USING THE SCALING AMBIGUITY FOR FILTER SHORTENING IN CONVOLUTIVE BLIND SOURCE SEPARATION

Radoslaw Mazur and Alfred Mertins

Institute for Signal Processing
University of Lübeck, 23538 Lübeck, Germany

ABSTRACT

In this paper, we propose to use the scaling ambiguity of convolutive blind source separation for shortening the unmixing filters. An often used approach for separating convolutive mixtures is the transformation to the time-frequency domain where an instantaneous ICA algorithm can be applied for each frequency separately. This approach leads to the so called permutation and scaling ambiguity. While different methods for the permutation problem have been widely studied, the solution for the scaling problem is usually based on the minimal distortion principle. We propose an alternative approach that allows the unmixing filters to be as short as possible. Shorter unmixing filters will suffer less from circular-convolution effects that are inherent to unmixing approaches based on binwise ICA followed by permutation and scaling correction. The results for the new algorithm will be shown on a real-world example.

***Index Terms*—** Blind source separation, convolutive mixture, frequency-domain ICA, scaling problem.

1. INTRODUCTION

Blind source separation (BSS) is a method for recovering signals from observed mixtures. Usually, neither the mixing system nor the original signals are known. The instantaneous case has been widely studied and there exist several efficient algorithms [1, 2, 3].

In a real-world scenario in an echoic environment, the situation becomes more difficult. As the signals arrive several times with different time lags, the mixing process becomes convolutive. This can be modelled using FIR filters, where a realistic scenario requires lengths of several thousand taps. For doing BSS in such a scenario, an estimation of an inverse system of similar length is required. One way is to estimate the unmixing filters directly in the time domain [4, 5]. The drawbacks of this approach are a high computational cost and difficulties of convergence, as the algorithm often gets trapped in a local minimum. Therefore, another approach is widely used: After transformation to the time-frequency domain, the convolution becomes a multiplication [6], and each frequency bin can be separated using an instantaneous method. This simplification has a major disadvantage though. As every separated bin can be arbitrarily permuted and scaled, a correction is needed. When the permutation is not correctly solved the separation of the entire signals fails. A variety of different approaches has been proposed to solve this problem utilizing either the time structure of the signals [7, 8, 9] or the properties of the unmixing matrices [10, 11, 12].

When the scaling is not corrected, a filtered version of the signals is recovered. In [13, 14] the authors proposed a post-filter method that is able to recover the signals as they have been recorded at the microphones, accepting the distortions of the mixing system while not adding new ones.

The circular-convolution problem of convolutive BSS has been explicitly addressed in [15]. As the unmixing matrices are calculated independently for each frequency, the desired linear convolution may turn into a circular one. To reduce these effects the authors applied a smoothing to the filters in the time domain.

In this paper, we propose a new method for solving the scaling ambiguity with the aim of making the unmixing filters as short as possible. In order to achieve this, we calculate the dependency between the scaling factors and the impulse responses of the unmixing filterbank and select the scaling factors that minimize a certain optimality criterion. The new filters are then short enough to avoid the circularity problem.

2. MODEL AND METHODS

2.1. BSS for instantaneous mixtures

In this section, we describe the instantaneous unmixing process that we used in frequency bins of the convolutive one. The instantaneous mixing process of N sources into N observations is modeled by an $N \times N$ matrix $\mathbf{A}$. With the source vector $\boldsymbol{s}(n) = [s_1(n), \dots, s_N(n)]^T$ and negligible measurement noise, the observation signals $\boldsymbol{x}(n) = [x_1(n), \dots, x_N(n)]^T$ are given by

$$\boldsymbol{x}(n) = \mathbf{A} \cdot \boldsymbol{s}(n). \quad (1)$$

The separation is again a multiplication with a matrix $\mathbf{B}$:

$$\boldsymbol{y}(n) = \mathbf{B} \cdot \boldsymbol{x}(n) \quad (2)$$

with $\boldsymbol{y}(n) = [y_1(n), \ldots, y_N(n)]^T$. The only source of information for the estimation of $\mathbf{B}$ is the observed process $\boldsymbol{x}(n)$. The separation is successful when $\mathbf{B}$ can be estimated so that $\mathbf{BA} = \mathbf{D\Pi}$ with $\mathbf{\Pi}$ being a permutation matrix and $\mathbf{D}$ being an arbitrary diagonal matrix. These two matrices stand for the two ambiguities of BSS. The signals may appear in any order and can be arbitrarily scaled.

