AN EFFICIENT ALGORITHM FOR ADAPTIVE SEPARATION OF MIXTURE OF SPEECH SIGNALS

Satoshi NAKASHIMA[*] and Kiyotoshi MATSUOKA[**]

[*]Department of Control Engineering
[**]Department of Brain Science and Engineering
Kyushu Institute of Technology
Sensui-cho 1-1, Tobata, Kitakyushu, Japan

ABSTRACT

This paper proposes an efficient algorithm for blind source separation (BSS) of mixture of speech signals. Conventional on-line algorithms for blind separation usually assume that the sources are iid or linear processes. Since however speech signals have strong nonlinearity, those algorithms are not efficient with respect to convergence and sometimes induce instability. In order to solve these issues we introduce a more suitable probabilistic model for speech signals, namely, a speech signal is modeled as an amplitude modulation of a stationary random process. Based on the model, a new BSS algorithm is derived. A couple of examples reveal that the proposed algorithm determines a desired separator within a considerably short time.

1. INTRODUCTION

Blind source separation (BSS) or independent component analysis has been attracting a great deal of attention as a new topic of signal processing. It is a technique that separates a set of source signals from their mixtures observed by (at least) the same number of sensors as the sources. If the transfer function (matrix) of the mixing process is known beforehand, then the source signals can be recovered by applying its inverse to the observed signals, of course. The difficulty of BSS exists in the restriction that the mixing process must be identified from the observed signals only.

Recently, some algorithms efficient with respect to computing time have been proposed, but almost all of them are off-line. In many actual applications of BSS, adaptive processing is indispensable, but on-line algorithms proposed so far usually need an unendurably large number of iteration steps, particularly for convolutive mixture of sources.

A basic approach of BSS is as follows. Let q_i be a certain statistical model of the i-th source signal. Then the separator is determined such that the separator's outputs not only be mutually independent but also each of them has a distribution as close as possible to q_i. A foremost problem in this approach is how to choose a model q_i for each source. Inappropriate choice of q_i may cause slow convergence and moreover instability in the execution of on-line computation. Choice of q_i is task-dependent, of course. This paper proposes an efficient adaptive BSS algorithm for mixture of speech signals, but it can be applied to other signals that have a similar property.

Almost every conventional algorithm for BSS is built based on the assumption that the sources are iid (independent, identically distributed) processes or linear ones. However, such a signal as speech has a strong nonlinearity as shown below. In this paper we shall derive a new algorithm, based on the assumption that a speech signal should be modeled as an amplitude modulation of a stationary process. In the implementation of the algorithm, the second-order statistics are only required. The basic idea of our approach was first given by Kawamoto et al. [7,8], but they only dealt with the case of two sources. The present algorithm is applicable for more sources and its stability is proved. A remarkable feature of the algorithm is rapid convergence; the separation can be attained within several seconds of speech.

2. THE MIXING PROCESS AND THE DEMIXING PROCESS

Let us consider a situation where statistically independent random signals $s_i(t)$ ($i = 1,\ldots,N$) 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 sensor's outputs $x_i(t)$ ($i = 1,\ldots,N$) are given by a linear mixing process:

$$\mathbf{x}(t) = \sum_{\tau} \mathbf{A}_{\tau}\mathbf{s}(t-\tau) = \mathbf{A}(z)\mathbf{s}(t) \tag{1}$$

where $\mathbf{s}(t) \triangleq [s_1(t),\ldots,s_N(t)]^T$, $\mathbf{x}(t) \triangleq [x_1(t),\ldots,x_N(t)]^T$, and $\mathbf{A}(z) \triangleq \mathbf{A}_{\tau}z^{-\tau}$. Here, z represents the time-shift operator ($z^{-1}\boldsymbol{s}\,(t) \triangleq \boldsymbol{s}\,(t-1)$) and is also used as a complex variable. It is known that, in order to realize BSS, at most one source signal is allowed to be Gaussian. For the mixing process we assume two conditions: $\sum_{\tau}\|\mathbf{A}_{\tau}\| < \infty$ and non-singularity of $\mathbf{A}(z)$ for $|z| =$ 1. The first condition states that the mixing process is stable, and the second one claims that $\mathbf{A}(z)$ is invertible though the inverse $\mathbf{A}^{-1}(z)$ may not be a causal system.

