Subband Blind Audio Source Separation Using a Time-Domain Algorithm and Tree-Structured QMF Filter Bank*

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. T-ABCD is a time-domain method for blind linear separation of audio sources proposed by Koldovský and Tichavský (2008). The method produces short separating filters (5-40 taps) and works well with signals recorded at the sampling frequency of 8-16 kHz. In this paper, we propose a novel subband-based variant of T-ABCD, in which the input signals are decomposed into subbands using a tree-structured QMF filter bank. T-ABCD is then applied to each subband in parallel, and the separated subbands are re-ordered and synthesized to yield the final separated signals. The analysis filter of the filter bank is carefully designed to enable maximal decimation of signals without aliasing. Short filters applied within subbands then result in sufficiently long filters in fullband. Using a reasonable number of subbands, the method yields improved speed, stability and performance at an arbitrary sampling frequency.

1 Introduction

Blind separation (BSS) of simultaneously active audio sources is a challenging problem within audio signal processing. The goal is to retrieve d audio sources from their convolutive mixtures recorded by m microphones. The model is described by

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

where $x_1(n), \ldots, x_m(n)$ are the observed signals on microphones and $s_1(n), \ldots, s_d(n)$ are the unknown original (audio) signals. This means that the mixing system is a MIMO (multi-input multi-output) linear filter with source-microphone

* 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. 25–32, 2010.

impulse responses h_{ij}'s each of length M_{ij}. Linear separation consists in finding a MIMO filter that inverts the mixing process (1) and yields estimates of the original signals $s_1(n), \ldots, s_d(n)$. It is convenient to assume the independence of $s_1(n), \ldots, s_d(n)$, and the separation can be based on Independent Component Analysis (ICA) [1]. Indeterminacies that are inherent to the ICA cause that the original colorations of $s_1(n), \ldots, s_d(n)$ cannot be retrieved. The goal is therefore to estimate their microphone responses (images), which only have properly defined colorations. The response of the kth source on the ith microphone is

$$s_k^i(n) = \sum_{\tau=0}^{M_{ik}-1} h_{ik}(\tau) s_k(n-\tau). \tag{2}$$

To apply the ICA, the convolutive mixture (1) must be transformed into an instantaneous one. This is done either directly in the time-domain (TD) by decomposing a matrix usually constructed of delayed copies of signals from microphones, or in the frequency-domain (FD) where the signals are transformed by the Short-Time Fourier Transform (STFT) that converts the convolution operation into the ordinary multiplication. Weaknesses of both approaches are well known from literature. The FD approach meets the so-called permutation problem [2] due to inherent indeterminacies in ICA and requires long data to generate sufficient number of samples for each frequency bin. On the other hand, TD methods are computationally more expensive due to simultaneous optimization of all filter coefficients, which restrict their ability to compute long filters.

A reasonable compromise is the subband approach [3] that consists in decomposing the mixed signals into subbands via a filter bank, separating each subband by a TD method, permuting the separated subbands, and synthesizing the final signals. If a moderate number of subbands is chosen, the permutation problem becomes less difficult compared to the FD approach. Since the subband signals are decimated, the length of separating filters is multiplied.

Several subband approaches have already been proposed in literature using various filter banks. The method in [5] uses a uniform DFT filter bank. Araki et al. [3] use a polyphase filter bank with a single side-band modulation. In [6,7], uniform FIR filter banks were used. All the referenced methods do not apply the maximal decimation of signals in order to reduce the aliasing between subbands. This restrains both the computational efficiency and the effective length of separating filters.

We propose a novel subband method designed to be maximally effective in this respect. The signals are decomposed uniformly into 2^M subbands using a two-channel QMF filter bank applied recursively in the full-blown 2-tree structure with M levels [4]. The signals are decimated by 2 in each level of the 2-tree so they are finally decimated by 2^M, which means *maximal decimation*. Through a careful design of a halfband FIR filter, which determines the whole filter bank, the aliasing is avoided. The blind separation within subbands is then carried out independently by the T-ABCD method [10], which is robust and effective in estimating short separating filters. The permutation problem due to the random order of separated signals in each subband is solved by comparing correlations

