A SPARSITY BASED CRITERION FOR SOLVING THE PERMUTATION AMBIGUITY 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 present a new algorithm for solving the permutation ambiguity in convolutive blind source separation. A common approach for separation of convolutive mixtures is the transformation to the time-frequency domain, where the convolution becomes a multiplication. This allows for the use of well-known instantaneous ICA algorithms independently in each frequency bin. However, this simplification leads to the problem of correctly aligning these single bins previously to the transformation to the time domain. Here, we propose a new criterion for solving this ambiguity. The new approach is based on the sparsity of the speech signals and yields a robust depermutation algorithm. The results will be shown on real-world examples.

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

1. INTRODUCTION

Blind Source Separation (BSS) of linear and instantaneous mixtures can be performed using the Independent Component Analysis (ICA). For this case, numerous algorithms have been proposed [1, 2, 3].

When dealing with real-world recordings of speech, this simple approach is not effective anymore. As the signals arrive multiple times with different delays, the mixing procedure becomes convolutive. These characteristics can be modeled using FIR filters. For realistic scenarios these filters can reach lengths of up to several thousand coefficients. In this case, the separation is only possible when the unmixing system is again a set of FIR filters with at least the same length.

The calculation of such filters directly in the time domain is very demanding [4, 5]. Furthermore these algorithms often get trapped in local minima. Due to these problems, another approach is often used. When transformed to the time-frequency domain the convolution becomes a multiplication [6], and the separation can be performed in each frequency bin independently by using an instantaneous algorithm.

However, this simplification has a major disadvantage. The separated signals usually have arbitrary scaling and are randomly permuted across the frequency bins. Without the correction of the scaling, only filtered versions of the signals are restored. This ambiguity is often solved using the minimal distortion principle [7] or inverse postfilters [8]. This method accepts the filtering done by the mixing system without adding new distortions. Newer approaches solve the scaling ambiguity with the aim of filter shortening [9] or shaping [10].

The random permutation of the single frequency bins has an even bigger impact. Without a correct alignment, different signals appear in the single outputs causing the whole process to fail. Many different approaches for solving this problem have been proposed. One class of algorithms make use of the properties of the unmixing matrices. In [11] the authors propose to use these as beamformers. This allows for the calculation of direction of arrival. By arranging the single frequency bins to these directions, depermutation could be achieved for most of the bins. An alternative formulation with the use of directivity patterns has been proposed in [12] and [13]. The major drawback of this approach is the assumption of sources originating from specific directions, which is only valid in low reverberant rooms with no diffuse background noise. In [14] the authors proposed to utilize the sparsity of the unmixing filters. However, in case of real world examples, this assumption is only valid for small parts of the filters.

The other group of algorithms uses the time structure of the separated bins. Here, a common idea is the assumption of high correlation between neighboring bins. This has been used for example in [8] and [15]. In [16] the authors use the amplitude modulation correlation for getting a separation criterion which avoids the permutation problem. Other approaches include a statistical modeling of the single bins using the generalized Gaussian distribution. Small differences of the parameters lead to a depermutation criterion in [17] and [18].

In low reverberant environments the algorithms from the first group usually perform better. With longer reverberation times the assumption of a single direction for the single sources at different frequencies is no longer valid. In this case the algorithms from the second group have to be used.

In this work, we propose a new criterion which is based on the sparsity of the time domain representation of speech signals. For improved robustness a dyadic depermutation scheme as in [15] is used. The improved performance will be shown on real world examples.

2. MODEL AND METHODS

2.1. BSS for instantaneous mixtures

In this section, we describe the instantaneous unmixing process that we use 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 $\boldsymbol{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ and negligible measurement noise, the observation signals $\boldsymbol{x}(n) = [x_1(n), \ldots, x_N(n)]^T$ are given by

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

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

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

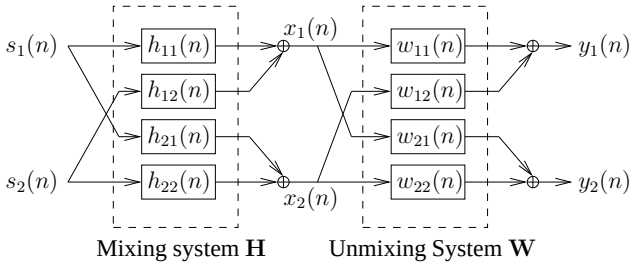

Fig. 1. BSS model with two sources and sensors.

with $\boldsymbol{y}(n) = [y_1(n), \ldots, y_N(n)]^T$. The single 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 $\boldsymbol{g}(\boldsymbol{y}) = (g_1(y_1), \ldots g_n(y_n))$ is a component-wise vector function of nonlinear score functions $g_i(s_i) = -p_i'(s_i)/p_i(s_i)$ 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 [19].

