Time-Domain Blind Audio Source Separation using Advanced ICA Methods

Zbyněk Koldovský[1] *and Petr Tichavský*[2]

[1]Institute of Information Technology and Electronics,
Technical University of Liberec, Liberec, Czech Republic
[2]Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic, Czech Republic
zbynek.koldovsky@tul.cz, tichavsk@utia.cas.cz

Abstract

In this paper, a prototype of novel algorithm for blind separation of convolutive mixtures of audio sources is proposed. The method works in time-domain, and it is based on the recently very successful algorithm EFICA for Independent Component Analysis, which is an enhanced version of more famous FastICA. Performance of the new algorithm is very promising, at least, comparable to other (mostly frequency domain) algorithms. Audio separation examples are included.

Index Terms: independent component analysis, multichannel blind deconvolution, audio source separation

1. Introduction

Blind source separation (BSS) consists in extraction of individual signals from their mixture using no prior knowledge about their nature. Here, we address the blind separation of audio sources by means of Independent Component Analysis (ICA) [3], which is a popular method for BSS using the assumption that the original sources are mutually independent. In recent years, this field attracts large attention in audio signal processing due to various application areas such as crosstalk removal in multi-microphone recordings, speech enhancement, beamforming, direction of arrival estimation, etc. A typical application example is the so-called *cocktail party problem* where individual speeches should be extracted from mixtures of several speakers recorded in a common acoustic environment [1, 9, 5, 11].

The mixing process is described as a linear MIMO system

$$x_i(n) = \sum_{j=1}^{d} \sum_{\tau=0}^{M_{ij}} h_{ij}(\tau) s_j(n-\tau), \qquad (1)$$

where $x_1(n), \ldots, x_m(n)$ are the observed signals from sensors (microphones), $s_1(n), \ldots, s_d(n)$ are the unknown original (audio) signals, and h_{ij} are impulse responses of the system of the corresponding length M_{ij}, which is always finite in practice, but may be seriously long in reverberant environments. We will assume that $m \geq d$ (the same or more number of sensors than sources) due to the identifiability issues [1]. In practice, where finite number of samples $n = 1, \ldots, N$ is available, the goal of BSS is to find an FIR MIMO inverse of (1) in order to retrieve the original signals as best as is possible. Then, the estimated signals are

$$\widehat{s}_i(n) = \sum_{j=1}^{m} \sum_{\tau=0}^{L} w_{ij}(\tau) x_j(n-\tau), \qquad (2)$$

where $w_{ij}(n)$ $n = 0, \ldots, L$ are coefficients of the inverse, and L is the length of the inverse filters.

Direct utilization of ICA for the audio BSS is not possible, because most of ICA algorithms [4, 2] are designed for instantaneous mixtures, where no delays are considered ($M_{ij} = 0$ $\forall i, j$). Moreover, properties of audio sources often do not match assumptions required by models of signals (e.g., i.i.d. sequences) that are retrievable by means of standard blind deconvolution approaches.

Therefore, some reformulation or generalization is needed to allow successful application of methods for separation of instantaneous mixtures. In this fashion, the audio BSS can be performed via ICA either in frequency domain [5, 1, 11] or in time domain [14]. The former approach needs a successful solution of the so-called permutation problem, because the order of the signal components can be different in each frequency bin.

In this paper, we consider the separation in the time-domain and propose a new method for audio BSS that utilizes a powerful ICA algorithm EFICA [7] as a decomposition method for the time-domain separation. The method is demonstrated by several experiments with real audio mixtures.

2. Separation via ICA in the time-domain

The instantaneous mixture ($M_{ij} = 0$ $\forall i, j$) considered for the underlying problem of ICA can be written in a matrix form $\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n)$, where $\mathbf{x}(n)$ and $\mathbf{s}(n)$ are vectors of the observed signals $[x_1(n), \ldots, x_m(n)]^T$ and the original signals $[s_1(n), \ldots, s_m(n)]^T$, respectively, and $\mathbf{A}$ is a matrix of attenuations $\mathbf{A}_{ij} = h_{ij}(0)$.