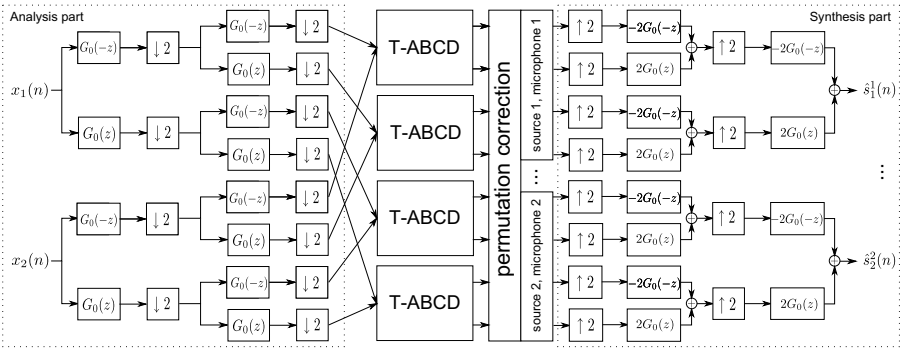

Fig. 1. An illustration of the proposed subband BSS algorithm using the QMF tree-structured filter bank with $M = 2$. Here, two-microphone recording is separated into two responses of each of two original sources.

of absolute values of signals [2]. Finally, the reordered signals are synthesized to yield the estimated responses (2). The flow of the method is illustrated in Fig. 1.

The following section gives more details on the proposed method, and Section 3 demonstrates its performance by experiments done with real-world signals.

2 Proposed Subband BSS Method

2.1 Subband Decomposition

The building block of the tree-structured subband decomposition applied in the proposed method is a two-channel bank that separates the input signals into two bands. In general, a two-channel bank consists of two analysis filters and two synthesis filters whose transfer functions are, respectively, $G_0(z)$, $G_1(z)$, $H_0(z)$ and $H_1(z)$. The input signal is filtered by $G_0(z)$ and $G_1(z)$ in parallel, and the outputs are decimated by 2 giving the subband signals. After the subband processing, the signals are expanded by 2 and passed through the synthesis filters and are added to yield the output signal.

The analysis filters of a Quadrature Mirror Filter (QMF) bank satisfy

$$G_1(z) = G_0(-z). \tag{3}$$

$G_0(z)$ should be a low-pass filter with the pass band $[-\pi/2, \pi/2]$ so that the decimated signals are not aliased. The synthesis filters may be defined as

$$H_0(z) = 2G_1(-z), \qquad H_1(z) = -2G_0(-z). \tag{4}$$

Then the whole two-channel QMF bank is determined by $G_0(z)$.

(4) is a sufficient condition for eliminating the aliasing from synthesized signals provided that no subband processing is done, i.e., when the signals are expanded immediately after the decimation (equation (12.58) in [4]). In such special case,

the transfer function of the two-channel QMF bank is $[G_0(z)]^2 - [G_0(-z)]^2$. It follows that the bank does not possess the perfect reconstruction property in general, which is nevertheless not as important in audio applications. While phase distortions are avoided provided that $G_0(z)$ has a linear phase, amplitude distortions can be made inaudible by a careful design of the filter[1].

To decompose the signal into more than two bands, the analysis part of the two-channel QMF bank can be applied recursively to split each band into two subbands etc. If the depth of the recursion is M, the filter bank splits the spectrum uniformly into 2^M subbands. This approach is utilized in the proposed method as demonstrated by Fig. 1. After the processing of subbands, the synthesis is done backwards then the analysis.

2.2 Separation Algorithm: T-ABCD

T-ABCD is an ICA-based method for blind separation of audio signals working in time-domain. It is based on the estimation of all independent components (ICs) of an *observation space* by an incorporated ICA algorithm. The observation space is spanned by rows of a data matrix $\mathbf{X}$ that may be defined in a general way [10]. For simplicity, we will consider the basic definition that is common to other TD methods [12]: Rows of $\mathbf{X}$ contain L time-shifted copies of each observed signal $x_1(n), \ldots, x_m(n)$. The number of rows of $\mathbf{X}$ is mL, which is the dimension of the observation space. Linear combinations of rows of $\mathbf{X}$ correspond to outputs of FIR MISO filters of the length L (hence also the ICs of $\mathbf{X}$). The steps of T-ABCD are as follows.

