Solving the Permutation Problem in Convolutive Blind Source Separation[⋆]

Radoslaw Mazur and Alfred Mertins

Institute for Signal Processing, University of Lübeck, 23538 Lübeck, Germany
{mazur,mertins}@isip.uni-luebeck.de

Abstract. This paper presents a new algorithm for solving the permutation ambiguity in convolutive blind source separation. When transformed to the frequency domain, the source separation problem reduces to independent instantaneous separation in each frequency bin, which can be efficiently solved by existing algorithms. But this independency leads to the problem of correct alignment of these single bins which is still not entirely solved. The algorithm proposed in this paper models the frequency-domain separated signals using the generalized Gaussian distribution and utilizes the small deviation of the exponent between neighboring bins for the detection of correct permutations.

1 Introduction

Blind Source Separation (BSS) is used to recover signals from observed mixtures without prior knowledge of the sources nor the mixing system. For the case of linear instantaneous mixtures, a number of different efficient approaches has been proposed [1,2]. When aiming at real-world mixtures of audio signals like speech, the situation becomes much more difficult. In this case, the mixing process is convolutive and can be modeled using FIR filters, where, for realistic scenarios, the length of these filters can be up to several thousands taps. The unmixing then has to be done using FIR filters of similar length. It is possible to calculate such filters directly in the time domain [3,4], but this approach suffers from high computational cost and difficulties of convergence. The most successful approach is to transform the signals to the frequency domain, where the convolution becomes multiplication [5]. Then the separation can be done independently in each frequency bin, which is a much simpler task. The major drawback of this approach is that the separated bins usually have different scaling and are arbitrarily permuted. Therefore they have to be correctly equalized and aligned, because otherwise the entire process of separation will fail.

While it is possible to obtain a proper scaling for the frequency components [6], there is still no algorithm that can tackle the permutation problem in all cases. One idea for solving the permutation problem is based on the assumption that neighboring bins have alike time structure [7]. Correlation coefficients for neighboring

[⋆] This work has been supported by the German Research Foundation under Grant No. ME 1170/1.

M.E. Davies et al. (Eds.): ICA 2007, LNCS 4666, pp. 512–519, 2007.

bins then yield a criterion for correct permutation. Another approach uses the unmixing matrices as beamformer. After computation of the directions of arrival for all bins, most of them can be aligned properly [8]. Unfortunately, if there are more than two sensors in a nonuniform array, the computation becomes very difficult.

In this paper we present a new approach for solving the permutation problem based solely on the statistics of the signals. The new algorithm models the single frequency bins using the generalized Gaussian Distribution (GGD) and utilizes the small changes of the shape parameter of the GGD between neighboring bins.

2 Model and Methods

2.1 BSS for Instantaneous Mixtures

In the instantaneous case the mixing process of N sources into N observations can be modeled by an $N \times N$ matrix $\boldsymbol{A}$. Given the source vector $\boldsymbol{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ and assuming negligible measurement noise, the vector of observation signals $\boldsymbol{x}(n) = [x_1(n), \ldots, x_N(n)]^T$ can be described as

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

The separation can be written as a multiplication with a $N \times N$ matrix $\boldsymbol{B}$:

$$\boldsymbol{y}(n) = [y_1(n), \ldots, y_N(n)]^T = \boldsymbol{B} \cdot \boldsymbol{x}(n) \tag{2}$$

The aim of BSS is to find $\mathbf{B}$ from the observed process $\boldsymbol{x}(n)$ so that $\mathbf{BA} = \mathbf{D\Pi}$ where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{D}$ an arbitrary diagonal matrix. These matrices represent the two ambiguities of BSS: (a) the separated signals appear in arbitrary order and (b) they are scaled versions of the sources.

We here consider the well known gradient-based update rule [1]

$$\Delta \boldsymbol{B} \propto (\boldsymbol{I} + E\left\{\boldsymbol{g}(\boldsymbol{y})\boldsymbol{y}^T\right\})\boldsymbol{B} \tag{3}$$

with $\boldsymbol{g}(\boldsymbol{y}) = (g_i(y_i), \ldots, g_n(y_n))$ being a component-wise vector function of nonlinear score functions g_i of the assumed source probability densities $p_i(s_i)$:

$$g_i = \frac{p_i'(s_i)}{p_i(s_i)} \tag{4}$$

In order to achieve good separation performance, the probability density function of the sources has to be known or at least well approximated [9].

