Multistage Convolutive Blind Source Separation for Speech Mixture

Yanxue Liang, Ichiro Hagiwara.

Department of Mechanical Sciences and Engineering
Graduate School of Science and Engineering of Tokyo Institute of Technology, Japan
yxliang@stu.mech.titech.ac.jp, hagiwara@mech.titech.ac.jp

Abstract

Blind source separation for convolutive mixture of speech signals has been addressed in many literatures. However, widely applied Multichannel Blind Deconvolution (MBD) method suffers whitening effect or arbitrary filtering problem which results in dramatic decrease of Automatic Speech Recognition system's performance. In present paper, a new MBD based multistage method is proposed, in which contributions of each source to every microphone are final goal rather than original signals. In detail, MBD is first implemented using entropy maximization criterion combined with Natural Gradient (NG) algorithm, then compensation matrix is constructed, based on which sources are recovered to its contribution to every microphone, i.e., whitening effect or arbitrary filtering problem has been transformed to fixed filtering problem. After compensation processing, for a certain source, it becomes Single Input and Multi-Output (SIMO) problem. Thus, not only spatial quality of source can be preserved, but also SIMO blind deconvolution can be further applied to fully recover temporal structure of speech signal. Finally, experiment shows validity and superiority over other methods in both spectra preservation efficiency and speed.

Index Terms: Multistage, Convolutive Source Separation, spectra compensation

1. Introduction

Blind Source Separation (BSS) is a field that has received considerable attention in the latest decade in many research fields, such as communications, speech separation and signal processing and control [1] and [2].

Recently, much research has been devoted to the convolutive case. Methods for such scenario, generally, can be classified into two classes. The one is MBD method, and the other is Convolutive BSS (CBSS). The difference lies in the former attempts to make system's outputs both spatially and temporally independent and later just performs separation without explicitly deconvolving outputs. For speech case, MBD is obviously unsuitable due to its side effect of temporally whitening the outputs; Comparatively, CBSS is more appropriate to solve cocktail party problem.

To overcome whitening effect in cocktail problem, various methods have been proposed. K.Torkkola [3] and Choi [4] proposed recurrent network method, by which filtered version estimation can be obtained. However, it is hard to design a feedback type separator so as to guarantee its stability, especially for non-minimum phase mixing process. As for nonholonomic method [5], it releases diagonal constrains on temporal structure of outputs. Indeed, it can not prevent arbitrary filtering problem radically. LP based MBD [6] [7] method works well only under the precondition that inherent dependencies of speech signal does not share same delay scope with echo delay. Otherwise, LP processing would destroy mixing information such that separation can not be implemented correctly. Due to early reflection in real environment, such condition is quite difficult to be satisfied. Another class method is two-stage method using newly designed criterion to estimate the contribution of each source to every microphone, such as SIMO-ICA [8] [13]and Minimal Distortion Principle [9]. However cost function of the MDP does not equal zero even in optimal point, which brings difficulties in finding equilibrium point blindly. High computational complexity of SIMO-ICA's is relatively surprising.

In this paper, we present novel multi-stage method to recover original signal up to its contribution to every microphone by employing a compensation matrix. In the proposed framework, conventional blind separation is first carried out using entropy maximization method combined with nonholonomic NG. After separation, the arbitrarily filtered version estimation and separation matrix are available. We then construct a compensation matrix in frequency domain, based on which contributions of each source to every microphone can be retrieved. Thus, for a certain source, the problem becomes a number of SIMO problems that is relatively suitable for post processing, such as deconvolution, because an exact inversion of the mixing system and thus, a perfect reconstruction of the source is possible for such case if multi-path impulse response does not share common zeros in z-plane [10]. Applying proposed method, not only temporal structure can be repaired, but also spatial quality of each source can be maintained. The validity of suggested approach is verified through experimental and performance comparison.

2. Problem statement and conventional solution

2.1. Mixing process

In the convolutive blind source separation task, m statistically independent source signals $\mathbf{s}(t) = [s_1(t) \quad ..... \quad s_m(t)]^T$ are mixed by a causal linear multichannel system and corresponding n mixture observations $\mathbf{x}(t) = [x_1(t) \quad ..... \quad x_n(t)]^T$ are given by

$$\mathbf{x}(t) = \sum_{k=0}^{K} \mathbf{H}(k)\mathbf{s}(t-k) \tag{1}$$

where $\{\mathbf{H}(k)\}$ is a sequence of $n \times m$ matrices that denotes the Impulse Response (IR) of acoustic environment with coefficient $h_{ji}(k)$ being IR between i -th source and j -th sensor. For

simplicity, in this study, we assume that $m = n$ and background noise is negligible.