1. Find all mL independent components of $\mathbf{X}$ by an ICA algorithm.
2. Group the components into clusters so that each cluster contains components corresponding to the same original source.
3. For each cluster, use components of the cluster to reconstruct microphone responses (images) of a source corresponding to the cluster.

For more details on the method see [9] and [10].

A shortcoming of T-ABCD is that its computational complexity grows rapidly with L. On the other hand, T-ABCD is very powerful when L is reasonably low ($L = 1, \ldots, 40$). This is because all ICs of $\mathbf{X}$ are estimated without applying any constraint to the separating MISO filters (step 1), and all ICs are used to reconstruct the sources' responses (steps 2 and 3). The performance of T-ABCD is robust as it is independent of an initialization provided that the applied ICA algorithm in step 1 is equivariant. Consequently, the use of T-ABCD within the subband separation is desirable, because the separating filters in subbands are shorter than those in fullband [3].

[1] We have chosen $G_0(z)$ as an equiripple FIR filter [4] with 159 taps having the minimum attenuation of 60 dB in the stopband. To eliminate the aliasing, the stop-frequency was shifted slightly from $\pi/2$ to the left by $\epsilon \approx 0.01$, which is small enough so that the cut-off band around $\pi/2$ is very narrow and results in inaudible distortions of signals.

2.3 The Permutation Problem

The estimated responses of sources by T-ABCD are randomly permuted due to indeterminacy of ICA or, more specifically, due to the indeterminacy of the order of clusters identified by step 2. Since the permutation might be different in each subband, the estimated signals in subbands must be aligned before synthesizing them.

Let $\hat{s}^i_{k,j}(n)$, $k = 1, \ldots, d$ be the not yet sorted estimates of responses of the sources at the ith microphone in the jth subband. We wish to find permutations $\pi_j(k)$, $j = 1, \ldots, M$ such that $\hat{s}^i_{\pi_j(k),j}(n)$ is the estimated response of the kth source at the microphone in the subband. We shall assume, for convenience, that the order of the components in one, say in the j_1th subband (e.g. $j_1 = 1$), is correct. Therefore we set $\pi_{j_1}(k) = k$, $k = 1, \ldots, d$. Permutations in all other subbands can be found by maximizing the following criterion,

$$d(p,q,r,s) = \sum_{i=1}^{m} \left|\mathrm{cov}\left(|\widehat{s}^i_{p,q}(n)|, |\widehat{s}^i_{r,s}(n)|\right)\right| = \\ = \sum_{i=1}^{m} \frac{1}{T} \sum_{n=1}^{T} \left(|\widehat{s}^i_{p,q}(n)| - \frac{1}{T}\sum_{t=1}^{T} |\widehat{s}^i_{p,q}(t)|\right) \left(|\widehat{s}^i_{r,s}(n)| - \frac{1}{T}\sum_{t=1}^{T} |\widehat{s}^i_{r,s}(t)|\right) \quad (5)$$

that compares dynamic profiles (absolute values) of the signals [2], as follows.

1. Put $\mathcal{S} = \{j_1\}$, a set of already permuted subbands.
2. Find $j_2 = \arg\max_{s \notin \mathcal{S}} \{\max_{p,r} d(p, j_1, r, s)\}$.
3. Use the greedy algorithm to find $\pi_{j_2}(\cdot)$ by maximizing $d(\cdot, j_1, \cdot, j_2)$. Namely, define $\mathcal{P} = \emptyset$ and $\mathcal{R} = \emptyset$, and repeat
   (a) $(p, r) = \arg\max_{p \notin \mathcal{P}, r \notin \mathcal{R}} d(p, j_1, r, j_2)$
   (b) put $\pi_{j_2}(p) = r$
   (c) $\mathcal{P} = \mathcal{P} \cup \{p\}$, $\mathcal{R} = \mathcal{R} \cup \{r\}$
   until $\mathcal{P} \subsetneq \{1, \ldots, M\}$
4. $\mathcal{S} = \mathcal{S} \cup \{j_2\}$, $j_1 = j_2$.
5. If $\mathcal{S} \subsetneq \{1, \ldots, M\}$, go to 2.