A straightforward generalization of an ICA algorithm for blind deconvolution, that was primarily proposed for the instantaneous mixtures, is given when it applies on data

$$\widetilde{\mathbf{x}}(n) = [x_1(n), x_1(n-1), \ldots, x_1(n-L), x_2(n), \ldots, x_m(n-L)]^T \qquad (3)$$

see e.g. [14]. Under the assumption that the original data are i.i.d. non-gaussian sequences, and L is large enough, it can be shown that the resulting independent components are estimates of re-ordered, attenuated, and delayed copies of the original signals. It is also a well-known fact that being the original

[0]This work was partly supported by the Czech Science Foundation (GA ČR) through projects 102/05/0278 and 102/07/P384 and by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M0572.

sources weak-sense stationary processes, the ICA decomposition of (3) results in innovation sequences of the original signals [14]. In these respects, we highlight good properties of algorithm EFICA [7] that yields excellent results in simulations with artificial data.

In real applications with audio signals, none of the previous assumptions is exactly fulfilled. For instance, the original signals are not stationary, or the length of the inverse filters L in (2) is not large enough for exact inversion of (1). Our experiments show that, in case of audio sources, the independent components of (3) are strongly filtered copies of the original signals, however, useless for being their estimates. Luckily, there are always groups of components that correspond to individual sources, with suppressed inter-group interference. This gives rise to the possibility to apply a clustering-reconstruction mechanism to retrieve the original signals in some useful quality.

3. Proposal of the Separation Procedure

Our proposal of the method that performs audio BSS in the time-domain follows the fashion described in previous section.

In the first stage, it applies algorithm EFICA on data $\widetilde{\mathbf{x}}(n)$, which yields $m(L+1)$ independent components $\mathbf{c}(n) = [c_1(n), \ldots, c_{m(L+1)}(n)]^T$ via the resulting decomposition matrix $\mathbf{W}$, i.e.

$$\mathbf{c}(n) = \mathbf{W}\widetilde{\mathbf{x}}(n). \tag{4}$$

The main idea of the algorithm relies on the hypothesis that the independent components in $\mathbf{c}(n)$ can be grouped so that each group contains filtered versions of one original signal only. A relation of components should be defined through some distance measure, and the grouping may be done via clustering.

The distance between the independent components (say between the i-th and $j-$th component) can be measured as

$$D_{ij} = \hat{\mathrm{E}}[\mathbf{P}_i c_j(n)]^2, \tag{5}$$

where $\mathbf{P}_i$ denotes a projector on subspace spanned by $[c_i(n-L), \ldots, c_i(n+L)]$, and $\hat{\mathrm{E}}$ denotes the sample mean operator.

In other words, defining vectors $\mathbf{c}_i$ and time shift operator D as

$$\mathbf{c}_i = \begin{bmatrix} c_i(L+1) \\ c_i(L+2) \\ \vdots \\ c_i(N-L-1) \\ c_i(N-L) \end{bmatrix}, \; D^k\mathbf{c}_i = \begin{bmatrix} c_i(L+k+1) \\ c_i(L+k+2) \\ \vdots \\ c_i(N-L+k-1) \\ c_i(N-L+k) \end{bmatrix}$$

for $k = -L, \ldots, L$, and a matrix

$$\mathbf{C}_i = [D^{-L}\mathbf{c}_i, D^{-L+1}\mathbf{c}_i, \ldots, D^{L-1}\mathbf{c}_i, D^L\mathbf{c}_i],$$

then the "distance" D_{ij} (that is not symmetric, in general) can be written as

$$D_{ij} = \|\mathbf{c}_j - \mathbf{C}_i(\mathbf{C}_i^T\mathbf{C}_i)^{-1}\mathbf{C}_i^T\mathbf{c}_j\|^2. \tag{6}$$

Here, it was assumed that the independent components were normalized to have unit variance.

