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

Independent Component Analysis and Its Applications to Sound Signal Separation

K. Matsuoka
Kyushu Institute of Technology
Hibikino, Wakamatsu-ku, Kitakyushu, Japan
matsuoka@brain.kyutech.ac.jp

1. Introduction

Independent component analysis (ICA) or blind source separation (BSS) is a method for recovering a set of statistically independent signals from the observation of their mixtures without any prior knowledge about the mixing process. It has been receiving a great deal of attention from various fields as a new signal processing technique. In this talk I would like to focus on an approach to BSS for sound signals.

Blind separation of sound signals has a lot of (potential) applications: voice control of personal computers, noise canceling in vehicles, robots' ears, hand-free telephone systems, hearing aid instruments, sound source tracking, etc. In order to realize these applications we need to pay attentions to some particular aspects which are characteristic of sounds signal processing.

(1) On-line algorithm

In view of the level of complexity, the mixing process can be classified into two types: instantaneous mixture and convolutive mixture. While early works for BSS dealt with the former type, recent works are mainly concerned with the latter type, which is much more difficult from theoretical as well as computational points of view. In the case of sound separation the mixing process must be deal with as a convolutive one, of course.

An approach to BSS for convolutive mixture is to transform raw sound data into a set of frequency components, using a filter bank, and to apply some instantaneous BSS method to each component. However, such methods are not suitable for on-line processing, which is requisite of realistic applications of sound separation. Moreover, those methods suffer from the problem of permutation. In this sense the time-domain algorithms are more convenient than the frequency-domain ones. The algorithm I would like to address in this talk adopts a completely time-domain approach.

(2) Preservation of sound quality

Inherently BSS has an indeterminacy. Given a data set produced by a mixing process, since any linear transform of a source signal can also be considered a source signal, there exist an infinite number of valid separators that extract the source signals. In the case of instantaneous mixture the indeterminacy it is not so a serious problem because it is just a scaling problem. However in the case of convolutive mixture it cannot be overlooked because indeterminacy increases up to filtering indeterminacy. Particularly in the case of sound separation, it is concerned with quality of separated sounds.

In the next section I would like to address a particular type of normalization principle for the separator, which we call minimal distortion principle (MDP). Among the set of valid separators the principle chooses the separator such that its output be the least subject to distortion. Separators based on this principle have some favorable features, particularly for sound separation. First, separation can be attained without inducing the temporal whitening of the observed signals. Second, the obtained separator is free of fluctuation even for such a nonstationary signal as speech.

(3) Robustness

In the authors' experience, although most conventional methods for BSS are able to achieve separation for artificially synthesized data, they do not necessarily work well for real-world data. The results of separation are often unsatisfactory and, what is worse, they sometimes suffer from incomprehensible computational instability. We are often faced with the following phenomenon: when applying an iterative ICA algorithm to a data set, the algorithm appears to behave in a desired manner in the beginning of the iterative calculation, but suddenly some instability occurs.

Although a lot of reasons are conceivable, in section 3 we discuss the instability induced by the singularity of the mixing matrix, which is often neglected but occurs frequently. Namely, we consider the case that the frequency response of the mixing process becomes almost singular at some part of frequency range. The task of ICA is basically to find the inverse of the mixing matrix and to apply it to the observation. So, if the mixing matrix is nearly singular in some frequency ranges, then the norm of the demixing matrix becomes very large and some numerical instability can occur. Even if the inverse has been obtained successfully, the separator obtained becomes too sensitive to the noise that contains the frequency components for which the norm of the demixing matrix is very large.

In blind separation of speech signals, if the microphones are located close to each other, the mixing matrix, i.e. the transfer function matrix from the speakers to the microphones, becomes almost singular, particularly for low frequencies. Conventional methods for ICA usually neglect such a problem.

(4) Actual implementation

To realize BSS for real-world data, many other aspects must be considered such as the computational cost. Also we need to take into account the case that the number of the sources is not constant and unknown.

2. Minimal Distortion Principle

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 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., $\mathbf{W}_\tau \neq \mathbf{O}$ ($\tau < 0$).

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 is considered here. 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 $\mathbf{A}(z)\mathrm{diag}\{e_1^{-1}(z),\dots,e_N^{-1}(z)\}$.

Due to this indeterminacy we can consider any 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 $d_i(z)$.

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. In the set of valid separators 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 can be characterized by either of the following two propositions.

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

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 subject to distortion among 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 $\mathbf{y}(t) = \mathrm{diag}\mathbf{A}(z) \cdot \mathbf{A}^{-1}(z)\mathbf{A}(z)\mathbf{s}(t) = \mathrm{diag}\mathbf{A}(z) \cdot \mathbf{s}(t)$. 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 will be convenient for interpretation of the signals separated and later processing.

