# An On-line Algorithm for Blind Source Separation on Speech Signals

Noboru Murata and Shiro Ikeda

*RIKEN Brain Science Institute*
*Hirosawa 2-1, Saitama, 351-0198 Japan*
*Noboru.Murata,Shiro.Ikeda@brain.riken.go.jp*

***Abstract*—** In this article, we propose an on-line algorithm for Blind Source Separation of speech signals, which is recorded in a real environment. This on-line algorithm makes it possible to trace the changing environment. The idea is to apply some on-line algorithm in the time-frequency domain. We show some results of experiments.

## I. Introduction

Recently, blind source separation (BSS) has attracted a great deal of attention in the engineering field. BSS is a problem to separate the independent sources from mixed observations, where mixing process is unknown. It is widely noticed that there are many possible applications such as noise reduction, removing crosstalk in telecommunication, preprocessing for multi-probed radar/sonar, analyzing biomedical data.

As a fundamental research, many algorithms have been developed for instantaneous mixtures, where only simple mixing process without time delay is considered. They have shown very good abilities to separate signals which are suitably thought as non-time delayed mixing, such as MEG (Magnetencephalograph) data. However, for separating acoustic signals recorded in a real environment, convolutive mixture have to be taken account of.

We have proposed a BSS method for temporal structured signals, such as speech signals recorded in a real environment [4, 8]. Our basic idea is as follows. First we transform mixed signals to the time-frequency domain, which is familiar with the name of spectrogram. Then we apply instantaneous BSS algorithm for each frequency channel independently. Next, we determine the correspondence of separated components in each frequency based on temporal structure of signals, and construct separated spectrogram of the source signals.

In this paper, we extend our algorithm for separating convolutive signals to on-line version. It aims at a situation in which a person is speaking in a room and moving around.

## II. Blind Source Separation Problem

Here we give a formulation of the BSS problem.

Source signals are denoted by a vector

$$\boldsymbol{s}(t) = (s_1(t), \cdots, s_n(t))^T, \quad t = 0, 1, 2, \ldots \tag{1}$$

and we assume that each component of $\boldsymbol{s}(t)$ is independent of each other. The independence of the sources are defined by

$$\begin{aligned} &p(s_1(t), \ldots, s_1(t-\tau), s_2(t), \ldots, s_n(t-\tau)) \\ &= \prod_i^n p(s_i(t), s_i(t-1), \ldots, s_i(t-\tau)), \end{aligned} \tag{2}$$

for any $\tau$, that is, the joint distribution of signals can be factorized by their marginals. Without loss of generality, we assume the source signal $\boldsymbol{s}(t)$ to be zero mean.

Observations are represented by

$$\boldsymbol{x}(t) = (x_1(t), \cdots, x_n(t))^T. \tag{3}$$

They correspond to the recorded signals. In the basic BSS problem, we assume that observations are linear mixtures of source signals:

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t), \tag{4}$$

where $A$ is an unknown linear operator. A typical example of linear operators is an $n \times n$ real valued matrix, which represents non-delayed mixing, and various learning algorithm are proposed for this setting (for example, [3]). In the case of real-room recording, a matrix of FIR filters is used as a linear operator [6, 9]. In this paper we focus on this problem, i.e.

$$\begin{aligned} \boldsymbol{x}(t) = A * \boldsymbol{s}(t) &= \left( \sum_k a_{ik} * s_k(t) \right), \\ \text{where } a_{ik} * s_k(t) &= \sum_{\tau=0}^{\tau_{max}} a_{ik}(\tau) s_k(t-\tau), \end{aligned} \tag{5}$$

The goal of BSS is to find a linear operator $B$ such that the components of the reconstructed signals

$$\boldsymbol{y}(t) = B\boldsymbol{x}(t) \tag{6}$$

are mutually independent, without knowing operator $A$ and the probability distribution of source signal $\boldsymbol{s}(t)$. Ideally $B$ should be the inverse of operator $A$, however, because of lack of information about the amplitude of the source signals and their order, there remains indeterminacy of permutation and dilation factors

## III. Proposed Algorithm

It is known that the human voice is stationary for a period shorter than a few 10msecs [5]. If it is longer than a few 10msecs and around 100msec, the frequency components of the speech will change its structure, and is not stationary. Therefore, first, we apply the windowed-Fourier transform to convolutive mixed signals (see Figure 1) and obtain the spectrogram,

$$\hat{\boldsymbol{x}}(\omega, t_s) = \sum_t e^{-j\omega t} \boldsymbol{x}(t) w(t - t_s), \tag{7}$$
$$\omega = 0, \tfrac{1}{N}2\pi, \ldots, \tfrac{N-1}{N}2\pi, \quad t_s = 0, \Delta T, 2\Delta T, \ldots$$

where $\omega$, $N$ and $t_s$ denote the frequency, the number of points of the discrete Fourier transform and the window position, respectively, $w$ is a window function (we used Hamming window) and $\Delta T$ is the shifting interval of moving windows.