The objective of BSS is to estimate source signal from mixture observations without any prior information. Utilizing the causal Finite Impulse Response (FIR) filter matrix, separation process can be represented in the form

$$\mathbf{y}(t) = \sum_{l=0}^{L} \mathbf{W}(l)\mathbf{x}(t-l) \tag{2}$$

where $\mathbf{y}(t) = [y_1(t) \quad ..... \quad y_m(t)]^T$ contains estimations of source signals without interference. There are two kinds of results in term of relation between $\mathbf{y}(t)$ and $\mathbf{s}(t)$, one is

$$y_i(t) = d_{ij} s_j(t - \Delta_i) \tag{3}$$

where d_{ij} is a scaling factor only from $j \to i$ and Δ_j is a time delay. Such results can preserve waveform of source exactly. The other one takes the form

$$y_i(t) = \sum_{l=0}^{L-1} d_{ij}(l) s_j(t-l) \tag{4}$$

which means that estimation is a filtered version of source, spectra characteristics has been destroyed in procedure of separation. Undoubtedly, the first solution is ideal for speech separation task.

2.2. Information maximization approach BSS with nonholonomic constraints

Until now, a number of approaches have been proposed to achieve BSS using different optimization criteria, such as Second Order Statistics (SOS) method, High Order Statistics (HOS) method. Well known information maximization approach derived from independence assumption is rather popular one due to its success in a number of applications. This method attempts to make information represented by outputs maximized through reducing redundancy in both inner symbol and inter symbol, finally outputs of system are independent spatially and temporally. It is also equivalent to minimizing the mutual information (K-L divergence) between the components of network output to render them independent.

Employing information maximization criterion and natural gradient, et al, resulting coefficient updates in time domain are

$$\mathbf{W}_l(k+1) = \mathbf{W}_l(k) + \eta\left[\mathbf{W}_l(k) - \mathrm{f}(\mathbf{y}(k-L))\mathbf{u}^T(k-l)\right] \tag{5}$$

$$\mathbf{y}(k) = \sum_{l=0}^{L} \mathbf{W}_l(k)\mathbf{x}(k-l) \tag{6}$$

$$\mathbf{u}(k) = \sum_{q=0}^{L} \mathbf{W}_{L-q}(k)\mathbf{y}(k-q) \tag{7}$$

where $\mathbf{W}_l(k)$ is l -th order of separation matrix at k -th iteration, η is learning rate, $\mathrm{f}(\mathbf{y}) = [\mathrm{f}_1(y_1) \quad ..... \quad \mathrm{f}_m(y_m)]^T$ represents a vector of nonlinear function acting on every component of outputs and is defined as $f_i(y_i) = \tanh(y_i)$ here. Through analysis, equilibrium point of above learning algorithm is

$$E\{f_i(y_i(k))y_j(k-l)\} = \delta_{ij}\delta_l \tag{8}$$

Equation (8) indicates outputs are both temporally and spatially independent when iteration arrives at stationary point, which is side effect for speech case.

To avoid whitening effect, also overcome instability brought by nonstaionarity of signal, nonholonomic constraints method is proposed. The natural gradient algorithm updates with nonholonomic constraints takes the form

$$\mathbf{W}_l(k+1) = \mathbf{W}_l(k) + \eta \cdot \sum_{r=0}^{L} off - diag\left\{\mathrm{f}(\mathbf{y}(k-L))y^T(k-L-l-r)\right\}\mathbf{W}_r(k) \tag{9}$$

$$\mathbf{y}(k) = \sum_{l=0}^{L} \mathbf{W}_l(k)\mathbf{x}(k-l) \tag{10}$$

Correspondingly, stationary point of above equation satisfy

$$\begin{aligned} &E\{off - diag[\mathrm{f}(\mathbf{y}(k))\mathbf{y}^T(k-l)]\} = 0 \\ &E\{diag[\mathrm{f}(\mathbf{y}(k))\mathbf{y}^T(k-p)]\} = \mathbf{\Lambda} \end{aligned} \tag{11}$$

where $\mathbf{\Lambda}$ is diagonal matrix that is automatically adjusted during the learning process. To void the noncausality, L tap delay is adopted in practical implementation for equation (5)-(10). It can be seen form (9)-(10) that there are no particular constraints on temporal structure of outputs, which implies that whitening problem may become arbitrary filtering problem.