2.2 Statistical Source Models and Estimators

Speech signals usually follow a Laplacian distribution. Therefore, for instantaneous mixtures, the nonlinear function $g_i(\cdot)$ reduces to

$$g_i(y) = \frac{\mathrm{sgn}(y)}{\sigma}. \tag{5}$$

Unfortunately, this assumption does not hold for the time-frequency representation $\boldsymbol{X}(\omega_k, n)$. The probability density functions of the components in the bins ω_k can vary in a large range from being sub- to super-Gaussian. A sufficient approximation can be achieved by the generalized Gaussian distribution (GGD) [10]:

$$p_y(y) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|y|/\alpha)^\beta} \tag{6}$$

with $\alpha, \beta > 0$ and the Gamma function given by $\Gamma(y) = \int_0^\infty x^{y-1}e^{-x}dx$. The β-parameter of the GGD describes the overall structure of the distribution. With $\beta = 2$ the GGD reduces to standard Gaussian distribution, with $\beta = 1$ to a Laplacian distribution and with $\beta = 0.5$ to a Gamma distribution. Generally, a large value of β indicates a flat distribution, whereas a small value yields a spiky distribution. α is the generalized measure of the standard deviation.

Usually, nonlinearities for super-Gaussian distributions utilize sigmoidal functions like sgn() or tanh(). Using the GGD, this model can be more generalized. The nonlinear function $g_i(\cdot)$ becomes $g_i(x) = |x|^{\beta-1}\text{sgn}(x)$, and using $\text{sgn}(x) = x/|x|$, we obtain

$$g_i(x) = \frac{x}{|x|^{2-\beta}}. \tag{7}$$

As shown in [9], based on this nonlinear function, even mixtures of sub- and super-Gaussian signals can be separated. Although the authors used fixed values for β they could achieve good results.

The above approach has been extended in [11], where an adaptive algorithm has been proposed. Because, in the blind scenario, the sources are not available and therefore an accurate estimation of β is not possible, the authors proposed to calculate β based on the statistics of the separated signals. They used the method of moments [12] to estimate β after each iteration of (3) and used this new value for the next step. It was shown that the approach leads to improved overall performance in terms of better separation and faster convergence.

2.3 Convolutive Mixtures

In real-world acoustic scenarios, the mixing channels can be modeled by FIR filters of length L, where L can be 2000 or more, depending on the reverberation time and sampling rate. The convolutive mixing model reads

$$\boldsymbol{x}(n) = \boldsymbol{H}(n) * \boldsymbol{s}(n) = \sum_{l=0}^{L-1} \boldsymbol{H}(l)\boldsymbol{s}(n-l) \tag{8}$$

where $\boldsymbol{H}(n)$ 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 $M \geq L-1$ and obtain

$$\boldsymbol{y}(n) = \boldsymbol{W}(n) * \boldsymbol{x}(n) = \sum_{l=0}^{M-1} \boldsymbol{W}(l)\boldsymbol{x}(n-l) \tag{9}$$

with $\boldsymbol{W}(n)$ containing the unmixing coefficients.

Estimating $\boldsymbol{W}(n)$ in the time domain is a very difficult task, because the number of unknowns, MN^2, can reach several tens of thousands. Although there exist approaches to this problem [3,4] the results are not satisfying because of distortions introduced by the unmixing system.

Due to this problem another approach is widely used. After transforming the signals to the frequency domain, for example using the blockwise Short-Time-Fourier-Transform (STFT), the convolution becomes a multiplication [5]:

$$\boldsymbol{Y}(\omega_k, n) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k, n) \tag{10}$$

Instead of estimating all coefficients at once, in the frequency domain it is possible to separate every bin independently. However, since there is the scaling and permutation ambiguity in every bin, we obtain

$$\boldsymbol{Y}(\omega_k, n) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k, n) = \boldsymbol{D}(\omega_k)\boldsymbol{\mathit{\Pi}}(\omega_k)\boldsymbol{S}(\omega_k, n) \tag{11}$$

with $\mathbf{\Pi}(\omega_k)$ being a permutation matrix and $\mathbf{D}(\omega_k)$ a diagonal scaling matrix for frequency ω_k. Therefore, it is necessary to correct the amplitudes and solve the permutation before transforming the signals back to the time domain.