With an appropriate window length, Equation (5) is well approximated as

$$\hat{\boldsymbol{x}}(\omega, t_s) = \hat{A}(\omega)\hat{\boldsymbol{s}}(\omega, t_s), \tag{8}$$

where $\hat{A}(\omega)$ is the Fourier transform of operator $A(t)$, and $\hat{\boldsymbol{s}}(\omega, t_s)$ is the spectrogram of $\boldsymbol{s}(t)$. This shows a convolutive mixture is a simple instantaneous mixture for a fixed $\omega$.

For extracting independent components from the mixed signals in each frequency channel, we use a recurrent neural network architecture [7, 2], in which the output vector is described as

$$\hat{\boldsymbol{u}}(\omega, t_s) = \hat{\boldsymbol{x}}(\omega, t_s) - B(\omega, t_s)\hat{\boldsymbol{u}}(\omega, t_s),$$

where $B(\omega, t_s)$ is a matrix, whose $ij$ element is a connection from the $j$-th component of output $\hat{\boldsymbol{u}}(\omega, t_s)$ to the $i$-th component of input $\hat{\boldsymbol{x}}(\omega, t_s)$ and whose diagonal elements are fixed to 0, that means there is no self-recurrent connection in the network. Since $\hat{\boldsymbol{u}}(\omega, t_s) = (B(\omega, t_s) + I)^{-1}\hat{\boldsymbol{x}}(\omega, t_s)$, the source signals are completely extracted when $A(\omega) = I + B(\omega, t_s)$, where $I$ is the identity matrix.

In the experiment described below, we adopt the following learning rule (see [1] for derivation of the algorithm and its stability analysis),

$$B(\omega, t_s + \Delta T) = B(\omega, t_s) - \eta \left(B(\omega, t_s) + I\right) \left(\mathrm{diag}\left(\phi(\boldsymbol{z})\boldsymbol{z}^*\right) - \phi(\boldsymbol{z})\boldsymbol{z}^*\right), \quad \boldsymbol{z} = \hat{\boldsymbol{u}}(\omega, t_s) \tag{9}$$

where diag$(\cdot)$ makes a diagonal matrix with the diagonal elements of its argument, $^*$ denotes complex conjugate, and

$$\phi(z) = \tanh(\mathrm{Re}(z)) + i \cdot \tanh(\mathrm{Im}(z)) \tag{10}$$

which operates component-wise to a column vector [9]. With using estimated matrix $B(\omega, t_s)+I$ and one independent component we obtain separated independent components of observation in each frequency as

$$\hat{\boldsymbol{v}}_\omega(t_s; i) = (B(\omega, t_s) + I)(0, \ldots, \hat{u}_i(\omega, t_s), \ldots, 0)^T. \tag{11}$$

Because of inherent indeterminacy of BSS problem, correspondence of $\hat{\boldsymbol{v}}_\omega(t_s; i)$ with another frequency is ambiguous. In our approach, individually separated frequency components are combined again based on the common temporal structure of original source signals. We assume that different frequency components from the same signal are under the influence of a similar modulation in amplitude. Defining an envelope

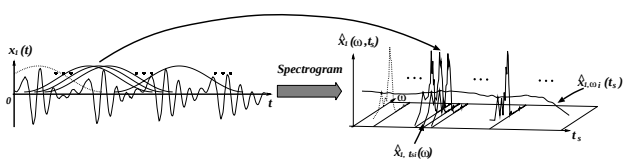


Figure 1: Windowed-Fourier transform (spectrogram)

