Time-Domain Blind Audio Source Separation Method Producing Separating Filters of Generalized Feedforward Structure[*]

Zbyněk Koldovský[1,2], Petr Tichavský[2], and Jiří Málek[1]

[1] Institute of Information Technology and Electronics
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
zbynek.koldovsky@tul.cz
http://itakura.ite.tul.cz/zbynek

[2] Institute of Information Theory and Automation, Pod vodárenskou věží 4,
P.O. Box 18, 182 08 Praha 8, Czech Republic
p.tichavsky@ieee.org
http://si.utia.cas.cz/Tichavsky.html

Abstract. Time-domain methods for blind separation of audio signals are preferred due to their lower demand for available data and the avoidance of the permutation problem. However, their computational demands increase rapidly with the length of separating filters due to the simultaneous growth of the dimension of an *observation space*. We propose, in this paper, a general framework that allows the time-domain methods to compute separating filters of theoretically infinite length without increasing the dimension. Based on this framework, we derive a generalized version of the time-domain method of Koldovský and Tichavský (2008). For instance, it is demonstrated that its performance might be improved by 4dB of SIR using the Laguerre filter bank.

1 Introduction

Blind Audio Source Separation (BASS) aims at separating unknown audio sources, which are mixed in an acoustical environment according to the convolutive model. The observed mixed signals are

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

where $\star$ denotes the convolution, m is the number of microphones, $s_1(n), \ldots, s_d(n)$ are the original sources, and h_{ij} are source-microphone impulse responses each of length M_{ij}. The linear separation consists in finding de-mixing filters that separate original sources in its outputs. Since many methods for finding the filters

[*] This work was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M0572 and by Grant Agency of the Czech Republic through the project 102/09/1278.

V. Vigneron et al. (Eds.): LVA/ICA 2010, LNCS 6365, pp. 17–24, 2010.

formally assume instantaneous mixtures, i.e., $M_{ij} = 1$ for all i, j, the convolutive model needs to be transformed. This can be done either in the frequency or time domain.

Time-domain approaches, addressed in this paper, consist in decomposing the *observation matrix* defined as [1]

$$\mathbf{X} = \begin{bmatrix} x_1(N_1) & \ldots\ldots & x_1(N_2) \\ x_1(N_1-1) & \ldots\ldots & x_1(N_2-1) \\ \vdots & \vdots\ \vdots & \vdots \\ x_1(N_1-L+1) & \ldots\ldots & x_1(N_2-L+1) \\ x_2(N_1) & \ldots\ldots & x_2(N_2) \\ x_2(N_1-1) & \ldots\ldots & x_2(N_2-1) \\ \vdots & \vdots\ \vdots & \vdots \\ x_m(N_1-L+1) & \ldots\ldots & x_m(N_2-L+1) \end{bmatrix}, \tag{2}$$

where N stands for the number of available samples, and $1 \leq N_1 < N_2 \leq N$ determine the segment of data used for computations, and L is a free parameter. The decomposition of $\mathbf{X}$ is done by multiplying it by a matrix $\mathbf{W}$. This way FIR filters of the length L whose elements correspond to rows of $\mathbf{W}$ are applied to the mixed signals $x_1(n), \ldots, x_m(n)$. This is due to the structure of $\mathbf{X}$ given by (2). The subspace of dimension mL in $\mathbb{R}^{N_2-N_1+1}$ spanned by rows of $\mathbf{X}$ will be called the *observation space.*

It is desired to decompose the observation space into linear subspaces where each of them represents one original signal. It can be done either by some independent subspace analysis (ISA) technique or by an independent component analysis (ICA) method, which is followed by the clustering of the components [2]. Performance of some ISA and ICA methods was studied in [12]. Some other methods utilize block-Sylvester structure of $\mathbf{A} = \mathbf{W}^{-1}$ [1,4]. Computational complexity of all these methods increases most ideally with L^3, which means that L cannot be too large. On the other hand, the frequency response of ordinary rooms is typically several hundreds of taps [3]. Therefore, longer filters would be desired.