The scaling ambiguity can be resolved to an acceptable degree using the method proposed by Ikeda and Murata [6]. The central idea is to recover the signals as they have been recorded by the sensors. Matusuoka and Nakashima [13] showed that this is the optimal approach, as it minimizes $E\{|y(t) - x(t)|^2\}$. Their Minimal Distortion Principle uses the following unmixing matrix:

$$\boldsymbol{W}'(\omega_k) = \mathrm{diag}(\boldsymbol{W}^{-1}(\omega_k)) \cdot \boldsymbol{W}(\omega_k) \tag{12}$$

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

The correction of the permutation ambiguity is even more important. Even if every bin is perfectly separated, different permutations at different frequencies make both signals appear in every output channel.

3 Resolving the Permutation Ambiguity

One of the first ideas used for the permutation problem is based on the statistics of the separated signals [6,7]. The key assumption is that the envelopes of all bins of one source are highly correlated. With $\boldsymbol{V}(\omega_k, n) = |\boldsymbol{Y}(\omega_k, n)|$ the correlation between two bins k, l is defined as

$$\rho_{qp}(\omega_k, \omega_l) = \frac{\sum_{n=0}^{N-1} \boldsymbol{V}(\omega_k, n)\boldsymbol{V}(\omega_l, n)}{\sqrt{\sum_{n=0}^{N-1} \boldsymbol{V}^2(\omega_k, n)}\sqrt{\sum_{n=0}^{N-1} \boldsymbol{V}^2(\omega_l, n)}} \tag{13}$$

with p, q being the indices of the separated signals. To decide if two bins are permutated equally, the value of

$$r = \frac{\rho_{pp}(\omega_k, \omega_l) + \rho_{qq}(\omega_k, \omega_l)}{\rho_{pq}(\omega_k, \omega_l) + \rho_{qp}(\omega_k, \omega_l)} \tag{14}$$

can be used. If $r > 1$, then the bins are sorted correctly. Otherwise, with $r < 1$, a permutation has occurred. With more than two sources the value of r has to be estimated for all pairs, which means that $N!$ calculations have to be performed.

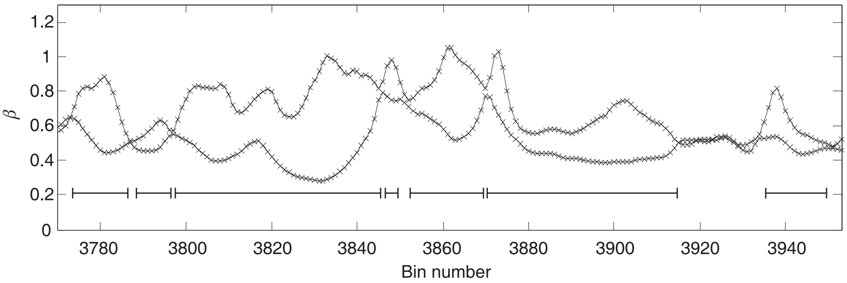

Fig. 1. Beta values of two signals over the frequency index. The detected clusters are indicated with bars ⊢⊣.

Although there are algorithms with less complexity, the practical use is restricted to only few sources [7].

Trying to sort all bins with respect to r for all p and q usually does not work for speech signals. The reason for this is that the key assumption of highly correlated envelopes often does not hold for frequencies which are not close together. Restricting the test to only neighboring frequencies is also not a solution, because at some frequencies, the envelopes of the individual signals do not differ enough to allow for correct sorting. A compromise is the dyadic sorting [7], which starts with pairwise correlation of two neighboring bins and then successively builds groups of bins in a dyadic fashion. This algorithm utilizes the fact that, in a sorted group, a few outliers do not preponderate, and the groups can be aligned properly. But like other proposals that rely only on the correlation of the separated signals, this algorithm suffers if there are too many poorly separated bins close to each other. Because of this, some of the first small groups are often not sorted properly, which then propagates while building the larger groups. The results are block permutations and the separation of the whole signal fails.