To recover the source signals from the sensor signals, we consider a demixing process (which will be referred to as the separator) of the form

$$\mathbf{y}(t) = \sum_{\tau} \mathbf{W}_{\tau} \mathbf{x}(t-\tau) = \mathbf{W}(z)\mathbf{x}(t) \tag{2}$$

where $\mathbf{y}(t) \triangleq [\, y_1(t),\ldots,y_N(t)\,]^T$ and $\mathbf{W}(z) \triangleq \mathbf{W}_{\tau}\, z^{-\tau}$. 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. An essential difficulty in BSS is that $\mathbf{A}(z)$ or $\mathbf{A}^{-1}(z)$ must be estimated from the observed data $\boldsymbol{x}(t)$ only. Besides, the impulse response $\{\mathbf{W}_{\tau}\}$ might need to take a noncausal form in general, i.e., $\mathbf{W}_{\tau} \neq \mathbf{O}$ ($\tau < 0$). This problem is resolved by designing the separator so as to recover the source signals with a time lag.

In BSS the definition of the source signals has indeterminacy. Namely, if $s_1(t),\cdots,s_N(t)$ are source signals, their arbitrarily linear-filtered signals $e_1(z)s_1(t),\cdots,e_N(z)s_N(t)$ can also be considered source signals because they are also mutually independent. The mixing process is then $\mathbf{A}(z)\mathrm{diag}\{e_1^{-1}(z),\ldots,e_N^{-1}(z)\}$. There is no way to distinguish between $\{s_i(t)\}$ and $\{e_i(z)s_i(t)\}$ because the only information we are given a priori is the fact that the sources are mutually independent and the mixing process is a linear one.

A source signal $s_i(t)$ is called a *linear process* if it can be expressed as $s_i(t) = c_i(z)e_i(t)$, where $c_i(z)$ is a linear filter and $e_i(t)$ is an iid signal. Conventional methods usually assume this linearity of the sources, and the separator is designed to provide $e_i(t)$. In the context of blind separation of linearly mixed signals, there is no substantial difference between 'iid sources' and 'linear sources'. As shown in the next section, speech signals, which we want to deal with in this paper, are far from linear.

3. A CONVENTIONAL ALGORITHM

To make clear the peculiarity of our approach, we start with an approach proposed by S. Amari et al. [2,3]. Define

$$I(\mathbf{W}(z)) \triangleq -\sum_{i=1}^{N} E\left[\log q_i(y_i(t))\right] - h\left[\mathbf{y}(t)\right], \tag{3}$$

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 and $q_i(u)$ approximates well the real pdf of $s_i(t)$, then minimizing $I(\mathbf{W}(z))$ provides a desired solution. The minimization can be performed by the following iterative calculation (natural gradient learning):

$$\Delta\mathbf{W}_{\tau} = \alpha\left\{\mathbf{W}_{\tau} - \varphi(\mathbf{y}(t))\sum_{r}\mathbf{y}(t-\tau+r)\mathbf{W}_r\right\}, \tag{4}$$

where $\boldsymbol{\varphi}(\mathbf{y}(t)) \triangleq [\, \varphi_1(y_1(t)),\ldots,\varphi_N(y_N(t))\,]^T$ and φ_i is defined as $\varphi_i(u) \triangleq -d\log q_i(u)/du$. α is a small positive constant.

When deriving the above algorithm it has been assumed that each of the source signals is an iid process or more generally a linear process. As shown below, however, speech signals have strong nonlinearity.

4. BSS FOR MIXTURE OF SPEECH SIGNALS

4.1 Nonlinearity of speech signals

Here we see how far speech signals are from linear processes. To this end, we attempted to decorrelate a speech signal $x(t)$, using a decorrelator:

$$y(t) = w(z)x(t) = \sum_{\tau} w_{\tau} x(t-\tau)\,. \tag{5}$$

To determine $w(z)$ or $\{w_{\tau}\}$, we utilized a single-input, single-output version of (4):

$$\Delta w_{\tau} = \alpha\left\{w_{\tau} - \varphi(y(t))\sum_{r} y(t-\tau+r)w_r\right\}, \tag{6}$$

where $\varphi(u) = (1-e^{-u})/(1+e^{-u})$. If the speech signal were a linear process, then the output $y(t)$ of the decorrelator ought to have been iid signal. Namely, the scatter diagram of $y(t)$ and $y(t+\tau)$ would show a similar shape for every τ, and moreover it would be cross-shaped because speech signals can be considered super-Gaussian.