Now, the distance may be symmetrized (or not) and used to group the individual components to clusters. The selection of a proper clustering algorithm is still subject to an intensive research. We have used standard agglomerative hierarchical clustering with an average linking strategy [13]. See example of clustered matrix of symmetrized distances (5) in Figure 1.

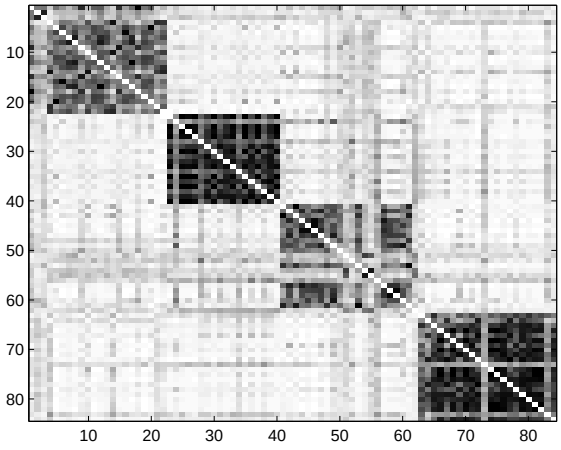

Fig.1 The matrix of distances (5) of clustered independent components that were obtained by EFICA ($L = 20$) from Sawada's mixture of four sources.

Once the independent components were clustered, each cluster, say j-th, is used for reconstruction of the corresponding original signal. Let $\mathbf{c}^{(j)}(n)$ be equal to $\mathbf{c}(n)$ but those rows that do *not* correspond to the components from the j-th cluster are zero. Then, we define reconstructed signals

$$\widetilde{\mathbf{x}}^{(j)}(n) = \mathbf{W}^{-1}\mathbf{c}^{(j)}(n)\,. \tag{7}$$

Such signals are (in an ideal case) delayed versions of the j-th source received on all microphones; see (3), (4), and the definition of $\mathbf{c}^{(j)}(n)$. We constructively sum the delayed versions of the signal on each microphone separately, so that the contribution of the source on the $i-$th microphone is

$$\widetilde{s}_i^{(j)}(n) = \sum_{p=1}^{L+1} \widetilde{\mathbf{x}}^{(j)}_{(i-1)(L+1)+p}(n+p-1) \tag{8}$$

The final estimator of the j-th source $\widehat{s}_j(n)$ is obtained as a simple delay-and-sum beamformer, that forms contributions of the j-th source (8) on all microphones together.

The whole convolutive BSS algorithm can be summarized as follows.

1. For a suitable number of delays L, form the auxiliary data matrix $\widetilde{\mathbf{x}}(n)$ in (3)
2. Apply EFICA on $\widetilde{\mathbf{x}}(n)$, and obtain the decomposition matrix $\mathbf{W}$.
3. Compute the distances in (6), and form clusters among the estimated components.
4. For each cluster reconstruct $\widetilde{\mathbf{x}}^{(j)}(n)$ in (7), then sum the delayed versions of $\widetilde{\mathbf{x}}^{(j)}(n)$ for each microphone separately by (8), and finally compute $\widetilde{s}_j(n)$ by combining outcome of all microphones.

4. Performance Measurement

In this section, we briefly describe performance measures that will be used for evaluation of separation quality or quality of components obtained by an ICA algorithm.

In blind scenarios, the exact performance evaluation is possible only if true system parameters are known, i.e., in simulations with artificially mixed data. This is not possible in real audio BSS experiments since the impulse response of the room is not known. In spite of being more or less measurable, a more convenient approach is to evaluate the performance via knowing the original sources. Therefore, in real simulations of audio BSS, the playback via loudspeakers is used instead of real human voices; see the simulations section.