3 Experiments

To demonstrate the performance of the proposed method, we test it on selected data from the SiSEC 2010 campaign[2]. The data consists of two-microphone real-world recordings of, respectively, two male and two female speakers played over loudspeakers (signal combinations #1 and #2) placed in room #1 in position #1 shown in Fig. 2. Each source was recorded separately to obtain its microphone responses, and the signals were summed to obtain the mixed signals; the original sampling rate was 44.1kHz.

[2] The task "Robust blind linear/non-linear separation of short two-sources-two-microphones recordings" in the "Audio source separation" category; see `http://sisec.wiki.irisa.fr/tiki-index.php`

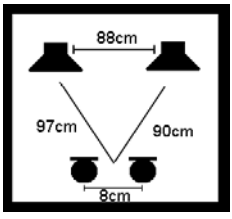

Fig. 2. The position of microphones and loudspeakers in the experiment

For evaluation of the separation, we use the standard Signal-to-Interference Ratio (SIR), as defined in [13]. The evaluation is computed using the full length of recordings, which is about 2 seconds, but only the first second of the data was used for computations of separating filters.

We compare the original T-ABCD from [9] working in fullband with the proposed subband T-ABCD decomposing signals into 2, 4, 8, and 16 subbands, that is, with $M = 1, \ldots, 4$. The fullband T-ABCD is applied with $L = 20$, while in subbands $L = 10$ is taken. The other parameters of T-ABCD are the same both in fullband and subband; namely, the weighting parameter is $\alpha = 1$, and the BGSEP algorithm from [11] is used for finding ICs of $\mathbf{X}$.

Fig. 3 shows the results of experiments done with signals resampled to the sampling rates f_s =8, 16, 32, and 44.1 kHz, respectively. The performance of the fullband T-ABCD decreases with the growing f_s. This is due to the fact that the effective length of separating filters decreases as L is fixed to 20. A comparable length of filters is applied in the 2-subbands method, where $L = 10$ in each subband. The performance of the 2-subbands method is either comparable ($f_s = 8$ and 32 kHz) to the fullband method or even better ($f_s = 16$ and 44.1 kHz) and does not decrease until $f_s \leq 16$. This points to the fact that the fullband method suffers from increased bandwidth of signals when f_s grows.

As can be seen from Fig. 3, the performance of the subband method does not automatically increase with the number of subbands. This is mainly caused by the permutation problem, which becomes more difficult with the growing number of subbands. The results indicate that the optimal bandwidth of subbands is between 2-5 kHz. Namely, (1) the 4-subbands method performs best at $f_s = 16$ and 32 kHz, (2) the 8-subbands method provides the best results when $f_s = 32$ and 44.1 kHz, and (3) the 16-subbands method seems to be effective if $f_s = 44.1$ kHz. On the other hand, the decomposition of signals into 16 subbands seems to be inadequate when $f_s = 8$ or 16 kHz, as the 16-subbands method yields unstable performance here.

3.1 Computational Aspects

The methods were running on a PC with quad-core i7 2.66 GHz processor in Matlab^{TM} with Parallel Computing Toolbox^{TM}. There were four running workers, i.e. one for each core of the processor, which means that up to four T-ABCDs

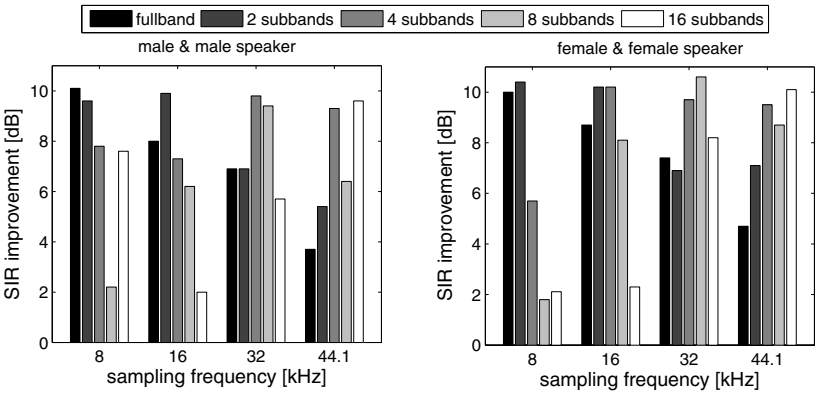

Fig. 3. SIR improvement achieved by the separation. Each value is an average over both separated sources and estimated microphone responses.