Longer separating filters can be obtained by the subband-based separation [3,5]. In this paper, however, we propose to increase the length of the separating filters by changing the definition of the observation space. For a given set of invertible filters $f_{i,\ell}$, $\mathbf{X}$ is defined as

$$\mathbf{X} = \begin{bmatrix} \{f_{1,1} \star x_1\}(N_1) & \ldots\ldots & \{f_{1,1} \star x_1\}(N_2) \\ \{f_{1,2} \star x_1\}(N_1) & \ldots\ldots & \{f_{1,2} \star x_1\}(N_2) \\ \vdots & \vdots\ \vdots & \vdots \\ \{f_{1,L} \star x_1\}(N_1) & \ldots\ldots & \{f_{1,L} \star x_1\}(N_2) \\ \{f_{2,1} \star x_2\}(N_1) & \ldots\ldots & \{f_{2,1} \star x_2\}(N_2) \\ \vdots & \vdots\ \vdots & \vdots \\ \vdots & \vdots\ \vdots & \vdots \\ \{f_{m,L} \star x_m\}(N_1) & \ldots\ldots & \{f_{m,L} \star x_m\}(N_2) \end{bmatrix}. \tag{3}$$

Linear combinations of rows of $\mathbf{X}$ defined in this way correspond to outputs of MIMO filters with a generalized feed-forward structure introduced in [8], where the filters $f_{i,\ell}$ are referred to as *eigenmodes*. Note that if $f_{i,\ell}$ realizes backward time-shift by $\ell - 1$ samples, i.e. $f_{i,\ell}(n) = \delta(n - \ell + 1)$, where $\delta(n)$ stands for the unit impuls function, the construction of $\mathbf{X}$ given by (3) coincides with (2)[1].

The proposed definition (3) extends the class of filters that are applied to signals $x_1(n), \ldots, x_m(n)$ when multiplying $\mathbf{X}$ by $\mathbf{W}$. Time-domain BSS methods searching $\mathbf{W}$ via ICA can thus apply long separating (even IIR) filters without increasing L.

When $\mathbf{X}$ is defined by (2), $\mathbf{A}$ or $\mathbf{W}$ can be assumed to have a special structure (e.g. block-Sylvester) [1,2,4]. In general, the structure does not exist if $\mathbf{X}$ is defined according to (3). It is necessary to apply a separating algorithm that does not rely on the special structure - such as the method from [6,7], referred to as T-ABCD[2]. An extension of T-ABCD working with $\mathbf{X}$ defined through (3) is proposed in the following section. Then, a practical version of T-ABCD using Laguerre eigenmodes is proposed in Section 3, and its performance is demonstrated by Section 4. In Section 5, we present a semi-blind approach to show another potential of the generalized definition of $\mathbf{X}$.

2 Generalized T-ABCD

2.1 The Original Version of T-ABCD

Following the minimal distortion principle, T-ABCD estimates microphone responses of the original signals, $s_k^i(n) = \{h_{ik} \star s_k\}(n)$, $i = 1, \ldots, m$, which are signals measured on microphones when the kth source sounds solo. First, we briefly describe the original version of T-ABCD from [6] that proceeds in four main steps.

1. Form the observation matrix $\mathbf{X}$ as in (2).
2. Decompose $\mathbf{X}$ into independent components, i.e., compute the $M \times M$ decomposing matrix $\mathbf{W}$ by an ICA algorithm, $M = mL$.
3. Group the components (rows of) $\mathbf{C} = \mathbf{WX}$ into clusters so that each cluster contains components that correspond to the same original source.
4. For each cluster, use only components of the cluster to estimate microphone responses of a source corresponding to the cluster.