(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].

(iii) In actual implementation the separator is usually embodied with an FIR filter. Then, it is desirable that the filter's degree (length) is as low as possible. MDP determines a valid separator such that its output becomes as close to the input as possible. So, it can be expected that the (FIR) separator will be realized with a relatively short filter length though not the shortest.

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

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

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

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

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.

We have further

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

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

for every τ.

This characterization of MDP is important for actual implementation of MDP.

In actual implementation of MDP the separator needs to be embodied by a FIR filter as $\mathbf{W}(z) \triangleq \sum_{\tau=-L_1}^{L_2} \mathbf{W}_\tau z^{-\tau}$. Then the output of the separator is given by $\mathbf{y}(t-L_1) \triangleq \sum_{\tau=-L_1}^{L_2} \mathbf{W}_\tau \mathbf{x}(t-L_1-\tau)$. The algorithm derived by MDP becomes

$$\begin{aligned}\Delta\mathbf{W}_\tau(t) = \\ -\alpha_\tau \sum_{r=-L_1}^{L_2} \{&\text{off-diag}\,\varphi(\mathbf{y}(t-L_3))\mathbf{y}^T(t-L_3-\tau+r) \\ &+ \beta \operatorname{diag}\left(\mathbf{y}(t-L_3) - \mathbf{x}(t-L_3)\right)\mathbf{y}^T(t-L_3-\tau+r)\}\mathbf{W}_r(t)\end{aligned} \tag{8}$$

where $L_3 = 2L_1 + L_2$. The first term of the equation comes from the algorithm in [2] and the second term attains MDP.

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

$$\mathbf{u}(t-L_0) = \sum_{r=-L_1}^{L_2} \mathbf{W}_r^T(t)\mathbf{y}(t-L_0+r) \tag{9}$$

$$\mathbf{V}(t-L_0) = \sum_{r=-L_1}^{L_2} \operatorname{diag} \mathbf{y}(t-L_0+r)\cdot\mathbf{W}_r(t) \tag{10}$$

$$\begin{aligned}\Delta\mathbf{W}_\tau(t) = -\alpha_\tau \{\ &\varphi(\mathbf{y}(t-L_3))\mathbf{u}^T(t-L_3-\tau) \\ &- \operatorname{diag}\varphi(\mathbf{y}(t-L_3))\cdot\mathbf{V}(t-L_3-\tau) \\ &+ \beta \operatorname{diag}\left(\mathbf{y}(t-L_3) - \mathbf{x}(t-L_3)\right)\cdot\mathbf{V}(t-L_3-\tau)\ \}\end{aligned} \tag{11}$$

where $L_0 \triangleq L_1 + L_2$.

3. ε-Minimal Distortion Principle

Although the idea of the normalization based on MDP seems to be natural, it has a serious problem in common with other conventional algorithms for BSS. Namely, when the mixing matrix is almost singular, the norm of the separating matrix becomes very large and it can induce some numerical instability.

In this section, so as to solve the singularity problem, we introduce a generalized form of the original MDP, which we call ε-MDP. Based on ε-MDP, we derive a new ICA algorithm, in which a kind of regularization term is incorporated so as to obtain a certain robustness. Roughly speaking, it determines the separator such that its gain be sufficiently small at frequencies for which the mixing matrix is singular or almost singular.

Since there is an indeterminacy in the definition of the mixing process, the words “the mixing matrix is almost singular” is somewhat ambiguous. To eliminate the ambiguity we introduce a normalized form of the mixing process as follows. (A1) The source signals are white signals with zero mean and unity variance; (A2) Sensor signal $x_i(t)$ has been scaled to be of order unity. On these assumptions we want to give a definition about the singularity of the mixing process. Although it might not be mathematically rigorous, it is enough for the present purpose. Let the i-th row of $\mathbf{A}^{-1}(z)$ be $\mathbf{b}_i(z)$. We say that the mixing matrix $\mathbf{A}(z)$ is almost singular with respect to source i if $\|\mathbf{b}_i(z)\|$ is very large.

Let $\mathbf{w}_i^*(z)$ be the i-th row of $\mathbf{W}^*(z)$, then we have $\mathbf{w}_i^*(z) = a_{ii}(z)\mathbf{b}_i(z)$. This implies that if **A**(z) is almost singular with respect source i, the norm of $\mathbf{w}_i^*(z)$ becomes very large. This can cause instability when we execute an algorithm for BSS. Moreover, even if separation has been attained successfully, the separator obtained becomes very sensitive to noise. Suppose that the sensors’ signals are corrupted with noise **d**(t). Then the output of the separator becomes

$$\begin{aligned}y_i(t) &= \mathbf{w}_i^*(z)(\mathbf{A}(z)\mathbf{s}(t) + \mathbf{d}(t)) \\ &= a_{ii}(z)s_{ii}(t) + \mathbf{w}_i^*(z)\mathbf{d}(t).\end{aligned} \tag{12}$$

So, when $\|\mathbf{b}_i(z)\|$ and hence $\|\mathbf{w}_i^*(z)\|$ are very large, $y_i(t)$ becomes undesirably sensitive to noise.