Figure 1 shows the actual distribution of $y(t)$ and $y(t+\tau)$ for $\tau = 1, 10, 100, 1000$ (one step corresponds to 0.1ms). For small τ the scatter diagrams are circular, suggesting that $y(t)$ and $y(t+\tau)$ are not independent for small τ. This can be seen more clearly in Figure 2, which shows two kinds of forth-order cross cumulants between $y(t)$ and $y(t+\tau)$ as a function of τ. For small τ the cross cumulants are very large while they are nearly zero for large τ.

The above result suggests that a speech signal cannot be made iid by a linear filter, implying that it is not a linear process. We may well imagine that the pdf of a speech signal (or more accurately, the speech signal that is made as iid as possible) takes a form as shown in Figure 4.

If a BSS algorithm based on the linearity of the source signals is applied to nonlinear signals, two kinds of issues can arise:

(i) Instability: as shown by Ohata and Matsuoka [9], such an iterative algorithm as described in §3 does not stably give a desired separator if the source signals are far from linear.

(ii) Inefficiency: even if it provides a desired separator, the convergence will become very slow.

4.2 A model for speech signals

We want to propose the following model for speeches uttered in common situations, that is, a speech signal $s(t)$ is modeled as an amplitude-modulation of a stationary signal:

$$s(t) = \sigma(t)r(t)\,, \tag{7}$$

where

(i) $r(t)$ and $\sigma(t)$ are mutually independent;

(ii) $r(t)$ is a stationary, white, Gaussian process with zero mean and unity variance;

(iii) $\sigma(t)$ is a stationary random process and, moreover, is slowly varying with time.

According this assumption, given $\sigma(t)$, the conditional pdf of $s(t)$ becomes

$$q(s(t)\,|\,\sigma(t)) = \frac{1}{\sqrt{2\pi\sigma^2(t)}}\exp\left\{-\frac{s^2(t)}{2\sigma^2(t)}\right\}. \qquad (8)$$

Note that as for $\sigma(t)$ its probabilistic property is not given specifically.

Signal $s(t)$ has the following properties:

(a) It is a super-Gaussian signal except that $\sigma^2(t)$ is constant with time.

(b) It is a white signal in the sense of second-order statistics; $E[\, s(t)s(\, t\, +\tau)\,] = 0$ for $\tau \neq 0$.

(c) For large τ, $s(t)$ and $s(\, t+\tau\,)$ are mutually independent while for small τ, they are not independent.

Here we show an example. It can be shown that if $\sigma(t)$ is a stationary Gaussian process with zero mean, then the (normalized) kurtosis of $\sigma(t)$ becomes 9. The actual kurtosis of the speech signal dealt with in 4.1 was 9.82, being very close to 9. Moreover, $y(t)$ and $y(\, t+\tau\,)$ becomes independent for τ greater than several hundreds. So, we here assume that $\sigma(t)$ is a Gaussian process given by the following AR model:

$$\sigma(t) = 0.997\sigma(t-1) + 0.0774n(t), \qquad (9)$$

where $n(t)$ is a stationary, white, Gaussian process with zero mean and unity variance. Figure 3 shows the scatter diagrams of $s(t)$ and $s(t+\tau)$. They remarkably resemble the scatter diagrams shown in Figure 1. Note, however, that below we will not pay any attention to the probabilistic characteristics of $\sigma(t)$ except that it is a slowly varying function.

4.3 A new BSS algorithm

If we knew the variance $\sigma_i^2(t)$ of the i-th source, we could employ (8) as a 'target' pdf in (3). Then, we would obtain the following algorithm:

$$\Delta\mathbf{W}_\tau = \alpha\left\{\mathbf{W}_\tau - \mathbf{R}^{-1}(t)\mathbf{y}(t)\sum_r \mathbf{y}^T(t-\tau+r)\mathbf{W}_r\right\}, \qquad (10)$$

where $\mathbf{R}(t) = \mathrm{diag}\{\sigma_1^2(t),\cdots,\sigma_N^2(t)\}$.