The details of the fourth step are as follows. For the kth cluster,

$$\widehat{\mathbf{S}_k} = \mathbf{W}^{-1}\mathrm{diag}[\lambda_1^k, \ldots, \lambda_M^k]\,\mathbf{W}\,\mathbf{X} = \mathbf{W}^{-1}\mathrm{diag}[\lambda_1^k, \ldots, \lambda_M^k]\,\mathbf{C}, \tag{4}$$

[1] A further practical generalization is if different number of eigenmodes were considered for a given i, that is $f_{i,\ell}$ for $\ell = 1, \ldots, L_i$. For simplicity, we will consider the case $L_1 = \cdots = L_m = L$ only.

[2] Time-domain Audio sources Blind separation based on the Complete Decomposition of the observation space.

where $\lambda_1^k, \dots, \lambda_M^k$ denote positive weights from $[0, 1]$, reflecting degrees of affiliation of components to the kth cluster. Ideally, $\widehat{\mathbf{S}_k}$ is equal to $\mathbf{S}_k$, which is a matrix defined in the same way as $\mathbf{X}$ but consists of the contribution of only the kth source, which is, of the time-shifted copies of the responses $s_k^1(n), \dots, s_k^m(n)$. Note that since $x_i(n) = s_1^i(n) + \cdots + s_d^i(n)$, it holds that $\mathbf{X} = \mathbf{S}_1 + \cdots + \mathbf{S}_d$.

Taking the structure of $\mathbf{S}_k$ (the same as (2)) into account, the microphone responses are estimated from $\widehat{\mathbf{S}_k}$ as

$$\widehat{s}_k^i(n) = \frac{1}{L}\sum_{\ell=1}^{L} \psi_{k,(i-1)L+\ell}(n + \ell - 1), \tag{5}$$

where $\psi_{k,p}(n)$ is equal to the (p, n)th element of $\widehat{\mathbf{S}}_k$. To clarify, note that $\psi_{k,p}(n)$ provides an estimate of $s_k^i(n - \ell + 1)$ for $p = (i - 1)L + \ell$. See [6] for further details on the method[3].

2.2 Generalization

In the first step of generalized T-ABCD, $\mathbf{X}$ is constructed according to (3). Further steps of the method are the same as described above up to the reconstruction formula given by (5), which is given as follows.

Let $f_{i,\ell}^{-1}$ be the inverse of the filter $f_{i,\ell}$. As $\psi_{k,p}(n)$ defined by the (p, n)th element of $\widehat{\mathbf{S}_k}$, $p = (i - 1)L + \ell$, provides an estimate of $\{f_{i,\ell} \star s_k^i\}(n)$, the microphone responses of the kth separated source are estimated as

$$\widehat{s}_k^i(n) = \frac{1}{L}\sum_{\ell=1}^{L} \{f_{i,\ell}^{-1} \star \psi_{k,(i-1)L+\ell}\}(n). \tag{6}$$

Obviously, (6) coincides with (5) if $f_{i,\ell}(n) = \delta(n - \ell + 1)$.

3 T-ABCD Using Laguerre Filters

In [9,10], Laguerre filters having the feed-forward structure [8] were shown to yield better separation than the ordinary FIR filters, apparently, thanks to increased effective length of their impulse response for certain values of a parameter μ. These filters can be applied within T-ABCD when the eigenmodes $f_{i,\ell}$ in (3) (now we may omit the first index i) are defined through their transfer functions F_ℓ recursively as

$$F_1(z) = 1, \tag{7}$$

$$F_2(z) = \frac{\mu z^{-1}}{1 - (1 - \mu)z^{-1}}, \tag{8}$$

$$F_n(z) = F_{n-1}(z)G(z), \quad n = 3, \dots, L, \tag{9}$$

[3] Note the missing factor $1/L$ in the formula (9) in [6].