2.2. Convolutive mixtures

When dealing with real-world acoustic scenarios it is necessary to consider 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

$$\boldsymbol{x}(n) = \mathbf{H}(n) * \mathbf{s}(n) = \sum_{l=0}^{L-1} \mathbf{H}(l)\boldsymbol{s}(n-l) \tag{5}$$

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

$$\boldsymbol{y}(n) = \mathbf{W}(n) * \boldsymbol{x}(n) = \sum_{l=0}^{M-1} \mathbf{W}(l)\boldsymbol{x}(n-l) \tag{6}$$

with $\mathbf{W}(n)$ containing the unmixing coefficients. Fig. 1 shows the scenario for two sources and sensors.

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]:

$$\boldsymbol{Y}(\omega_k, \tau) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k, \tau), \quad k = 0, 1, \ldots, K-1, \tag{7}$$

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:

$$\boldsymbol{Y}(\omega_k, \tau) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k, \tau) = \boldsymbol{D}(\omega_k)\mathbf{\Pi}(\omega_k)\boldsymbol{S}(\omega_k, \tau) \tag{8}$$

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

Without correction of scaling, a filtered version of the sources is recovered. The already mentioned minimal distortion principle uses unmixing matrix

$$\boldsymbol{W}'(\omega) = \mathrm{dg}(\boldsymbol{W}^{-1}(\omega)) \cdot \boldsymbol{W}(\omega) \tag{9}$$

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

The correction for permutation is essential, as otherwise different signals will be restored at different frequencies and the whole process will fail. In the next section we will review the correlation approach for solving the permutation problem and the dyadic scheme improvements.

3. DEPERMUTATION ALGORITHM

There exist many algorithms that relay on the statistics of the separated signals. Usually the high correlation of neighboring bins is assumed [8]. With $\boldsymbol{V}(\omega, \tau) = |\boldsymbol{Y}(\omega, \tau)|$, the correlation between two bins k and l is defined as

$$\rho_{qp}(\omega_k, \omega_l) = \frac{\sum_{\tau=0}^{\mathcal{T}-1} V_q(\omega_k, \tau)V_p(\omega_l, \tau)}{\sqrt{\sum_{\tau=0}^{\mathcal{T}-1} {V_q}^2(\omega_k, \tau)}\sqrt{\sum_{\tau=0}^{\mathcal{T}-1} {V_p}^2(\omega_l, \tau)}} \tag{10}$$

where p, q are the indices of the separated signals, $V_q(\omega_k, \tau)$ is the qth element of $\boldsymbol{V}(\omega_k, \tau)$, and $\mathcal{T}$ is the number of frames. The decision on aligning the bins is made on the basis of the ratio

$$r_{kl} = \frac{\rho_{pp}(\omega_k, \omega_l) + \rho_{qq}(\omega_k, \omega_l)}{\rho_{pq}(\omega_k, \omega_l) + \rho_{qp}(\omega_k, \omega_l)}. \tag{11}$$

It is assumed that with $r_{kl} > 1$ the bins are correctly aligned and otherwise a permutation has occurred. Aligning consecutive bins using (11) is not robust, as single wrong permutations lead to whole blocks of falsely permuted bins. The dyadic sorting scheme proposed in [15] approaches this problem. Here, at the first step, only pairs of bins are depermuted. In the second step, these pairs are aligned, and then the resulting quadruples are depermuted. This scheme is continued until all bins are processed. Within this procedure, single wrong permuted bins at the early stages do not outbalance the majority.

In [15] the depermutation of larger blocks is essentially based on the correlation of the single bins within these blocks. The increased robustness is only due to the fact that the maxima and respectively the minima of the correlation coefficients are compared. With too many wrong permutations at the early stages, or too many poorly separated bins, this method also fails.

In the next section, we propose a new depermutation criterion. It is based on the sparsity of the time domain representation. This formulation is able to use the dyadic sorting scheme and at the same time is able to calculate the depermutation of blocks much more robust.

4. NEW ALGORITHM

The principle of the new algorithm is based on two observations. (1) Speech signals are sparse. (2) A mixture of speech signals is less sparse than the single contributions. An example is given in Fig. 2. In Fig. 2(a) two single speech signals are shown. With some speech pauses between the words these signals are clearly sparse. In Fig. 2(b) the upper halves of the frequency bins are swapped between the two channels. In this case, both of the signals have less