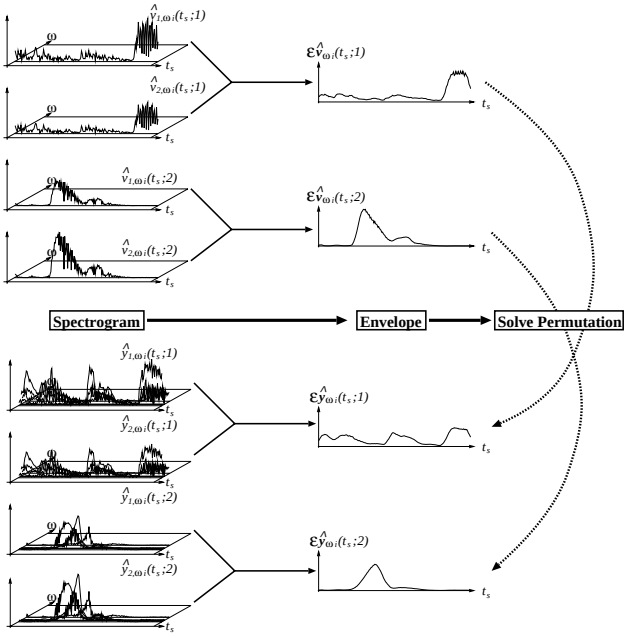


Figure 2: Construct separated spectrogram

making operator by

$$\mathcal{E}\hat{\boldsymbol{v}}_\omega(t_s;i) = \frac{1}{2M}\sum_{t'_s=t_s-M}^{t_s+M}|\hat{\boldsymbol{v}}_\omega(t'_s;i)|, \qquad (12)$$

where $M$ is a positive constant, we find a permutation $\sigma_\omega(i)$ which maximizes correlation between $\mathcal{E}\hat{\boldsymbol{v}}_\omega(t_s;\sigma_\omega(i))$ and $\mathcal{E}\hat{\boldsymbol{y}}(t_s;i) = \mathcal{E}\sum_{\omega'}\hat{\boldsymbol{v}}_{\omega'}(t_s;\sigma_{\omega'}(i))$ inductively (see Figure 2). For more detailed explanation about the practical implementation, see [4, 8].

## IV. Experiment

We applied the algorithm to a mixture of speech signals. Figure 3 shows the source signals which are recorded separately, and their spectrograms are shown in Figure 6. We mixed these signals as,

$$x_1(t) = s_1(t) + 0.3s_2(t-1) \qquad (13)$$
$$x_2(t) = s_2(t) + 0.3s_1(t-1). \qquad (14)$$

These inputs are shown in Figure 4, and their spectrograms are in Figure 7.

We applied our algorithm and as a result, separated signals are obtained in Figure 5, and their spectrograms are shown in Figure 8.

## V. Conclusion

We have proposed an on-line algorithm for convolutive mixture, based on the notion of temporal structure of speech signals. Thanks to the advantage of on-line learning, it can follow the changing environment in time and separate the signals. For example, it works for a situation in which a person is speaking in a room and moving around. Since our algorithm are constructed with rather simple procedures, i.e. Fourier transform and instantaneous BSS algorithms and it is easy to implement on a hardware, a possible application would be a system for tracking person's voice in real time.

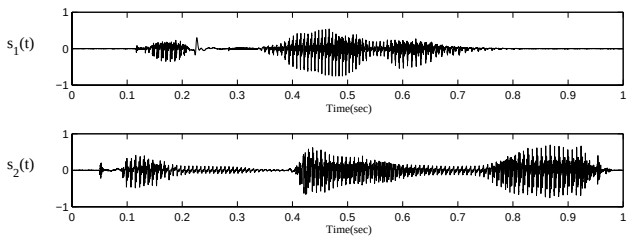


Figure 3: The source signals: each signal was spoken by a different male and recorded with sampling rate of 16kHz. $s_1(t)$ is a recorded word of "good morning" and $s_2(t)$ is a Japanese word "konbanwa" which means "good evening".

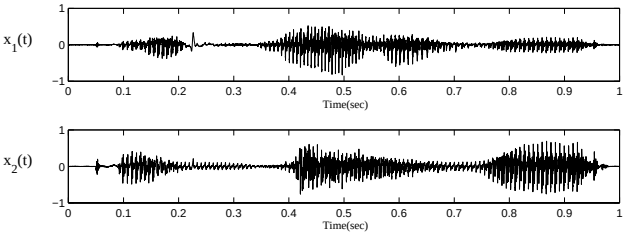


Figure 4: Input signals

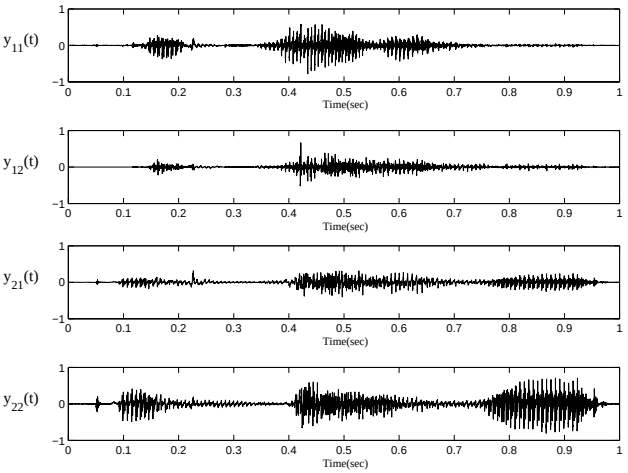


Figure 5: Output signals

## References

[1] S. Amari, T.-P. Chen, and A. Cichocki. Stability analysis of learning algorithms for blind source separation. *Neural Networks*, 10(8):1345–1351, 1997.