4 The Proposed Method

In this paper we propose to use the smoothness of the exponent β of the GGD. For this, we approximate the statistics of every bin by (6). Although the values of β vary in a significant way, the values in neighboring bins do not differ much. Furthermore, two different signals usually have distinct values in most bins, as can be seen in Fig. 1 for a typical situation. However, it can also be seen in Fig. 1 that there are some bins with almost the same value of β, like the bins around 3920. In this range, no differentiation of the two signals is possible on the basis of the value of β. But huge ranges like the bins 3800-3840, can be clustered with certainty. These clusters can be used to correctly de-permute wide frequency ranges. Afterwards, the remaining bins can be de-permuted using alternative methods.

The proposed method consists of three parts: (1) estimation of the boundaries of the clusters, (2) calculation of the permutation between the clusters, and (3) aligning the remaining bins. The algorithm is at first derived for two signals and then extended for multiple signals.

4.1 Calculation of the Cluster Boundaries

The first step is to estimate the β values for all bins of both estimated sources. One possibility is to estimate this parameter in every iteration of the BSS algorithm mentioned in Section 2.1. Alternatively, any other known BSS algorithms can be used, because β can be also estimated after separation.

The second step is to make a simple grouping. The bins are compared pairwise: the ones with higher values of β are assigned to one and the ones with lower values are assigned to the other source.

The third step is to determine the actual clusters. The idea for a simple and fast method is the following: Take an existing cluster and find out if the neighboring bin can be added to it. The decision is based on the assumption of the values of β being distinct and smooth.

The actual implementation is as follows:

1. Start at bin $l = 1$.
2. Test by comparing the β-values if the next bin $l + 1$ can be added.
3. If yes, then add this bin to the cluster, increase l and go to Step 2.
4. If not, then the end of the cluster has been found. If the cluster is large enough, mark it as being correctly permutated. Increase l, mark l as the beginning of a new cluster and go to Step 2.

The result of this algorithm is shown in Fig. 1.

If there are more than two signals, the algorithm can be extended. For this, the β values are sorted, and the two largest ones are assigned to $\beta_H(\omega_k)$ and $\beta_L(\omega_k)$, respectively. After clustering and removing $\beta_H(\omega_k)$, the same procedure can be applied to the remaining bins. An analogous procedure can be applied to the bottommost values for increased performance.

4.2 Calculation of Cluster Correlations and Aligning the Remaining Bins

The next step after the identification of the clusters is to determine the permutation between them. As the gaps between clusters are usually much smaller than the clusters themselves, the assumption of highly correlated envelopes can be used. Here we follow the idea of dyadic sorting and calculate the value of r, as defined in (14), for all combinations of all bins of two clusters. As the bins within the clusters are de-permuted with high confidence, the correct permutation between clusters can be determined by the highest or lowest value of r, as for the dyadic sorting in [7].

After calculating the correct permutation for the clusters, the remaining bins also have to be aligned. Again, a comparison of the correlation coefficients r for these bins with all coefficients for the bins in the neighboring clusters can be used.

5 Simulations

In a first simulation, the algorithm has been used on unmixed audio signals, which have been arbitrarily permuted in the frequency domain. This should simulate the behavior of the algorithm in ideal conditions, as if the blind separation stage in each frequency bin would be able to work perfectly. In this case, the algorithm was able to correctly de-permute all bins.

When using real-world data, the separation in the single bins is not always perfect. Therefore, the estimation of correct permutations is harder. In the experiments we used a data set where the individual contributions from the sources to the microphones were available [14], and the separation performance could be estimated using the signal-to-interference ratio

$$SIR_{y_i} = 10log_{10} \frac{E[(g_{ii}(n) * s_i(n))^2]}{E[(\sum_{j=1,j\neq i}^{N} g_{ij}(n) * s_j(n))^2]} \tag{15}$$

with $g_{ij}(n) = w_i(n) * h_j(n)$. In Fig. 2, the separation performance for the single bins is given.