3. Proposed multi-stage method

3.1. Motivation

In the previous research, many approaches have been suggested to deal with whitening problem for signal with temporal structure. It is shown that Minimal Distortion Principle (MDP) criterion and SIMO-ICA method are attractive. Their main idea lies in whole procedure is divided into two stages, contributions of source to microphones are estimated first, and then SIMO deconvolution is conducted at second stage. However, in first stage, high computational complexity and high sensitivity to the initial settings of separation filter is serious drawback. It motivates us to propose more convenient and efficient method to overcome such problem. In proposed novel multistage method, conventional BSS is conducted first, whitening effect or arbitrarily filtering problem would happen during separation. Sequentially, we construct a compensation matrix utilized to retrieve contribution of each source to every microphone. That is, separation is split into two stages, conventional separation process and compensation process. After separation and compensation, not only temporal structure of signal can be obtained by using deconvolution method but also spatial information of source can be preserved exactly. Moreover, many post processing can be applied, such as Direction of Arrival (DOA) and even Source Location (SL). Hence, multi-stage strategy can provide most convenience with quite low calculation cost. It can be combined with either time domain methods or frequency domain methods in a quite natural way. The whole framework is shown in fig.1

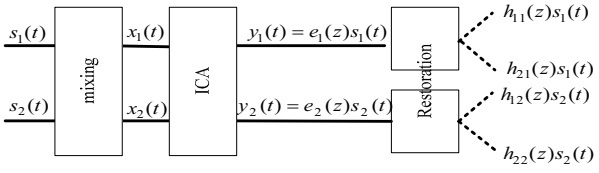

Figure 1 *Schematic diagram of multistage BSS*

3.2. Algorithm

Using (9)-(10) mentioned in section 2.2, conventional BSS is performed in time domain. After separation, we obtain separated signals whose components are mutually independent are given by

$$\begin{bmatrix} y_1(t) & \dots & y_m(t) \end{bmatrix}^T = W(z)\begin{bmatrix} x_1(t) & \dots & x_m(t) \end{bmatrix}^T \quad (11)$$

where $\begin{bmatrix} y_1(t) & \dots & y_m(t) \end{bmatrix}^T$ is output vector, $W(z)$ denotes z-transform of $m \times m$ separation matrix. Without considering permutation problem, relation between outputs and original signals is given by

$$y_i(t) = e_i(z)s_i(t) \quad (12)$$

where $e_i(z)$ represents arbitrary filter. An equivalent description of (11) in frequency domain is

$$\begin{bmatrix} y_1(f_i) & \dots & y_m(f_i) \end{bmatrix}^T = W(f_i)\begin{bmatrix} x_1(f_i) & \dots & x_m(f_i) \end{bmatrix}^T \quad (13)$$

where f_i is frequency bin. Without loss generality, we use f uniformly hereinafter. Assuming reverse matrix of $W(f)$ at every frequency bin exists, i.e., $R(f) = W(f)^{-1}$, we write equation (13) in the following form

$$\begin{bmatrix} x_1(f) & \dots & x_m(f) \end{bmatrix}^T = R(f)\begin{bmatrix} y_1(f) & \dots & y_m(f) \end{bmatrix}^T \quad (14)$$

When only one component of separated signal vector on right hand of (14) is kept, (14) becomes

$$\begin{bmatrix} x_{1i}(f) & \dots & x_{mi}(f) \end{bmatrix}^T = R(f)\begin{bmatrix} 0 & y_i(f) & 0 \end{bmatrix}^T \quad (15)$$

where $x_{ji}(f)$ is contribution of $y_i(f)$ to $x_j(f)$ without interference. It is interesting to note that mixing process can also be written in similar form as follows in which only one source is considered.

$$\begin{bmatrix} x'_{1i}(f) & \dots & x'_{mi}(f) \end{bmatrix}^T = H(f)\begin{bmatrix} 0 & s_i(f) & 0 \end{bmatrix}^T \quad (16)$$

From (15) and (16), combining with (12), we have conclusions that left hand of (15) corresponds to contribution of i-th source to every microphone. This happens to be objective of compensation. All the sources can be retrieved up to their contributions to microphones in same way.

4. Experiment and discusses

4.1. Experiments condition and evaluation

The experiments are conducted using speech signal convolved with the IR of room environment (Fig. 2) simulated by image method [11] with reverberation time $T60 = 100ms$. We use two speaker and two microphone with inter space 4cm. Sampling frequency is 8 KHz and each speech signal is 7 seconds. The two sources arrive one meter away from two directions, -45^o , 45^o respectively.