[2] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[3] J.-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. Signal Processing*, 44(12):3017–3030, December 1996.

[4] S. Ikeda and N. Murata. An approach to blind source separation of speech signals. In *Proceedings of 1998 International Conference on Artificial Neural Networks*, Skovde, September 1998. ICANN'98.

[5] H. Kawahara and T. Irino. Exploring temporal feature representations of speech using neural networks. Technical Report SP88-31, IEICE, Tokyo, 1988. (in Japanese).

[6] T.-W. Lee, A. J. Bell, and R. H. Lambert. Blind separation of delayed and convolved sources. In

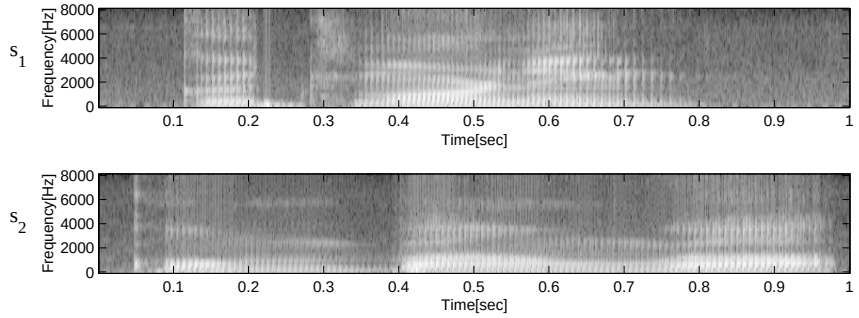


Figure 6: Spectrogram of the source signals

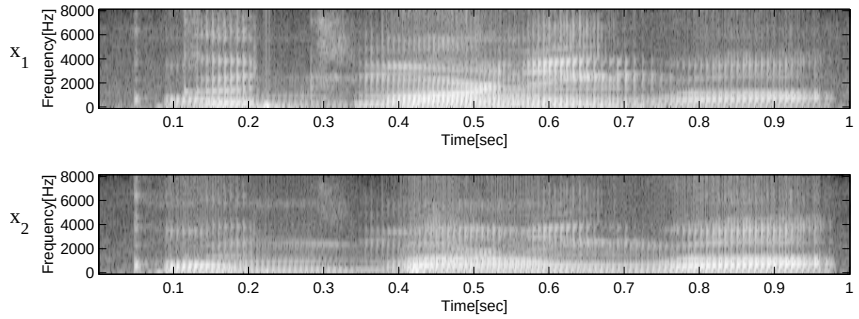


Figure 7: Spectrogram of the input signals

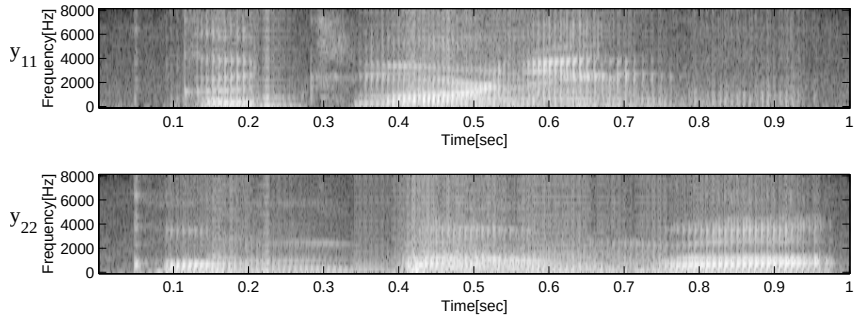


Figure 8: Spectrogram of the output signals $y_{11}$ and $y_{22}$

M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 758–764. MIT Press, Cambridge MA, 1997.

[7] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72(23):3634–3637, 1994.

[8] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. Technical Report BSIS Technical Reports No.98-2, RIKEN Brain Science Institute, 1998.

[9] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. In *International Workshop on Independence & Artificial Neural Networks*, University of La Laguna, Tenerife, Spain, February 1998.