The validation of the separation can be done by comparing the original sources with the acquired ones. Still this is a difficult task due to lack of objective measures, because there are several distortions in the estimated signals such as interferences, artifacts, or noise.

In this paper, we utilize a straightforward approach proposed in [16] that consists in numerical projections of the separated signals on subspaces, where each subspace is spanned by delayed duplicates of a corresponding original signal. Thus, each estimated signal $\widehat{s}_j(n)$ is decomposed as

$$\widehat{s}_j(n) = s_{\text{target}}(n) + e_{\text{interf}}(n) + e_{\text{artif}}(n), \tag{9}$$

where $s_{\text{target}}(n)$ corresponds to the original signal up to allowed distortions, and $e_{\text{interf}}(n)$ and $e_{\text{artif}}(n)$ are, respectively, the interferences and artifacts error terms. Here, we do not consider any additive noise. Two criteria, i.e. signal-to-distortion ratio and signal-to-interference ratio, will be used throughout the paper:

$$\text{SDR}_j = 10\log_{10}\frac{\|s_{\text{target}}(n)\|^2}{\|e_{\text{interf}}(n)+e_{\text{artif}}(n)\|^2} \tag{10}$$

$$\text{SIR}_j = 10\log_{10}\frac{\|s_{\text{target}}(n)\|^2}{\|e_{\text{interf}}(n)\|^2}. \tag{11}$$

The criteria are computed by means of functions bss_decomp_filt and bss_crit from BSS_EVAL Toolbox; see [16]. We select the dimension 256 of the projection subspaces, i.e. number of delays.

5. Experimental Results

In this section, results of real experiments with acoustical signals are described and demonstratively compared with other methods whose matlab code is available on the internet, namely, we tried BSS algorithm of Parra and Spence [9][1] and the frequency-domain algorithm FDICA [10][2].

5.1. Separation of own data

Original signals (two man's voices, a woman's voice, a Gaussian noise of SNR=0dB, and a typewriter sound) were played from two loudspeakers (one source from one speaker) and recorded by two microphones that were placed about 1.5m far from the speakers and 15cm far from each other. The length of the recorded data was $N = 18000$, and the length of data used for estimation of the ICA decomposition (in the second step of the proposed method) was 8000 samples.

Four combinations of the original sources were recorded at sampling rate 8kHz and separated by the proposed method (taking $L = 20$) and the Parra's and FDICA algorithm taking their default settings. For illustration, values of (11) and (10) achieved by the proposed method and Parra's algorithm are shown in Table 1. All sound files that contain the results from separations are included in the DVD proceedings.

[1] http://ida.first.fraunhofer.de/~ harmeli/download/download_convbss.html

[2] http://tsi.enst.fr/icacentral/Algos/prasad/Bss4Speech.zip

	Proposed method		Parra's algorithm	
	SIR_1 / SIR_2	SDR_1 / SDR_2	SIR_1 / SIR_2	SDR_1 / SDR_2
man's voice	10.44	6.1	6.16	4.64
man's voice	5.63	2.48	5.44	1.38
man's voice	2.16	5.98	9.79	2.97
woman's voice	4.11	1.67	6.97	4.12
man's voice	14.19	6.71	8.45	4.76
Gaussian noise	9.43	5.87	11.34	8.65
man's voice	17.89	6.81	7.82	2.69
typewriter	12.22	8.88	11.97	9.50

Table 1: Values of criteria (11) and (10) in dB of each separated source from corresponding mixture.

The table shows that the performance of the proposed approach is comparable with that of Parra's algorithm, however, note that the values of the criteria give directory information only due to difficulties of objective evaluation. For instance, performance of the proposed method in case of the mixture of man's voice with the Gaussian noise seems to be better, but significantly poorer performance is achieved when separating woman's and man's voices. However, listening to the extracted signals shows similar performance; the algorithm FDICA yields signals disturbed by numerous artifacts, which is likely due to the permutation problem [11].