In actual situations, however, $\sigma_i(t)$ is unknown, of course. Here we replace $\mathrm{diag}\{\sigma_1^2(t),\cdots,\sigma_N^2(t)\}$ in (10) with $\mathbf{D}(t) \triangleq \mathrm{diag}\{E[y_1^2(t)],\cdots, E[y_N^2(t)]\}$, leading to

$$\Delta\mathbf{W}_\tau = \alpha\left\{\mathbf{W}_\tau - \mathbf{D}(t)^{-1}\mathbf{y}(t)\sum_r \mathbf{y}^T(t-\tau+r)\mathbf{W}_r\right\}. \qquad (11)$$

$E[y_i^2(t)]$ needs to be estimated by a time averaging filter for $y_i^2(t)$. It should be noted that, before source separation is achieved, $\mathbf{D}(t)$ is not an estimate of $\mathbf{R}(t)$. But once the separation is achieved, $\mathbf{D}(t)$ becomes $\mathbf{R}(t)$ (To be exact, $E[y_i^2(t)]$ becomes proportional to $\sigma_i^2(t)$). So, it can be expected the behavior of (11) will be almost the same as that of (10) around a desired solution. A rough outline of stability analysis for (11) is given in Appendix.

5. EXAMPLES

Here, we show two examples. For the actual calculation the following algorithm was used:

$$\Delta\mathbf{W}_\tau = \alpha\left\{\mathbf{W}_\tau - \hat{\mathbf{D}}^{-1}(t-L)\mathbf{y}(t-L)\sum_{r=0}^{L}\mathbf{y}^T(t-L-\tau+r)\mathbf{W}_r\right\} \qquad (12)$$

$$\hat{\mathbf{D}}(t) = \frac{1}{2K+1}\sum_{k=-K}^{K} y_i^2(t-L-k) \qquad (13)$$

In the examples below, the parameters of the learning were chosen as $\alpha = 0.00001$, $K = 25$, and $L = 40$, and the initial condition of the demixing process were set as $\mathbf{W}_{L/2} = \mathbf{I}$, $\mathbf{W}_\tau = \mathbf{O}$ ($\tau \neq L/2$).

Example 1:

The sources are three speeches in radio news. Three convolutive mixtures were produced by the following artificial mixing process:

$$\mathbf{A}(z) = \begin{bmatrix} 1.0 & 0.6z^{-1} & 0.4z^{-2} \\ 0.6z^{-1} & 1.0 & 0.6z^{-1} \\ 0.4z^{-2} & 0.6z^{-1} & 1.0 \end{bmatrix}.$$

Figure 5 and 6 shows the result of $\mathbf{W}(z)$ obtained by the present algorithm and the total transfer function $\mathbf{V}(z) = \mathbf{W}(z)\mathbf{A}(z)$ from the sources to the outputs after around 6 seconds had elapsed. On the other hand, when using a conventional algorithm (4) instead, several tens times more iterations were required to attain a satisfactory separation.

Example 2:

In this example two speech data were taken with two microphones in a room. Figure 8 shows the original speeches, the observed signals, and the output of the separator. One can see that the waveforms of sources resemble very closely those of the outputs. Also in this case, the separation was attained within around several seconds of speech.

6. CONCLUSION

We have shown a new algorithm for BSS of mixture of speech signals. It is based on a nonlinear model for sources, which is more suitable for BSS of speech signals than conventional iid models. The experiment shows a considerable good performance in respect of convergence speed.

However there are a problem however in the present algorithm. Namely, since the equilibrium of the algorithm is semi-stable, there is a possibility that a numerical instability can occur. To suppress this tendency it is better to add some constraint such as Minimum Distortion Principle proposed by Matsuoka and Nakashima. [10]

REFERENCES

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

[2] S. Amari, S. C. Douglas, A. Cichocki and H. H. Yang: Multichannel blind deconvolution and equalization using the natural gradient, The First Signal Processing Workshop on Signal Processing Advances on Wireless Communications, Paris, France, pp.101-104, 1997.

[3] S. Choi, S. Amari, A. Cichocki and R. Liu: Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels, International Workshop on Independent Component Analysis and Blind Signal Separaton (ICA'99), Aussois, France, pp.371-376, 1999.

[4] S. Amari: Natural gradient learning works efficiently in learning, Natural Computation 10, pp.251-176, 1997.

[5] M. Girolami: *Self-Organizing Neural Networks*, 1999.

[6] J. F. Cardoso and A. Souloumiac: Blind beamforming for non Gaussian signals, IEE- Proceedings-F 140-6, pp.362-370, 1993.