For the separation we use the well known gradient-based update rule [1]

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \Delta\mathbf{B}_k \tag{3}$$

with

$$\Delta\mathbf{B}_k = \mu_k(\boldsymbol{I} - E\left\{\boldsymbol{g}(\boldsymbol{y})\boldsymbol{y}^T\right\})\mathbf{B}_k. \tag{4}$$

The term $\boldsymbol{g}(y) = (g_1(y_1), \ldots g_n(y_n))$ is a component-wise vector function of nonlinear score functions

$$g_i(s_i) = -\frac{p_i'(s_i)}{p_i(s_i)} \tag{5}$$

where $p_i(s_i)$ are the assumed source probability densities. These should be known or at least well approximated in order to achieve good separation performance [16].

2.2. Convolutive mixtures

When dealing with real-world acoustic scenarios it is necessary to consider the reverberation. The mixing system can be modeled by FIR filters of length L. Depending on the reverberation time and sampling rate, L can reach several thousand. The convolutive mixing model reads

$$\boldsymbol{x}(n) = \mathbf{H}(n) * \mathbf{s}(n) = \sum_{l=0}^{L-1} \mathbf{H}(l)\boldsymbol{s}(n-l) \tag{6}$$

where $\mathbf{H}(n)$ is a sequence of $N \times N$ matrices containing the impulse responses of the mixing channels. For the separation we use FIR filters of length M and obtain

$$\boldsymbol{y}(n) = \mathbf{W}(n) * \boldsymbol{x}(n) = \sum_{l=0}^{M-1} \mathbf{W}(l)\boldsymbol{x}(n-l) \tag{7}$$

with $\mathbf{W}(n)$ containing the unmixing coefficients. In order to achieve satisfying performance we choose $M \geq L-1$ [17].

Using the short-time Fourier transform (STFT), the signals can be transformed to the time-frequency domain, where the convolution approximately becomes a multiplication [6]:

$$\boldsymbol{Y}(\omega_k, \tau) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k, \tau), \quad k = 0, 1, \ldots, K-1, \tag{8}$$

where K is the FFT length. The major benefit of this approach is the possibility to estimate the unmixing matrices for each frequency independently, however, at the price of possible permutation and scaling in each frequency bin:

$$\boldsymbol{Y}(\omega_k, \tau) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k, \tau) = \boldsymbol{D}(\omega_k)\mathbf{\Pi}(\omega_k)\boldsymbol{S}(\omega_k, \tau) \tag{9}$$

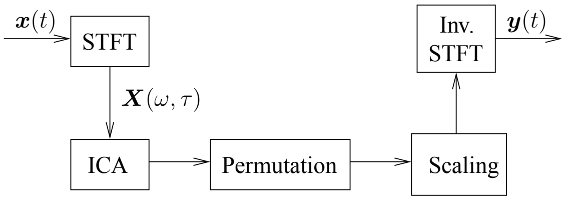


Fig. 1. Overview of frequency-domain BSS

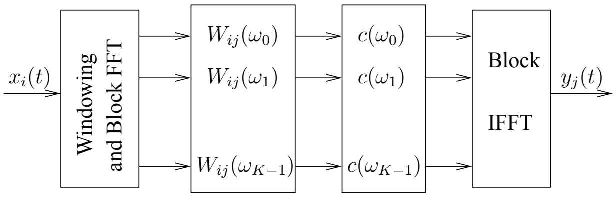


Fig. 2. Data flow from input i to output j.

where $\mathbf{\Pi}(\omega)$ is a frequency-dependent permutation matrix and $\boldsymbol{D}(\omega)$ an arbitrary diagonal scaling matrix.

The correction of the permutation is essential, because the entire unmixing process fails if different permutations occur at different frequencies. A number of approaches has been proposed to solve this problem. [7, 8, 9, 10, 11, 12].