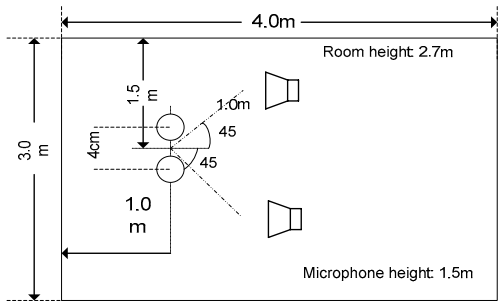

Figure 2 *layout of reverberant room used in experiment.*

To evaluate performance of proposed method, Signal-to-Noise Ratio Improvement (SNRI) defined as following is used to test separation performance.

$$SNRIi = 10\log 10\frac{\sum_t \left|C_{ii}(z)s_i(t)\right|^2}{\sum_t \left|C_{ij}(z)s_j(t)\right|^2}\frac{\sum_t \left|H_{ij}(z)s_j(t)\right|^2}{\sum_t \left|H_{ii}(z)s_i(t)\right|^2} \quad (17)$$

where $j \neq i$ and $C(z) = W(z)H(z)$ is global system function. Sound Quality (SQ) described below indicates spatial information preservation after separation.

$$SQij = 10\log 10\frac{\sum_t \left|x_{ij}(t)\right|^2}{\sum_t \left|x_{ij}(t) - C_{ij}(z)s_j(t)\right|^2} \quad (18)$$

Final SNRI and SQ scores is averages value of all channels.

Proposed method in this study is referred to as Multi-Stage method. Nevertheless, our method is also can be treated as a post processing method that can make matter easier to be tackled for post processing, such as binary masking for improvement of separation and SIMO deconvolution for removing reflection. It can be combined with almost all separation schemes seamlessly. In order to show its efficiency and superiority, different separation methods are first performed, and then, proposed method is applied as a post process to recover spectra of signal to some extent.

4.2. Results and discussion

Experiment 1: MBD with Nonholonomic constrain (Non MBD) method (9)–(10) is performed with learning rate 2×10^{-3} and separation matrix with $L = 64$ is initialized using null beamforming method, in which null direction is steered to $(-60^o, 60^o)$. Multi-Stage method is conducted as a post processing method combined with Non MBD method to show its efficiency.

Results are shown in Fig. 3 and table 1, from which we can see that resulting spectra of Non-MBD method is almost flat, whitening effect is relatively serious. Whereas, our proposed method improves SQ performance greatly, retrieved spectra are almost same as that of contribution of source to microphone. Tiny decrease of SNRI is mainly due to amplification of compensation matrix in low frequency range, which increases proportion of interference components simultaneously. In this sense, there is a trade-off between SNRI and SQ performance.

Experiment 2: To show superiority of proposed method, Comparison is carried out with SIMO-ICA method mentioned above. In order to test insensitivity to initialization, different Null Beamforming (NB) techniques are applied to initialize separation matrix and direction patterns of NB are assumed uniformly to be $(-60^o, 60^o)$. The one NB is that used in [12], which is referred to as NB-1 here. The other one is similar with first one except releasing constrain of target direction by setting direct path $w_{ii}(f) = 1, i = 1,2$ at every frequency bin and cross path $w_{ij}(f), i \neq j$ being calculated using null direction constrain. The second NB is referred to as NB-2 for simplicity.