[7] M. Kawamoto, K. Matsuoka, and N. Ohnishi: Real world blind separation of convolved speech signals, 1999 International Conference on Neural Networks, Paper No.2058, 1999.

[8] M. Kawamoto, A. K. Barros, A. Mansour, K. Matsuoka, and N. Ohnishi; Blind signal separation for convolved non-stationary signals, The Trans. of the Institute of Electronics, Information and Communication Engineers A, Vol. J82-A, No.8, pp.1320-1328, 1999.

[9] M. Ohata and K. Matsuoka; Stability analyses of a couple of blind separation algorithms when the sources are nonlinear processes, to appear in IEEE Trans. on Signal Processing.

[10] K. Matsuoka and S. Nakashima,; Minimal distortion principle for Blind Source Separation, to appear in this workshop.

APPENDIX: STABILITY ANALYSIS OF THE PROPOSED ALGORITHM

If the learning coefficient α is very small, the behavior of the dynamics (11) can be approximated by the following continuous equation:

$$\frac{d\mathbf{W}_\tau}{dt} = \mathbf{W}_\tau - \mathbf{D}(t)^{-1} E\left[\mathbf{y}(t)\sum_r \mathbf{y}^T(t-\tau+r)\right]\mathbf{W}_r . \qquad \text{(A-1)}$$

Let $\mathbf{V}(z) = \mathbf{W}(z)\mathbf{A}(z) = \sum_\tau \mathbf{V}_\tau z^{-\tau}$ be the transfer function matrix from the sources to the output terminal of the separator, then the impulse response $\mathbf{V}_\tau$ is written as

$$\mathbf{V}_\tau = \sum_k \mathbf{W}_{\tau-k}\mathbf{A}_k . \qquad \text{(A-2)}$$

Convolving both the sides of (A-1) with $\{\mathbf{W}_\tau\}$ from the right, we have

$$\frac{d\mathbf{V}_\tau}{dt} = \mathbf{V}_\tau - \mathbf{D}(t)^{-1} E\left[\mathbf{y}(t)\sum_r \mathbf{y}^T(t-\tau+r)\right]\mathbf{V}_r . \qquad \text{(A-3)}$$

It can easily be found that any impulse response of the following form is an equilibrium of (A-3):

$$\bar{\mathbf{V}}_0 = \text{ a nonsingular diaganal matrix}$$
$$\bar{\mathbf{V}}_\tau = \mathbf{0} \quad (\tau \neq 0)$$

In order to investigate local stability of this equilibrium. Consider a small perturbation $\Delta\bar{\mathbf{V}}_\tau$ around $\bar{\mathbf{V}}_\tau$ as

$$\mathbf{V}_\tau = \bar{\mathbf{V}}_\tau + \Delta\mathbf{V}_\tau \qquad \text{(A-4)}$$

Substituting (A-4) for (A-3), we obtain

$$\frac{d\Delta\mathbf{V}_\tau}{dt} = -\left(\bar{\mathbf{V}}_0\mathbf{R}(t)\bar{\mathbf{V}}_0^T\right)\left(\Delta\mathbf{V}_\tau\mathbf{R}(t)\bar{\mathbf{V}}_0^T + \bar{\mathbf{V}}_\tau\mathbf{R}(t)\Delta\mathbf{V}_{-\tau}^T\right)\bar{\mathbf{V}}_0 . \qquad \text{(A-5)}$$

Here we have assumed that $\mathbf{R}(t)$ alters very slowly with time.

We can see that (A-5) can be decomposed into a set of two-dimensional differential equation:

$$\frac{d}{dt}\begin{bmatrix}\Delta v'_{ij,\tau}\\ \Delta v'_{ji,-\tau}\end{bmatrix} = -\begin{bmatrix}\dfrac{\bar{v}_j^2}{\bar{v}_i^2}\dfrac{r_j(t)}{r_i(t)} & 1\\ 1 & \dfrac{\bar{v}_i^2}{\bar{v}_j^2}\dfrac{r_i(t)}{r_j(t)}\end{bmatrix}\begin{bmatrix}\Delta v'_{ij,\tau}\\ \Delta v'_{ji,-\tau}\end{bmatrix} \qquad \text{(A-6)}$$

where $\Delta v'_{ij,\tau} = \bar{v}_i\Delta v_{ij,\tau}$, $\Delta v'_{ji,\tau} = \bar{v}_j\Delta v_{ji,-\tau}$. It is easy to show that the coefficient matrix is negative semi-definite for any t. If $r_i(t)$ and $r_j(t)$ independently fluctuate sustainedly, then $\Delta v'_{ij,\tau}$ and $\Delta v'_{ji,-\tau}$ converge to zero.