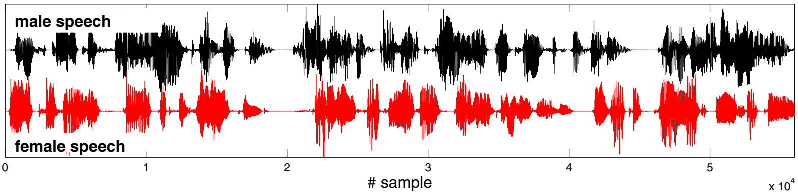

Fig. 1. Original signals used in experiments

where

$$G(z) = \frac{(\mu - 1) + z^{-1}}{1 - (1 - \mu)z^{-1}}, \tag{10}$$

and μ takes values from $(0, 2)$. Note that f_2 is either a low-pass filter (for $0 < \mu < 1$) or a high-pass filter (for $1 < \mu < 2$), and g is an all-pass filter.

The construction of $\mathbf{X}$ through Laguerre eigenmodes embodies (2) as a special case, because for $\mu = 1$, $F_2(z) = G(z) = z^{-1}$, that is $f_2(n) = g(n) = \delta(n - 1)$, consequently, $f_\ell(d) = \delta(n - L + 1)$. This is the only case where the Laguerre filters are FIR of the length L. For $\mu \neq 1$, the filters are IIR.

The effective length of the Laguerre filters denoted by L_* is defined as the minimum length needed to capture 90% of the total energy contained in the impulse response. For the Laguerre filters it approximately holds that [10]

$$L_* = (1 + 0.4|\mu - 1| \log_{10} L)L/\mu. \tag{11}$$

We can see that $L_* > L$ for $\mu < 1$ and vice versa. From here on, we will refer to T-ABCD as the variant proposed in this section as it encompasses the original algorithm when $\mu = 1$.

4 Experiments with Real-World Recordings

The proposed algorithm will be tested in the SiSEC evaluation campaign. The experiments in this paper examine mixtures of Hiroshi Sawada's original signals, which are available on the Internet[4]. The data are a male and a female utterance of the length 7 s recorded at the sampling rate 8kHz; see Fig. 1. For evaluations, we use two standard measures as in [13]: Signal-to-Interference Ratio (SIR) and Signal-to-Distortion Ratio (SDR). The SIR determines the ratio of energies of the desired signal and the interference in the separated signal. The SDR provides a supplementary criterion of SIR that reflects the difference between the desired and the estimated signal in the mean-square sense.

The performance of T-ABCD defined in the previous section was tested by separating Sawada's recordings of the original signals that were recorded in a room with the reverberation time of 130ms using two closely spaced microphones and two loudspeakers placed at a distance of 1.2 m. T-ABCD was applied to

[4] `http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html`

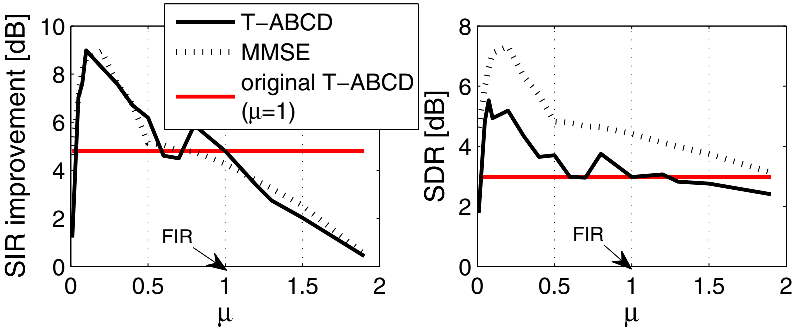

Fig. 2. Results of separation of Sawada's real-world recordings

separate the recordings with $L = 20$ and varying μ. Two seconds of the data were used for computations of separating filters, i.e., $N_1 = 1$ and $N_2 = 16000$. The ICA algorithm applied within T-ABCD is BGSEP from [11] that is based on the approximate joint diagonalization of covariance matrices computed on blocks of $\mathbf{X}$ (we consider blocks of 300 samples). The weighting parameter α for determining weights in (4) was set to 1. A similar setting was used in [6].