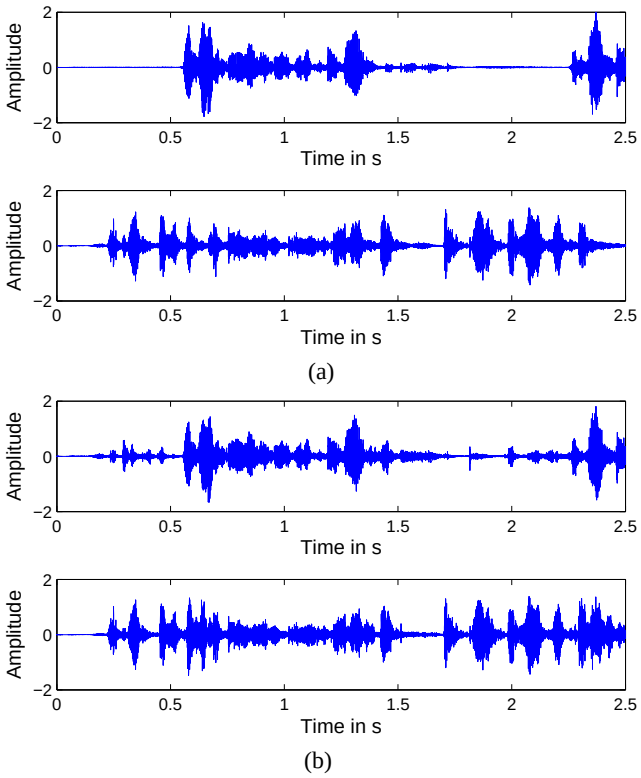

Fig. 2. The demonstration of the sparsity criterion. (a) Two signals with correct alignment of all bins. (b) Half of the bins are wrongly permuted. In this case the signals are less sparse.

pauses and are accordingly less sparse. This can be more formally verified using the fact that with wrong permutations the output signal is the superposition of two or more statistically independent signals, resulting in an amplitude distribution that is less supergaussian than those of the individual signals. With correct permutations, on the other hand, statistically dependent subband components are summed up, leading to a clearly supergaussian distribution.

For the derivation of the criterion, we define analogously to equation (10) a sparsity measure

$$\varrho_{qp}(\omega_{ks}, \omega_{lt}) = \|z_q(\omega_{ks}, n) + z_p(\omega_{lt}, n)\|_{\ell_p} \tag{12}$$

with $z(\omega_{ab}, n)$ being the time-domain representation of the bins in the frequency range $[a, b]$ of $Y(\omega, \tau)$, which may be obtained using the inverse STFT. For $a < b$, the time signal $z(\omega_{ab}, n)$ is a subband signal of the to be restored signal $y(n)$, composed of the frequency bins ω_a to ω_b. Otherwise, for $a = b$, $z(\omega_{ab}, n)$ represents only a single bin, and can be abbreviated as $z(\omega_a, n)$. The sparsity is measured using the ℓ_p pseudo norm

$$\|\boldsymbol{x}\|_{\ell_p} = \left(\sum_{i=0}^{N-1} |x(i)|^p \right)^{\frac{1}{p}} \tag{13}$$

with $0 \leq p \leq 1$. A good choice is $p = 0.1$. Analogously to equation (11) the ratio

$$r_{kl,st} = \frac{\varrho_{pp}(\omega_{ks}, \omega_{lt}) + \varrho_{qq}(\omega_{ks}, \omega_{lt})}{\varrho_{pq}(\omega_k s, \omega_{lt}) + \varrho_{qp}(\omega_{ks}, \omega_{lt})}. \tag{14}$$

for the determining the permutation can be defined. With $r_{kl,st} > 1$ ranges $[k, s]$ and $[l, t]$ of the separated signals q and p are correctly aligned among each other. Otherwise they are permuted.

Table 1. Comparison of the results for different depermutation algorithms in terms of separation performance in dB. Dataset 1 is taken from [20]. Dataset 2 is recorded in higher reverberant room [21].

Algorithm	Dataset 1	Dataset 2
Proposed	15.4	8.1
Dyadic sorting [15]	2.7	3.0
DOA-Approach [11]	17.3	3.4
$\alpha\beta$-Algorithm [17]	18.4	0.3
Non blind	18.4	9.4

The principle of the dyadic depermutation procedure using these definitions is shown in Fig. 3. At first, the depermutation for all successive pairs of single bins is estimated. This is quite alike the procedure from [15], but with the new depermutation criterion. At the next stages, where whole groups of bins are aligned among each other, the situation becomes quite different. With the new formulation, all relevant frequency bins are taken into account when calculating the ratio in equation (14). The method from [15] estimates the correlation coefficients for all single bins independently and relies solely on the maxima. With higher number of frequency bins at the higher stages of the dyadic sorting scheme, the proposed new formulation becomes even more robust.

5. SIMULATIONS

Simulations have been done on real-world data available at [20]. This data set consists of eight-seconds long speech recordings sampled at 8 kHz with individual contributions from the sources to the microphones. The chosen FFT length was $K = 2048$ and every bin has been separated using 200 iterations of (4). In Table 1 the results for different depermutation methods are summarized (Dataset 1). With the low reverberation, the direction of arrival approach and the $\alpha\beta$-algorithm from [17] are both able to depermute almost all bins, and the separation performance is almost as good as in the non-blind case. The new proposed algorithm could also depermute almost all bins. With 2dB less, the separation performance is only slightly reduced. The dyadic sorting based on correlation fails.