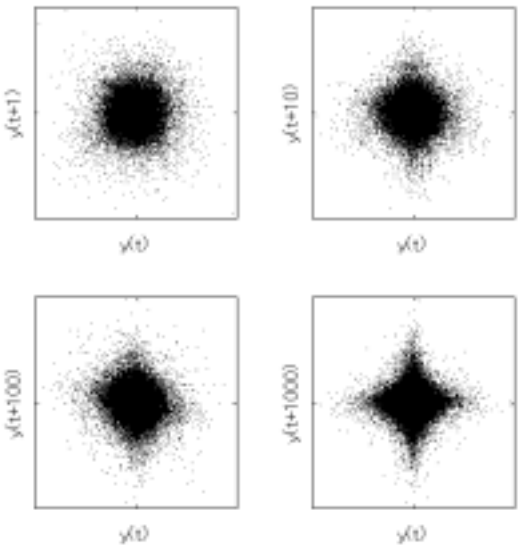


Figure 1 The scatter diagram of $y(t)$ and $y(t + \tau)$, where $y(t)$ is a decorrelated speech signal.

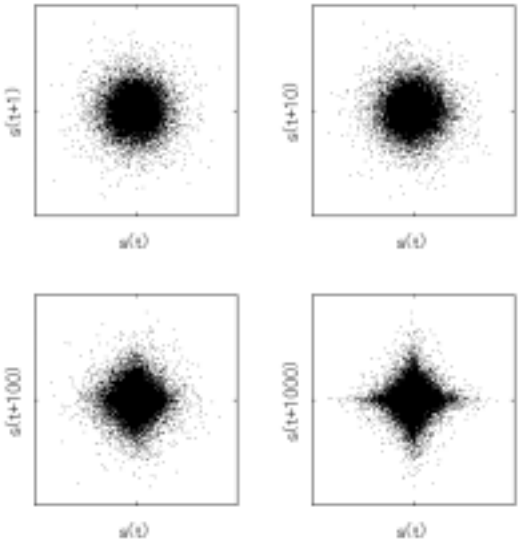


Figure 3 The scatter diagram of $s(t)$ and $s(t + \tau)$, where $s(t)$ is an amplitude modulation of a stationary, white signal.

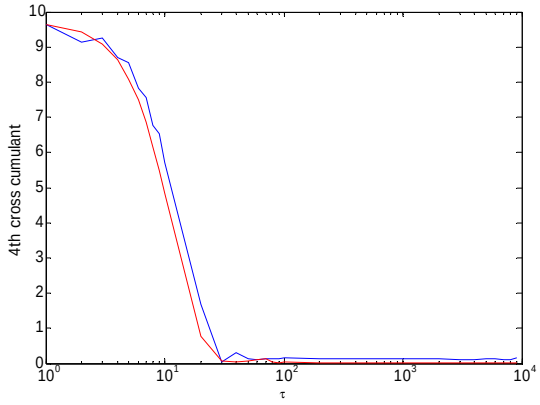


Figure 2 The forth-order cross cumulants between $y(t)$ and $y(t + \tau)$.

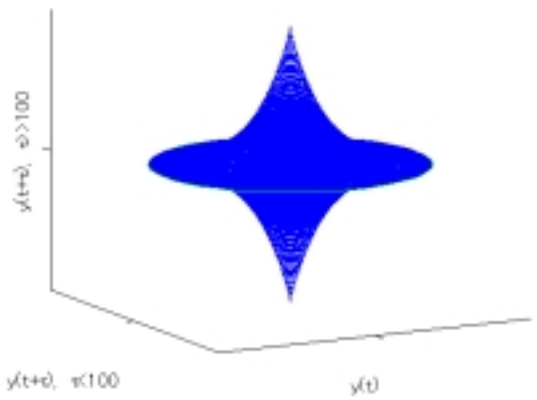


Figure 4 An image of the distribution of speech signals.

