)^{1/2}$.

Proposition 2: The optimal separator $\mathbf{W}^*(z)$ is a valid separator that minimizes $E\left[\|\mathbf{y}(t)-\mathbf{x}(t)\|^2\right]$.

These two propositions state the minimal distortion principle in two manners. Namely, the optimal separator is determined such that the overall transfer function $\mathbf{W}(z)\mathbf{A}(z)$ be as close to $\mathbf{A}(z)$ as possible, or equivalently the separator's output $\mathbf{y}(t)$ be as close to $\mathbf{x}(t)$ as possible. The optimal separator is 'optimal' in the sense that the separator's output is the least subjected to distortion among the set of all the valid separators.

The optimal separator has some properties that are favorable in actual implementation of BSS:

(i) The separator's output then becomes

$$\mathbf{y}(t) = \mathrm{diag}\mathbf{A}(z)\cdot\mathbf{s}(t). \qquad (4)$$

This implies that output $y_i(t)$ is $a_{ii}(z)s_i(t)$, which is the *i*-th source that would be observed at the *i*-th sensor when there were no other source signals. This property is very natural and convenient particularly for separation of voice signals.

(ii) The optimal separator does not depend on the properties of the sources; it depends on the mixing process $\mathbf{A}(z)$ only. So, even for such nonstationary signals as voices, the optimal separator is invariant with time as long as the mixing process is fixed. This property helps to enhance the stability of the algorithm, compared to the one proposed in [1].

The optimal separator can also be characterized as a direct constraint on matrix $\mathbf{W}(z)$.

Proposition 3: The optimal separator $\mathbf{W}^*(z)$ is a valid separator that satisfies

$$\mathrm{diag}\,\mathbf{W}^{-1}(z) = \mathbf{I}. \qquad (5)$$

Including the pioneering work by Herault and Jutten some studies on BSS have considered a separator of feedback structure:

$$\mathbf{y}(t) = \mathbf{x}(t) - \bar{\mathbf{W}}(z)\mathbf{y}(t), \qquad (6)$$

where $\bar{\mathbf{W}}(z)$ is a matrix whose diagonal elements are all zeros. This is equivalent to putting $\mathbf{W}(z) = \left(\mathbf{I} + \bar{\mathbf{W}}(z)\right)^{-1}$ in a feedforward-type separator, leading to $diag\,\mathbf{W}^{-1}(z) = \mathbf{I}$. So, the present normalization itself is not a new idea. What we want to stress is that the constraint (5) can be derived from the minimal distortion principle (Propositions 1 and 2). It is hard to design a feedback-type separator so as to guarantee its stability, particularly for non-minimum phase mixing processes. Using the following proposition, we can incorporate the constraint (5) easily in a multi-dimensional FIR filter, which is guaranteed to be stable.

Proposition 4: The optimal separator is a valid separator that satisfies

$$\mathrm{diag}\,E\left[\left(\mathbf{y}(t)-\mathbf{x}(t)\right)\mathbf{y}^T(t-\tau)\right] = \mathbf{0} \qquad (7)$$

for every τ.

As shown in the next subsection, the last proposition is the most important for implementation of MDP.

2. 3 An implementation of the minimal distortion principle

Here, we want to show how the proposed principle is implemented. We start with an approach proposed by Amari et al. [1]. Define

$$I(\mathbf{W}(z)) \triangleq -\sum_{i=1}^{N} E[\log q_i(y_i(t))] - h[\mathbf{y}(t)], \quad (8)$$

where $h[\mathbf{y}(t)]$ is the entropy *rate* of $\mathbf{y}(t)$ and $q_i(u)$ is a pdf assumed for source signal $s_i(t)$. If the source signals are iid (or liner processes in general) and $q_i(u)$ approximates well the real pdf of $s_i(t)$, then minimizing $I(\mathbf{W}(z))$ provides a valid solution. In actual computation, however, the separator must be embodied by a FIR filter as $\mathbf{W}(z) \triangleq \sum_{\tau=-L_1}^{L_2} \mathbf{W}_\tau z^{-\tau}$. The minimization is then performed by the following iterative calculation (natural gradient learning):

$$\mathbf{y}(t-L_1) = \sum_{\tau=-L_1}^{L_2} \mathbf{W}_\tau(t)\mathbf{x}(t-L_1-\tau),$$