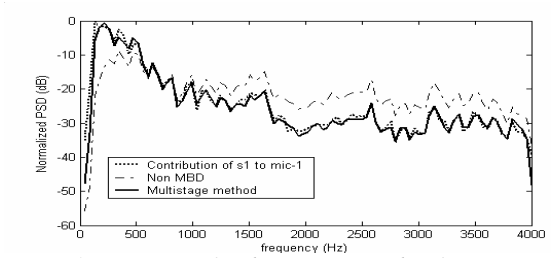

Figure 3 *normalized PSD compared with Non MBD*

Table 1. *Performance comparison. (dB)*

	Non MBD	Multistage
SNRI	13.71	12.35
SQ	0.29	8.22

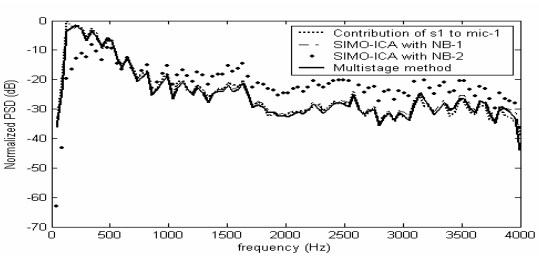

Figure 4 *normalized PSD compared with SIMO-ICA*

Table 2. *comparison with SIMO-ICA. (dB)*

	SNRI	SQ	Post-Pro SQ
NB-1	11.12	8.40	9.74
NB-2	12.12	1.81	8.96

Results shown in Fig 4 imply that SIMO-ICA method is highly sensitive to initialization. With 'good' initialization, SIMO-ICA can get satisfied SQ performance. However, from table 4, we known easily, Multi-stage method still owns superiority over SIMO-ICA method and it can improve SQ performance as a post processing method. Initialized with 'bad' value, SQ performance of SIMO-ICA declines seriously. Resorting to our method, SQ can be regained very well. As for later work of SIMO-ICA [13], proposed self-generator for initial filter can only be applied to HRTF case; therefore, comparison is not conduct here. Also, self-generator for conventional room transfer function will be considered in the future work.

From both experiments, it is found that separation in low frequency is still a challenge. Just as a result of bad performance of BSS in low frequency, SQ and SNRI is a pair of contradictory objective score, at least the case for speech signal.

5. Conclusions and future work

In this study, we proposed a novel and smart method to recover source to its contributions to every microphone. Until now, one step MBD have no enough ability to perform separation and deconvolution simultaneously well for signal with temporal structure, such as speech signal. Alternatively, two-stage, even multi-stages method is another suitable choice. Under such background, proposed method is valuable that can broad range of application. Validity has been shown in experiments and superiority also be demonstrated by comparison with other method. Moreover, simplicity of method makes it easily to be combined with other separation method naturally, either time domain algorithm or frequency domain algorithm.

In order to fully recover original signals without losing spectra characteristics, SIMO deconvolution is necessarily to remove reflection sequentially. Only in this way, can signals with temporal structure be expected to be deconvoluted fully.

6. References

[1] Simon Haykin., Unsupervised Adaptive Filtering; John Willey & Sons, INC, 2000.

[2] A. Hyvärinen, J. Karhunen and E. Oja., Independent Component Analysis; Wiley 2001.

[3] K.Torkkola., "Blind separation of convolved sources based on information maximization," Proc. IEEE Workshop Neural Networks for signal processing, pages 423-432,1996.

[4] S. Choi, et al., "Multichannel signal separation for cocktail party speech recognition: A dynamic recurrent network," Neurocomput., vol. 49, no. 1–4, pp. 299–314, Dec. 2002

[5] S. Amari, et al., "Non-holonomic constraints in learning algorithms for blind source separation," Neural Computation, 12:1463-1484, 2000

[6] X. Sun and S. C. Douglas., "A natural gradient convolutive blind source separation algorithm for speech mixtures," in Proc. 3rd Int. Conf. ICA and BSS, San Diego, CA, Dec. 9–13, 2001, pp. 59–64.

[7] K. Kokkinakis, Asoke K. Nandi., "Multichannel Blind Deconvolution for Source Separation in Convolutive Mixtures of Speech," IEEE Transactions on speech and audio processing, Vol. 14, No. 1, Jan 2006.

[8] T. Takatani, et al., "High-Fiedelity Blind Separation of Acoustic Signals Using SIMO-Model Based Independent Component Analysis," IEICE Trans Fundamentals, Vol. E87-A, No. 8, Aug 2004.

[9] K.Matsuoka, S.Nakashima., "Minimal distortion principle for blind source separation," Proc. International Conference on Independent Analysis and Blind Signal Separation, pp.722-727, Dec 2001.

[10] M. Miyaoshi et al., "Inverse Filtering of Room Acoustic", IEEE Trans. Acoutics, speech and signal process., Vol 36, no 2, pp-145-152,1988

[11] J. Allen and D. Berkeley, "Image method for efficiently simulating smallroom acoustics," Journal of the Acoustical Society of America, vol. 65, no. 4, pp. 943-950, April 1979.

[12] H. Saruwatari, et al., "Blind source separation combing independent component analysis and beamforming", EURASIP Journal on Applied Signal Processing, 1135-1146, Nov 2003.

[13] Tomoya Takatani, Satoshi Ukai, et al., "A self-generator method for initial filters of SIMO-ICA applied to blind separation of binaural sound mixtures," IEICE Trans. Fundamentals, Vol.E88-A, No.7, pp1673-1682, 2005