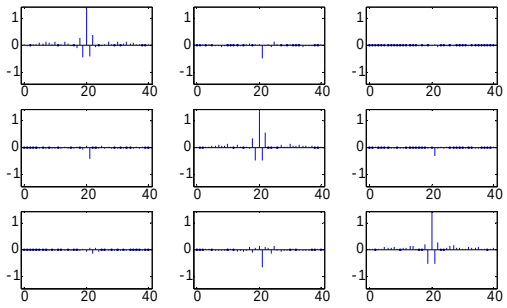


Figure 5 This figure shows the impulse responses of $\mathbf{W}(z)$.

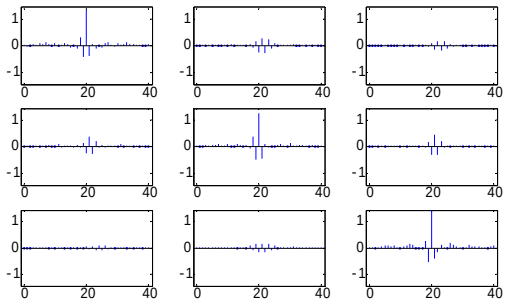


Figure 6 This figure shows the impulse response of total $\mathbf{V}(z)$.

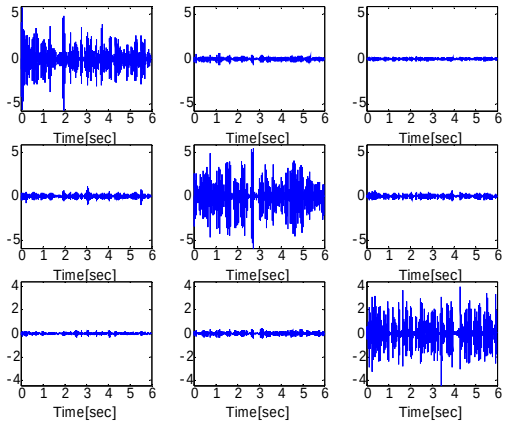
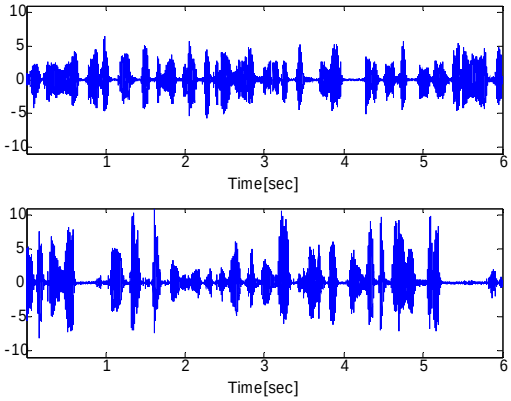
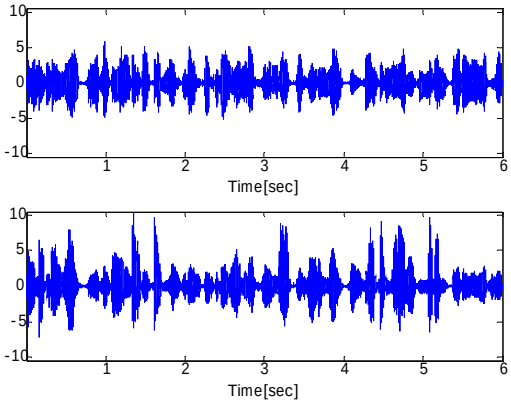


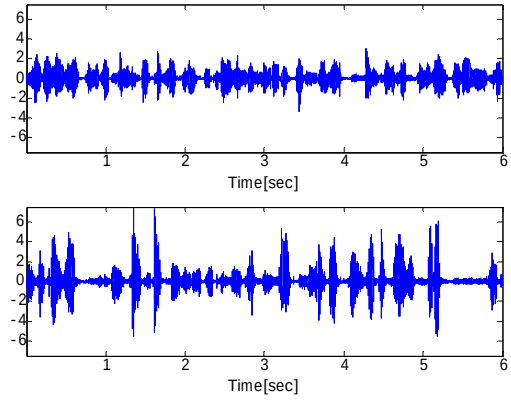
Figure 7 This figure shows to what degree the source signals are contained in each output of the separator. The figure in the *i*-th row and the *j*-th column indicates the *j*-th source signal contained in the *i*-th output.



(a) The original speeches



(b) The voices sensed by the microphones



(c) The output of the separator.

Figure 8 The result of Example 2