$$\Delta\mathbf{W}_\tau(t) = \alpha_\tau \left\{ \mathbf{W}_\tau(t) - \varphi(\mathbf{y}(t-L_3)) \sum_{r=-L_1}^{L_2} \mathbf{y}^T(t-L_3-\tau+r)\mathbf{W}_r(t) \right\}, \quad (9)$$

where $L_3 \triangleq 2L_1 + L_2$, $\varphi(\mathbf{y}(t)) \triangleq [\varphi_1(y_1(t)),...,\varphi_N(y_N(t))]^T$ and φ_i is defined as $\varphi_i(u) \triangleq -d\log q_i(u)/du$. α is a small positive constant. Note that the time shift of L_1 and L_3 has been introduced in consideration of the available data at time *t*.

We here introduce the idea of nonholonomic constraint $\mathrm{diag}\,\Delta\mathbf{W}(z)\mathbf{W}^{-1}(z) = \mathbf{O}$ [2], leading to

$$\Delta\mathbf{W}_\tau(t) = -\alpha \sum_{r=-L_1}^{L_2} \left\{\text{off-diag}\,\varphi(\mathbf{y}(t-L_3))\mathbf{y}^T(t-L_3-\tau+r)\right\}\mathbf{W}_r(t) \quad (10)$$

According to this algorithm, each output $y_i(t)$ of the separator becomes indeterminate with respect to linear transformation. Now we incorporate the minimal distortion principle (Proposition 2) into this system. We superimpose the (natural) gradient of $E\left[\|\mathbf{y}(t)-\mathbf{x}(t)\|^2\right]$ (or Proposition 4) to (15) as

$$\Delta\mathbf{W}_\tau(t) = -\alpha_\tau \sum_{r=-L_1}^{L_2} \{\text{off-diag}\,\varphi(\mathbf{y}(t-L_3))\mathbf{y}^T(t-L_3-\tau+r) + \beta\,\mathrm{diag}\left(\mathbf{y}(t-L_3)-\mathbf{x}(t-L_3)\right)\mathbf{y}^T(t-L_3-\tau+r)\}\mathbf{W}_r(t) \quad (11)$$

This algorithm gives the desired separator, independently of the initial condition of $\mathbf{W}(z)$.

In the above algorithm the computation time increases in proportion to L_3^2. With a slight modification, we can reduce the time considerably. The following algorithm allow for the computation time of order L_3.

$$\mathbf{u}(t-L_0) = \sum_{r=-L_1}^{L_2} \mathbf{W}_r^T(t)\mathbf{y}(t-L_0+r) \quad (12)$$

$$\mathbf{V}(t-L_0) = \sum_{r=-L_1}^{L_2} \mathrm{diag}\,\mathbf{y}(t-L_0+r)\cdot\mathbf{W}_r(t) \quad (13)$$

$$\Delta\mathbf{W}_\tau(t) = -\alpha_\tau \{\ \varphi(\mathbf{y}(t-L_3))\mathbf{u}^T(t-L_3-\tau) - \mathrm{diag}\,\varphi(\mathbf{y}(t-L_3))\cdot\mathbf{V}(t-L_3-\tau) + \beta\,\mathrm{diag}\left(\mathbf{y}(t-L_3)-\mathbf{x}(t-L_3)\right)\cdot\mathbf{V}(t-L_3-\tau)\ \} \quad (14)$$

where $L_0 \triangleq L_1 + L_2$.

3. An Experimental Result

The setup of an experiment is shown in Fig. 1. The source sounds were eight voices of a woman which were provided by eight loudspeakers. The reverberation time (the time for the sound intensity to decay by 60 dB) of the room was around 140 ms.

We applied the proposed algorithm to sound signals taken by 8 microphones at 10kHz sampling frequency. Since the length of the filter was 801 ($L_1 = 200$, $L_2 = 600$), totally 64 x 801 parameters had to be estimated to obtain a desired separator. Independent components were considered super-Gaussian, and φ_i is chosen as $\varphi_i(u) = \mathrm{sgn}(u)$. The learning coefficient α was set as $\alpha = 2\text{x}10^{-6}$ and parameter β was set as $\beta = 0.01$.