Computational burden of the proposed method strongly depends on the length of the inversion filter L. For instance, when separating the mixture of two man's voices, the computation time for L equal to 5, 10, 20, and 40 was, respectively, 4s, 6.1s, 18.2s, and 53s on a standard PC with 3GHz processor and 2GB of RAM. Parra's algorithm with the default settings needed about 10.4s.

5.2. Separation of public data

Here, we present results of separation of several data that are available on the internet. For objective comparison, we provide the results for audition in DVD Proceedings of this conference. The data include Te-Won Lee's [8] signals that are commonly used for demonstrations (two sources mixtures: speech-music and speech-speech mixtures), and the data provided by Hiroshi Sawada [12], where mixtures of 2, 3, and 4 sources from the same number of microphones are provided.

As an illustration of ability of the proposed method to separate the four sources from Sawada's recordings, Figure 1 shows clustered matrix that well reveals clusters corresponding to individual signals, and, in Figure 2, the separated signals are shown. In this example, the performance of the proposed method is evidently superior to that of Parra's algorithm; the available implementation of FDICA does not allow separation of more than two sources.

6. Conclusions

In this paper, we propose a prototype of method for audio source separation, which is based on ICA applied in the time-domain, that can be done via any appropriate algorithm. Here, we rely on fast and accurate algorithm EFICA working well even in high dimensions with short length of data. A novel clustering-reconstruction mechanism utilizing the independent components for original signals retrieval is proposed, and it yields good results in real experiments even for small length of the separating filters L. Thanks to the transparency of the

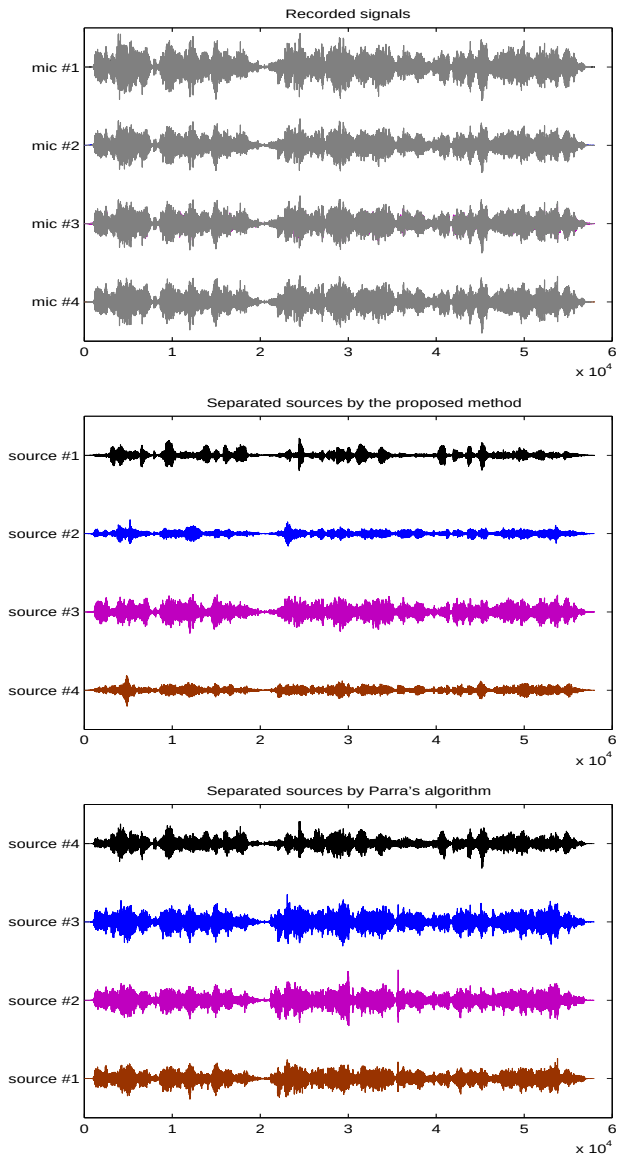

Fig.2 The recorded and the separated Sawada's 4-sources recordings with $L = 20$.

time-domain concept, the approach offers many possibilities for improvements that are subject to our future work. In speech separation, the method can be further extended by use of time-frequency masking [6, 5].