When the scaling ambiguity is not solved, filtered versions of the sources are recovered. A widely used approach has been proposed in [13]. The authors recovered the signals as they were recorded at the microphones accepting all filtering done by the mixing system. A similar technique has been proposed in [14] under the paradigm of the minimal distortion principle, which uses the unmixing matrix

$$\boldsymbol{W}'(\omega) = \mathrm{dg}(\boldsymbol{W}^{-1}(\omega)) \cdot \boldsymbol{W}(\omega) \tag{10}$$

with dg$(\cdot)$ returning the argument with all off-diagonal elements set to zero.

3. FILTER SHORTENING

The proposed method is to estimate a set of scaling factors $c(\omega)$ so that the filter lengths of the unmixing filters are reduced. The motivation for this comes from the fact that the conversion of time-domain convolution into frequency-domain multiplication is only then exactly equivalent when certain conditions on the filter and FFT lengths, known from fast-convolution algorithms, are satisfied. With arbitrary scaling factors, however, the frequency-domain multiplication will results in circular-convolution artifacts.

In Fig. 1 the overall BSS system is shown. It consists of $N \times N$ single channels as depicted in Fig. 2. In this representation the permutation has already been corrected. The dependency of time-domain filter coefficients of a filter vector $\boldsymbol{w}_{ij}$ and scaling factors $\boldsymbol{c} = [c(\omega_0), c(\omega_1), \ldots, c(\omega_{K-1})]^T$ can be calculated as follows:

$$\begin{aligned} \boldsymbol{w}_{ij} &= \sum_l \boldsymbol{E}_l \cdot \mathcal{F}^{-1} \cdot \boldsymbol{C} \cdot \boldsymbol{W}_{ij} \cdot \mathcal{F} \cdot \boldsymbol{D}_l \cdot \delta \\ &= \sum_l \boldsymbol{E}_l \cdot \mathcal{F}^{-1} \cdot \mathrm{diag}(\mathcal{F} \cdot \boldsymbol{D}_l \cdot \delta) \cdot \boldsymbol{W}_{ij} \cdot \boldsymbol{c} \qquad (11) \\ &= \boldsymbol{V}_{ij} \cdot \boldsymbol{c} \end{aligned}$$

where $\mathrm{diag}(\cdot)$ converts a vector to a diagonal matrix. The term δ a unit vector containing a single one and zeros otherwise. $\boldsymbol{D}_l$ is a diagonal matrix containing the coefficients of the STFT analysis window shifted to the lth position according to the STFT window shift. $\mathcal{F}$ is the DFT matrix. $\boldsymbol{W}_{ij}$ is a diagonal matrix containing the frequency-domain unmixing coefficients. $\boldsymbol{c}$ is a vector of the sought scaling factors, and $\boldsymbol{C}$ is a diagonal matrix made up as $\boldsymbol{C} = \mathrm{diag}(\boldsymbol{c})$. $\boldsymbol{E}_l$ is a shifting matrix corresponding to $\boldsymbol{D}_l$, defined in such a way that the overlapping STFT blocks are correctly merged. Note that for real-valued signals and filters, the above equation can be modified to exploit the conjugate symmetry in the frequency domain.

The calculation of an optimal scaling vector $\boldsymbol{c}$ that leads to filters $\boldsymbol{w}_{ij}$ of short length can be done by minimizing

$$\|\bar{\mathbf{w}} - \bar{\boldsymbol{V}}\boldsymbol{c}\|_{\ell_2} \qquad (12)$$

where $\bar{\boldsymbol{V}}$ and $\tilde{\mathbf{w}}$ are vertical concatenations of matrices $\boldsymbol{V}_{ij}$ and some desired filters $\tilde{\mathbf{w}}_{ij}$, respectively. In the proposed method, each vector $\bar{\mathbf{w}}_{ij}$ contains zeros and a single one at the position where the corresponding $\boldsymbol{w}_{ij}$ has its main peak. The solution is given by $\boldsymbol{c} = \bar{\boldsymbol{V}}^{+}\bar{\mathbf{w}}$, with $\bar{\boldsymbol{V}}^{+}$ beeing the pseudoinverse of $\bar{\boldsymbol{V}}$.

4. SIMULATIONS

Simulations have been done on real-world data available at [18]. This data set consists of eight-seconds long speech recordings sampled at 8 kHz. The chosen parameters were a Hann window of length 2048, a window shift of 256, and an FFT-length of $K = 4096$. 200 iterations of (4) for each frequency bin have been done. As the original sources are available for the considered data set, the permutation problem could be ideally solved, so that permutation ambiguities could not influence the results.