For comparison, minimum mean-square error (MMSE) solutions were computed as the best approximations of known responses of signals in the observation space defined by $\mathbf{X}$. It means that the MMSE solutions achieve the best SDR for given L and thus provide an experimental performance bound [10].

Fig. 2 shows resulting values of SIR and SDR averaged over both separated responses of both signals. The potential of Laguerre filters to improve the separation for $\mu < 1$ is demonstrated by the performance of the MMSE separator both in terms of SIR and SDR; similar results were observed in experiments in [10]. T-ABCD improves its performance when μ approaches 0.1 as well, with the optimum at around $\mu = 0.2$. For μ very close to zero ($\mu < 0.1$), the performance usually becomes unstable. Compared to the case $\mu = 1$, where $\mathbf{X}$ coincides with (2) and the separating filters are FIR, the separation is improved by 4dB of SIR and 2dB of SDR. This is achieved at essentially the same computational time (about 1.1 s in Matlab version 7.9 running on a PC, 2.6GHz, 3GB RAM), because the value of μ does not change the dimension of $\mathbf{X}$.

5 Semi-Blind Separation

The goal of this section is to provide another definition example of eigenmodes in (3) that utilizes prior information about the mixing system, otherwise known as the semi-blind approach. Consider the general $m = 2$ and $d = 2$ scenario

$$x_1(n) = \{h_{11} \star s_1\}(n) + \{h_{12} \star s_2\}(n) \tag{12}$$

$$x_2(n) = \{h_{21} \star s_1\}(n) + \{h_{22} \star s_2\}(n). \tag{13}$$

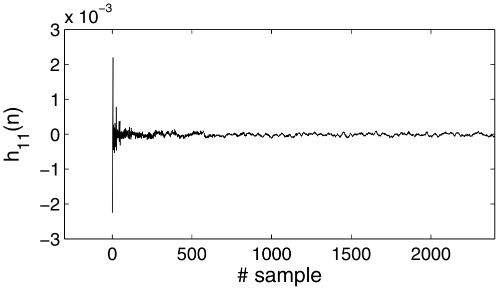

Fig. 3. The microphone-source impulse response $h_{11}(n)$

Almost perfect separation of this mixture can be achieved when taking $L = 2$ and defining $f_{11} = b \star h_{22}$, $f_{12} = -b \star h_{21}$, $f_{21} = -b \star h_{12}$, and $f_{22} = b \star h_{11}$, where $b = (h_{11} \star h_{22} - h_{21} \star h_{12})^{-1}$ assuming that the inversion exists. A trivial verification shows that combinations of signals $\{f_{11} \star x_1\}(n) + \{f_{21} \star x_2\}(n)$ and $\{f_{12} \star x_1\}(n) + \{f_{22} \star x_2\}(n)$ are independent, because they are equal to the original sources s_1 and s_2, respectively. If these combinations were unknown (e.g. when $f_{11}, \ldots, f_{22}$ were known up to a multiple by a constant), we could identify them blindly as independent components of $\mathbf{X}$ that would be defined through (3) with the eigenmodes $f_{11}, \ldots, f_{22}$. The dimension of such $\mathbf{X}$ is only 4, so the computation of ICA is very fast.

Additionally, we can define $f_{11}, \ldots, f_{22}$ with an arbitrary b, e.g., $b(n) = \delta(n)$. Note that b only affects the spectra of independent components of $\mathbf{X}$.

To demonstrate this, we recorded impulse responses of the length 300ms in a lecture room and mixed the original signals from Fig. 1 according to (12)-(13). An example of the recorded impulse response $h_{11}(n)$ is shown in Fig. 3.

The observation matrix $\mathbf{X}$ was constructed as described above with $b(n) = \delta(n)$. BGSEP was applied to $\mathbf{X}$ using only the first second of the recordings ($N_1 = 1$, $N_2 = 8000$) and yielded randomly permuted independent components of $\mathbf{X}$. Signal-to-Interference ratios of two of four components were, respectively, 28.3 dB subject to the male speech and 18.4 dB subject to the female speech, SIRs that represent a highly effective separation.