With another dataset [21], recorded by the authors of this paper in a higher reverberant room, the situation becomes quite different. As Table 1 shows (Dataset 2), the direction of arrival approach fails, as the assumption of the single direction for every source is not valid. The dyadic sorting scheme from [15] based on correlation also fails. The $\alpha\beta$-Algorithm from [17] fails, too, as one of its steps is the dyadic sorting of frequency groups with the same problems as the plain dyadic sorting. The new proposed algorithm, however, is still able to depermute almost all bins and nearly reaches the performance of the non-blind (i.e., Wiener filter) case. It is the only algorithm that is able to perform the depermutation in a satisfactory manner on such highly reverberated data.

6. CONCLUSIONS

In this paper we proposed a new criterion for solving the permutation ambiguity in convolutive blind source separation. This criterion is based on the sparsity of speech signals and may be used in the dyadic sorting scheme. The robustness of this approach has been shown on real world data, where it was able to significantly outperform other state-of-the-art algorithms.

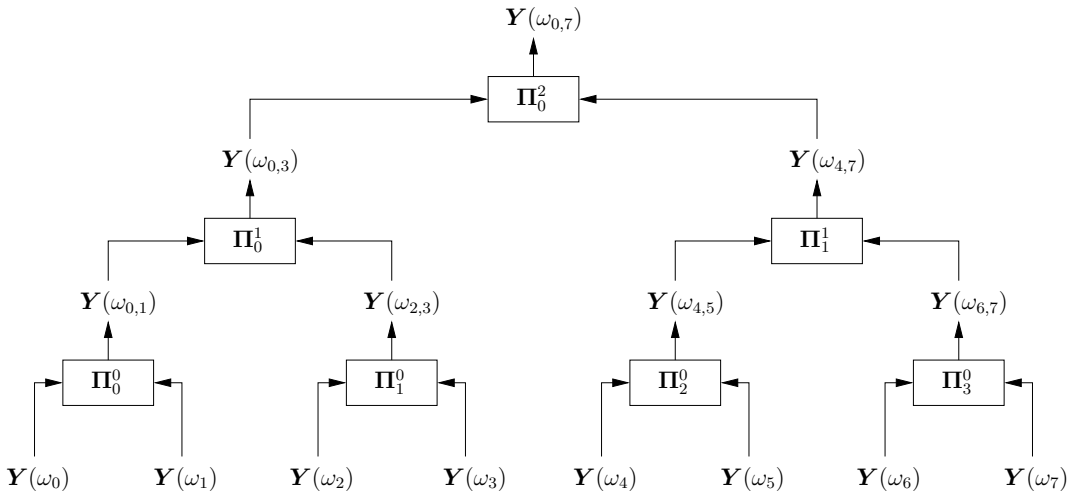

Fig. 3. Dyadic permutation sorting scheme for the case when the total number of frequency bins is $K = 8$.

7. REFERENCES

[1] S.-I. 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., MIT Press, Cambridge, MA, 1996, vol. 8, pp. 757–763.

[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, "On-line 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. 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.

[8] 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.

[9] R. Mazur and A. Mertins, "Using the scaling ambiguity for filter shortening in convolutive blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Taipei, Taiwan, April 2009, pp. 1709–1712.

[10] R. Mazur and A. Mertins, "A method for filter shaping in convolutive blind source separation," in *Independent Component Analysis and Signal Separation (ICA2009)*. 2009, vol. 5441 of *LNCS*, pp. 282–289, Springer.

[11] 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.

[12] 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.

[13] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: investigation and solutions," *IEEE Transactions on Speech and Audio Processing.*, vol. 13, no. 1, pp. 1–13, Jan. 2005.

[14] Prasad Sudhakar and Rémi Gribonval, "Sparse filter models for solving permutation indeterminacy in convolutive blind source separation," in *SPARS'09 - Signal Processing with Adaptive Sparse Structured Representations*, Rémi Gribonval, Ed., Saint Malo France, 2009, Inria Rennes - Bretagne Atlantique.

[15] 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.

[16] J. Anemller 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.

[17] R. Mazur and A. Mertins, "An approach for solving the permutation problem of convolutive blind source separation based on statistical signal models," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 117–126, Jan. 2009.

[18] R. Mazur and A. Mertins, "Simplified formulation of a depermutation criterion in convolutive blind source separation," in *Proc. European Signal Processing Conference*, Glasgow, Scotland, Aug 2009, pp. 1467–1470.

[19] 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.

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

[21] `http://www.isip.uni-luebeck.de/index.php?id=479`