In Figs. 3 and 4 the filters designed with the traditional method (10) and the proposed method are shown, respectively. The main difference is the clearly visible and significantly bigger main peak and the faster decay of the impulse responses designed with our method. As one can observe by comparing Figs. 3(b) and 4(b), the energy difference between the main peak and the tail of the impulse responde could be increased by about 25 dB.

The new filters are also able to significantly enhance the separation performance as shown in Table 1.

Table 1. Comparison of the signal-to-interference ratios in dB between the minimal distortion principle and the new algorithm.

	Left	Right	Overall
MDP	18.05	15.27	16.18
New Alg.	20.62	26.48	23.04

5. SUMMARY

In this paper, we have proposed the use of the scaling ambiguity of convolutive blind source separation for shortening the unmixing filters. We calculate a set of scaling factors that maximize the energy ratio of the main peak and the tail of the impulse response. On a real-world example, the energy decay could be improved by 25 dB, which also translated into better signal-to-interference ratios.

6. REFERENCES

[1] S. Amari, A. Cichocki, and H. H. Yang, “A new learning algorithm for blind signal separation,” in *Advances in Neural Information Processing Systems*, David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, Eds. 1996, vol. 8, pp. 757–763, The MIT Press.

[2] A. Hyvärinen and E. Oja, “A fast fixed-point algorithm for independent component analysis,” *Neural Computation*, vol. 9, pp. 1483–1492, 1997.

[3] J.-F. Cardoso and A. Soulomiac, “Blind beamforming for non-Gaussian signals,” *Proc. Inst. Elec. Eng., pt. F.*, vol. 140, no. 6, pp. 362–370, Dec. 1993.

[4] S. C. Douglas, H Sawada, and S. Makino, “Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 1, pp. 92–104, Jan 2005.

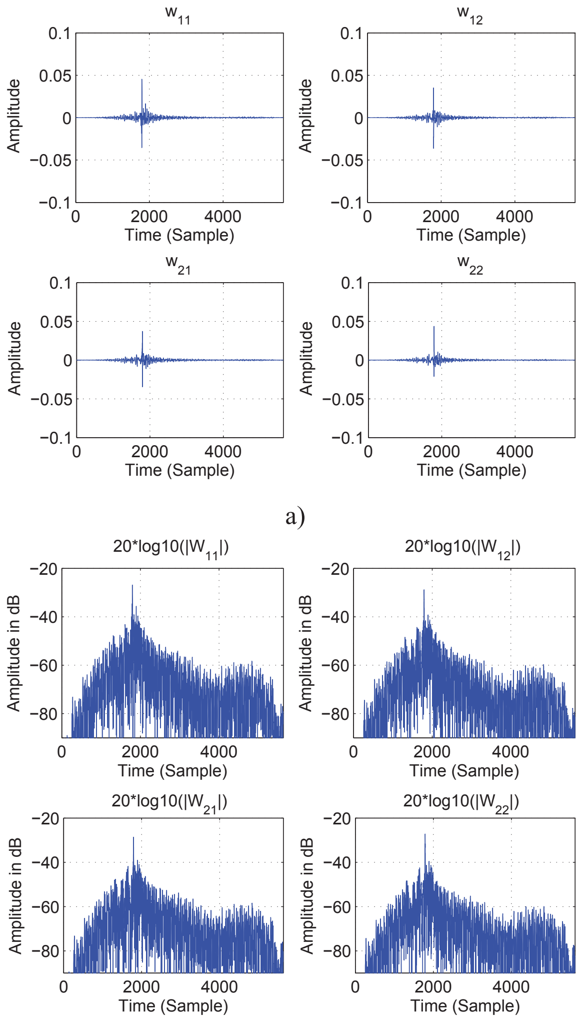
[5] R. Aichner, H. Buchner, S. Araki, and S. Makino, “Online time-domain blind source separation of nonstationary convolved signals,” in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003, pp. 987–992.

[6] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain.,” *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[7] K. Rahbar and J.P. Reilly, “A frequency domain method for blind source separation of convolutive audio mixtures,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 832–844, Sept. 2005.