In comparison, MMSE solutions obtained by optimum FIR filters of the length 20 ($L = 20$ and $\mu = 1$) achieve only 4.8 dB of average SIR subject to the male speech and 6.8 dB subject to the female speech. Although the independent components have different coloration then the original signals (they are close to twice reverberated original signals by the room impulse response), the example reveals the great potential of the general construction of $\mathbf{X}$ in theory. For instance, it is indicative of the possibility to tailor the eigenmodes $f_{i,\ell}$ to room acoustics if the impulse response of the room can be measured with sufficient accuracy.

6 Conclusions

We have proposed a general construction of the observation matrix $\mathbf{X}$ that allows for the application of long separating filters in time-domain BASS methods without increasing the dimension of the observation space. This approach preserves the computational burden as it mostly depends on that dimension. The T-ABCD method was generalized in this way, and its version using Laguerre separating filters was shown to improve the separation with $\mu < 1$, i.e., when the effective length of separating filters L_* is increased compared to ordinary FIR filters with the length L. Future research can be focused on optimizing the choice of the eigenmodes.

References

1. Buchner, H., Aichner, R., Kellermann, W.: A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. IEEE Trans. on Speech and Audio Proc. 13(1), 120–134 (2005)
2. Févotte, C., Debiolles, A., Doncarli, C.: Blind separation of FIR convolutive mixtures: application to speech signals. In: 1st ISCA Workshop on Non-Linear Speech Processing (2003)
3. Araki, S., Makino, S., Aichner, R., Nishikawa, T., Saruwatari, H.: Subband-based blind separation for convolutive mixtures of speech. IEICE Trans. Fundamentals E88-A(12), 3593–3603 (2005)
4. Xu, X.-F., Feng, D.-Z., Zheng, W.-X., Zhang, H.: Convolutive blind source separation based on joint block Toeplitzation and block-inner diagonalization. Signal Processing 90(1), 119–133 (2010)
5. Koldovský, Z., Tichavský, P., Málek, J.: Subband blind audio source separation using a time-domain algorithm and tree-structured QMF filter bank. In: Vigneron, V., et al. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 25–32. Springer, Heidelberg (2010)
6. Koldovský, Z., Tichavský, P.: Time-domain blind audio source separation using advanced component clustering and reconstruction. In: HSCMA 2008, Trento, Italy, vol. 2008, pp. 216–219 (2008)
7. Koldovský, Z., Tichavský, P.: Time-Domain blind separation of audio sources on the basis of a complete ICA decomposition of an observation space. Accepted for Publication in IEEE Trans. on Audio, Language, and Speech Processing (April 2010)
8. Principe, J.-C., de Vries, B., de Oliveira, G.: Generalized feedforward structures: a new class of adaptive filters. In: ICASSP 1992, vol. 4, pp. 245–248 (1992)
9. Stanacevic, M., Cohen, M., Cauwenberghs, G.: Blind separation of linear convolutive mixtures using orthogonal filter banks. In: ICA 2001, San Diego, CA (2001)
10. Hild II, K.-E., Erdogmuz, D., Principe, J.-C.: Experimental upper bound for the performance of convolutive source separation methods. IEEE Trans. on Signal Processing 54(2), 627–635 (2006)
11. Tichavský, P., Yeredor, A.: Fast approximate joint diagonalization incorporating weight matrices. IEEE Transactions of Signal Processing 57(3), 878–891 (2009)
12. Koldovský, Z., Tichavský, P.: A comparison of independent component and independent subspace analysis algorithms. In: EUSIPCO 2009, Glasgow, England, pp. 1447–1451 (2009)
13. Schobben, D., Torkkola, K., Smaragdis, P.: Evaluation of blind signal separation methods. In: ICA 1999, Aussois, France, pp. 261–266 (1999)