As the individual sources are known, the best possible unmixing can be estimated. In Fig. 3, the difference between this best approach and the result of the proposed algorithm is shown. As we can see, above 300 Hz the proposed algorithm produces exactly the same output as the ideal de-permutation. Below this frequency there occur permutations, but this is a frequency range where the separation has failed in several bins. Further inspection of the data showed that the estimation of clusters worked, but the cluster correlations were incorrect. This is a typical behavior for correlation-based approaches, when the separation is not perfect. The overall performance with swapped bins is an SIR of 13.16 dB. When leaving the low frequencies out and recovering only the signal components above 300 Hz, the overall performance increases to 20.03 dB.

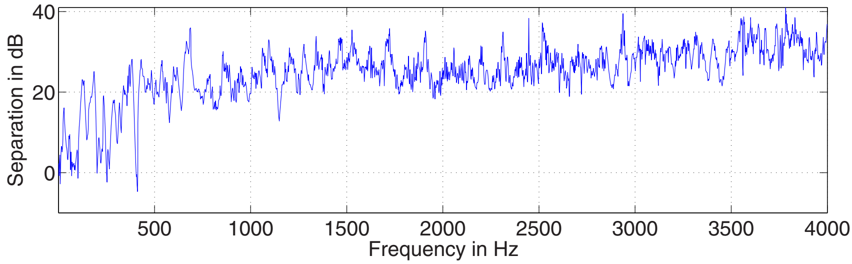

Fig. 2. Separation Performance

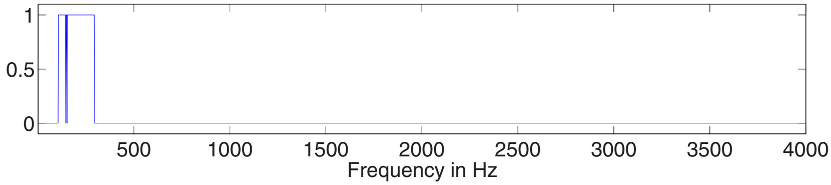

Fig. 3. Swap errors

6 Summary

In this paper, we presented a new approach for resolving the permutation problem, which occurs in convolutive blind source separation. For this we modeled every bin using the generalized Gaussian Distribution and used the exponent β for estimating the correct permutation. The performance of the algorithm has been studied on artificial and real word data.

References

1. Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.) Advances in Neural Information Processing Systems, vol. 8, pp. 757–763. The MIT Press, Cambridge (1996)
2. Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Computation 9, 1483–1492 (1997)
3. Douglas, S.C., Sawada, H., Makino, S.: Natural gradient multichannel blind deconvolution and speech separation using causal fir filters. IEEE Trans. Speech and Audio Processing 13(1), 92–104 (2005)
4. Aichner, R., Buchner, H., Araki, S., Makino, S.: 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 (April 2003), pp. 987–992 (2003)
5. Smaragdis, P.: Blind separation of convolved mixtures in the frequency domain. Neurocomputing 22(1-3), 21–34 (1998)
6. Ikeda, S., Murata, N.: A method of blind separation based on temporal structure of signals. In: Proc. Int. Conf. on Neural Information Processing, pp. 737–742 (1998)
7. Rahbar, K., Reilly, J.: A frequency domain method for blind source separation of convolutive audio mixtures. IEEE Trans. Speech and Audio Processing 13(5), 832–844 (2005)
8. 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 and Audio Processing 12(5), 530–538 (2004)
9. Choi, S., Cichocki, A., Amari, S.: Flexible independent component analysis. In: Constantinides, T., Kung, S.Y., Niranjan, M., Wilson, E. (eds.) Neural Networks for Signal Processing VIII, pp. 83–92 (1998)
10. Gazor, S., Zhang, W.: Speech probability distribution. IEEE Signal Processing Letters 10(7), 204–207 (2003)
11. Kokkinakis, K., Nandi, A.K.: Multichannel Speech Separation Using Adaptive Parameterization of Source PDFs. In: ICA 2004. LNCS, vol. 3195, pp. 486–493. Springer, Heidelberg (2004)
12. Varanasi, M.K., Aazhang, B.: Parametric generalized gaussian density estimation. Acoustical Society of America Journal 86(4), 1404–1415 (1989)
13. Matsuoka, K.: Minimal distortion principle for blind source separation. In: Proceedings of the 41st SICE Annual Conference. vol. 4, pp. 2138–2143 (August 5-7, 2002)
14. http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html