[8] J. Anemüller and B. Kollmeier, “Amplitude modulation decorrelation for convolutive blind source separation,” in *Proceddings of the second international workshop on independent component analysis and blind signal separation*, 2000, pp. 215–220.

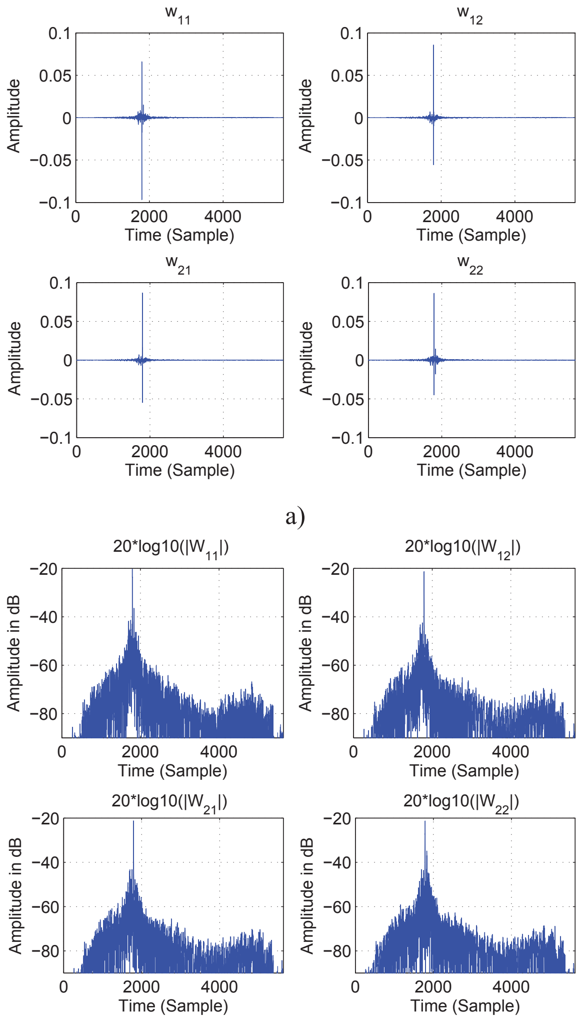
[9] R. Mazur and A. Mertins, “Solving the permutation problem in convolutive blind source separation,” in *Independent Component Analysis and Signal Separation*, M. E. Davies, Ch. J. James, S. A. Abdallah, and M. D. Plumbley, Eds. 2007, vol. 4666 of *LNCS*, pp. 512–519, Springer.



a)

b)

Fig. 3. Filter set using the Minimal Distortion Principle. (a) The unmixing filter impulse responses $w_{ij}(n)$. (b) $20\log|w_{ij}(n)|$.



a)

b)

Fig. 4. Filter set using new method. (a) The unmixing filter impulse responses $w_{ij}(n)$. (b) $20\log|w_{ij}(n)|$.

[10] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[11] W. Wang, J. A. Chambers, and S. Sanei, "A novel hybrid approach to the permutation problem of frequency domain blind source separation," in *Lecture Notes in Computer Science*. 2004, vol. 3195, pp. 532–539, Springer.

[12] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation of 3-d located many speech signals," in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2005, pp. 9–12.

[13] S. Ikeda and N. Murata, "A method of blind separation based on temporal structure of signals.," in *Proc. Int. Conf. on Neural Information Processing*, 1998, pp. 737–742.

[14] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proceedings of the 41st SICE Annual Conference*, 5-7 Aug. 2002, vol. 4, pp. 2138–2143.

[15] H. Sawada, R. Mukai, S. de la Kethulle, S. Araki, and S. Makino, "Spectral smoothing for frequency-domain blind source separation," *International Workshop on Acoustic Echo and Noise Control (IWAENC2003)*, pp. 311–314, Sep 2003.

[16] S. Choi, A. Cichocki, and S. Amari, "Flexible independent component analysis," in *Neural Networks for Signal Processing VIII*, T. Constantinides, S. Y. Kung, M. Niranjan, and E. Wilson, Eds., 1998, pp. 83–92.

[17] K. Rahbar and J.P. Reilly, "Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 7-11 May 2001, vol. 5, pp. 2745–2748.

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