may run simultaneously in subbands. The average computational burden summarizes Table 1 in the form A/B, where A and B denote the time needed for separation without and with the aid of parallel computations, respectively. The parallelization was realized through the parallel for-cycle (`parfor`).

The values in Table 1 prove the advantage of the subband method consisting in lower computational complexity. Although the parallelization by means of the Parallel Computing Toolbox^{TM} is not that effective, it points to the potential improvement in terms of speed. For example, the 4-subband method should be almost four-times faster when running in parallel, since about 80% of the computational burden is caused by T-ABCD, while the permutation correction takes about 3% and the rest is due to the filtering operations.

Table 1. Average time needed per separation without and with parallel computations

	computational time [s]			
	8kHz	16kHz	32kHz	44.1kHz
fullband	0.42/ -	0.90/ -	2.03/ -	2.84/ -
2-subband	0.25/0.17	0.46/0.31	0.90/0.69	1.35/0.97
4-subband	0.30/0.19	0.56/0.33	1.06/0.65	1.51/0.97
8-subband	0.40/0.25	0.66/0.42	1.26/0.83	1.83/1.13
16-subband	0.56/0.35	0.87/0.55	1.50/0.99	2.20/1.39

4 Conclusion

The proposed subband T-ABCD was shown to be an improved variant of T-ABCD in terms of speed and separation performance, especially, when working with signals sampled at sampling rates higher than 16 kHz. The method is able to separate one second of data in a lower time, which points to its applicability in a batch-online processing. The optimum number of subbands depends on the

sampling frequency, which was shown to correspond to the bandwidth of about 2-5kHz per subband. Experiments not shown here due to lack of space show that the subband T-ABCD might be combined with other filter banks (e.g. [3]) as well, but the analysis filters must be adjusted to avoid the aliasing in maximally decimated signals.

References

1. Comon, P.: Independent component analysis: a new concept? Signal Processing 36, 287–314 (1994)
2. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech Audio Processing 12(5), 530–538 (2004)
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. Porat, B.: A course in digital signal processing. John Wiley & Sons, New York (1997)
5. Grbić, N., Tao, X.-J., Nordholm, S.E., Claesson, I.: Blind signal separation using overcomplete subband representation. IEEE Transactions on Speech and Audio Processing 9(5), 524–533 (2001)
6. Russell, I., Xi, J., Mertins, A., Chicharo, J.: Blind source separation of nonstationary convolutively mixed signals in the subband domain. In: ICASSP 2004, vol. 5, pp. 481–484 (2004)
7. Kokkinakis, K., Loizou, P.C.: Subband-based blind signal processing for source separation in convolutive mixtures of speech. In: ICASSP 2007, vol. 4, pp. 917–920 (2007)
8. Duplessis-Beaulieu, F., Champagne, B.: Fast convolutive blind speech separation via subband adaptation. In: ICASSP 2003, vol. 5, pp. 513–516 (2003)
9. Koldovský, Z., Tichavský, P.: Time-domain blind audio source separation using advanced component clustering and reconstruction. In: HSCMA 2008, Trento, Italy, pp. 216–219 (2008)
10. 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)
11. Tichavský, P., Yeredor, A.: Fast approximate joint diagonalization incorporating weight matrices. IEEE Transactions of Signal Processing 57(3), 878–891 (2009)
12. Bousbia-Salah, H., Belouchrani, A., Abed-Meraim, K.: Jacobi-like algorithm for blind signal separation of convolutive mixtures. IEE Elec. Letters 37(16), 1049–1050 (2001)
13. Vincent, E., Févotte, C., Gribonval, R.: Performance measurement in blind audio source separation. IEEE Trans. Audio, Speech and Language Processing 14(4), 1462–1469 (2006)