7. References

[1] Buchner, H., Aichner, R., and Kellermann, W.: "A Generalization of Blind Source Separation Algorithms for Convolutive Mixtures Based on Second-Order Statistics", *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 1, pp. 120-134, January 2005.

[2] Cardoso, J.-F.: "High-order Contrasts for Independent Component Analysis", *Neural Computat.*, vol. 11, no. 1, pp. 157-192, 1999.

[3] Hyvärinen, A., Karhunen, J., and Oja, E.: *Independent Component Analysis*, Wiley-Interscience, New York, 2001.

[4] Hyvärinen, A., and Oja, E.: "A Fast Fixed-Point Algorithm for Independent Component Analysis", *Neural Computation*, 9(7):1483-1492, 1997.

[5] Mori, Y., Saruwatari, H., Takatani, T., Ukai, S., Shikano, K., Hiekata, T., Ikeda, Y., Hashimoto, H., and Morita, T.: "Blind Separation of Acoustic Signals Combining SIMO-Model-Based Independent Component Analysis and Binary Masking", *EURASIP Journal on Applied Signal Processing*, Vol. 2006, Article ID 34970, 17 pages, 2006.

[6] Koldovský, Z., Nouza, J., and Kolorenč, J.: "Continuous Time-Frequency Masking Method for Blind Speech Separation with Adaptive Choice of Threshold Parameter Using ICA", *Proc. of Interspeech 2006*, Pittsburgh PA, USA, 17.-21. September, pp. 2578-2581, 2006.

[7] Koldovský, Z., Tichavský, P., and Oja, E.: Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound, *IEEE Tr. Neural Networks*, **17** (2006) 1265–1277.

[8] Te-Won Lee's demo pages [online], http://www.cnl.salk.edu/˜ tewon/Blind/blind_audio.html.

[9] Parra, L., and Spence, C.: "Convolutive Blind Separation of Non-Stationary Sources", *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 3, pp. 320-327, May 2000.

[10] Prasad, R., Saruwatari, H., and Shikano, K.: "Blind Separation of Speech by Fixed-Point ICA with Source Adaptive Negentropy Approximation", *IEICE-Tran. Fund. Elec., Comm. & Comp. Sci.*, Vol. E88-A, Num. 7, pp. 1683-1692, July 2005.

[11] Sawada, H., Mukai, R., Araki, S., and Makino, S.: "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation", *IEEE Trans. Speech Audio Processing*, Vol. 12, No. 5, pp. 530-538, Sept. 2004.

[12] Hiroshi Sawada's demo pages [online], http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html.

[13] Späth, H.: *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, Chichester, Ellis Horwood 1980.

[14] Thomas, J., Deville, Y., Hosseini, S.: "Time-domain Fast Fixed-Point Algorithms for Convolutive ICA", *IEEE Signal Processing Letters*, **13**(4) (2006) 228–231.

[15] Tichavský, P., Koldovský, Z., and Oja, E.: "Speed and accuracy enhancement of linear ICA techniques using rational nonlinear functions", to-be published in *Proc. of ICA 2007*.

[16] Vincent, E., Gribonval, R., and Févotte, C.: "Performance Measurement in Blind Audio Source Separation", *IEEE Trans. on Speech and Audio Processing*, Vol 14, No 4, pp. 1462 - 1469, July 2006.

[17] Yilmaz, Ö, and Rickard, S., "Blind Separation of Speech Mixtures via Time-Frequency Masking", *IEEE Trans. on Signal Processing*, vol. 52, no. 7, July 2004.