To overcome these problems we extend the optimal separator as

$$\begin{aligned}\mathbf{W}^{**}(z) &\triangleq \mathbf{C}(z)\,\mathrm{diag}\mathbf{A}(z)\cdot\mathbf{A}^{-1}(z)\\ &= \mathbf{C}(z)\,\mathbf{W}^{*}(z)\end{aligned} \tag{13}$$

where $\mathbf{C}(z)$ is a diagonal matrix defined as

$$\begin{aligned}\mathbf{C}(z) &= \left\{\mathbf{I} + \varepsilon\ \mathrm{diag}\left(\mathbf{A}^{-1}(z)\mathbf{A}^{-H}(z)\right)\right\}^{-1}\\ &= \mathrm{diag}\left\{\frac{1}{1+\varepsilon\mathbf{b}_i(z)\mathbf{b}_i^H(z)}\right\}\end{aligned} \tag{14}$$

and ε is a small positive constant. We call this separator the ε-optimal separator. The ε-optimal separator is obviously valid, and the special case of $\varepsilon = 0$ reduces to the original optimal separator. Although the ε-optimal separator has lost some of the favorable properties held by the original optimal separator, it should be noted that the diagonal entries $\left\{1+\varepsilon\ \mathrm{diag}\left(\mathbf{b}_i(z)\mathbf{b}_i^H(z)\right)\right\}^{-1}$ of $\mathbf{C}(z)$ are zero-phase filters. It implies that every frequency component in the output of the ε-optimal separator receives no phase shift relative to that in the output of the optimal separator. An important property of the ε-optimal separator is that $\|\mathbf{w}_i^{**}(z)\|^2 \sim \frac{1}{\varepsilon}o(1)$. Namely, $\|\mathbf{w}_i^{**}(z)\|$ never exceeds a finite value of order $1/\sqrt{\varepsilon}$ (even for $\|\mathbf{b}_i(z)\| \to \infty$).

The ε-optimal separator can be characterized in similar ways to the original optimal separator.

Proposition 5: The ε-optimal separator is a valid separator that minimizes

$$Q_\varepsilon(\mathbf{W}(z)) \triangleq E\left[\|\mathbf{y}(t)-\mathbf{x}(t)\|^2\right] + \varepsilon\|\mathbf{W}(z)\|^2 . \tag{15}$$

Term $\varepsilon\|\mathbf{W}(z)\|^2$ is a kind of regularization term, which is often introduced in certain types of ill-posed optimization problems. We refer to the normalization based on minimization of $Q_\varepsilon(\mathbf{W}(z))$ as ε-MDP. Corresponding to Proposition 2, we have

Proposition 6: The ε-optimal separator is a valid separator that satisfies

$$\mathrm{diag}\left\{E\left[\left(\mathbf{y}(t)-\mathbf{x}(t)\right)\mathbf{y}^T(t-\tau)\right] + \varepsilon\sum_k \mathbf{W}_\tau\mathbf{W}_{\tau+k}^T\right\} = \mathbf{O} . \tag{16}$$

A similar algorithm to (11) can be derived by this proposition.

4. An Experiment

Many attempts have been made to perform BSS for mixed voice signals. 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 a 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 or three. It is very doubtful that the algorithms employed there work as well for a larger number of sound sources. Here we show a challenge to a much more difficult task; blind separation of eight sounds acquired in an ordinary office room.

We applied the proposed algorithm to 8 sound signals taken by 8 microphones at 10kHz sampling frequency. The source sounds were 8 voices of 'a' woman which were provided by eight loudspeakers. 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. The recovered sounds were considerably clear, which will be shown at the Workshop.

A remaining serious issue is that it takes a very long time to complete the calculation. In the case of two or three voices the algorithm can be executed in real time, but cannot in the case of 8 voices as yet. 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. International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99), pp. 371-376, 1999.

[3] K. Matsuoka and S. Nakashima, Minimal distortion principle for blind source separation, Proc. International Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001), pp. 722-727, 2001.

[4] K. Matsuoka, Principle for eliminating two kinds of indeterminacy in blind source separation, Proc. DSP2002, 147-150, 2002.