Fig. 2 shows the impulse responses of the mixing process $\mathbf{A}(z)$ and the overall process $\mathbf{W}(z)\mathbf{A}(z)$. From one can see a desired separator was successfully obtained. The recovered sounds were very clear, which will be shown at the Workshop.

4. Concluding Remarks

The proposed algorithm has proved to be very effective for BSS for convolutive mixture of many voices. A remaining serious issue is that it takes a very long time to complete the calculation; some hours for a 30-second data. Hardware implementation is a future work.

References

[1] S. Amari, S. C. Douglas, A. Cichocki and H. H.Yang, "Multichannel blind deconvolution and equalization using the natural gradient", Proc. IEEE International Workshop on Wireless Communication, pp. 101-104, 1997.

[2] S. Choi, S. Amari, A. Cichocki, and R. Liu, Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels, Proc. ICA99, pp. 371-376, 1999.

[3] K. Matsuoka and S. Nakashima, Minimal distortion principle for blind source separation. Proc. ICA2001, 722-727, 2001

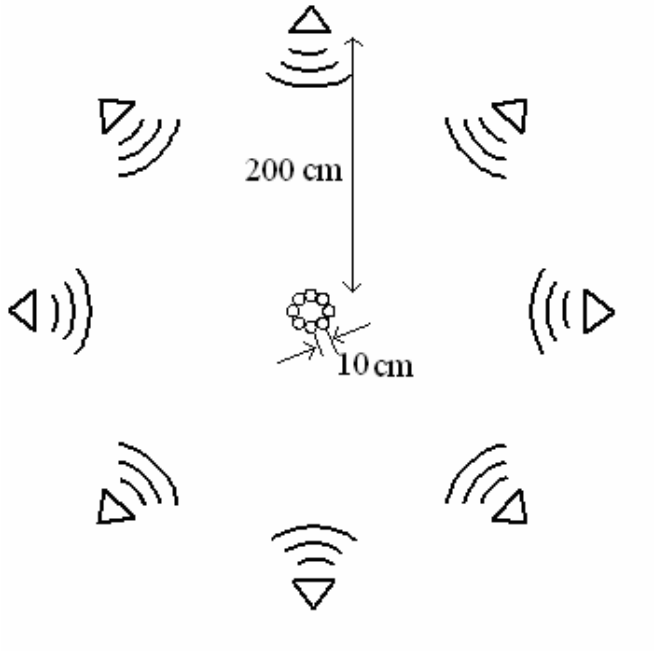
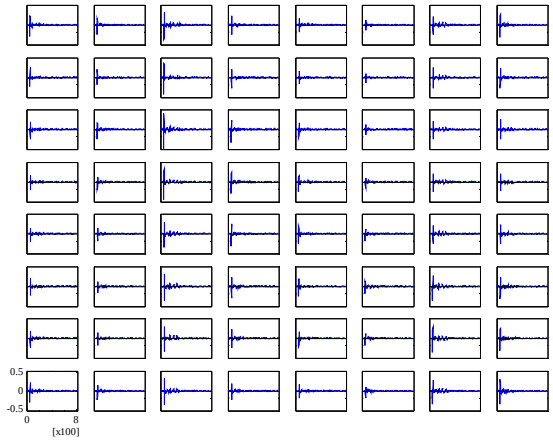
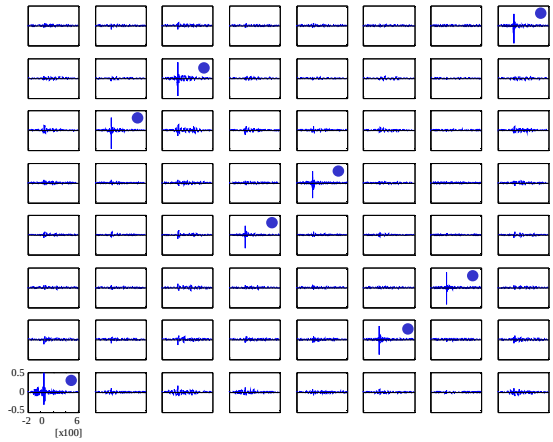


Fig. 1 A configuration of the loudspeakers and the microphones.



(a)



(b)

Fig. 2 The impulse responses of (a) the mixing process **A**(z) and (b) the overall process **W**(z)**A**(z).