A NEW CLUSTERING APPROACH FOR SOLVING THE PERMUTATION PROBLEM IN CONVOLUTIVE BLIND SOURCE SEPARATION

Radoslaw Mazur, Jan Ole Jungmann, and Alfred Mertins

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

ABSTRACT

In this paper we propose a new clustering approach for solving the permutation ambiguity in convolutive blind source separation. After the transformation to the time-frequency domain, the problem of separation of sources can be reduced to multiple instantaneous problems, which may be solved using independent component analysis. The drawbacks of this approach are the inherent permutation and scaling ambiguities, which have to be corrected before the transformation to the time domain. Here, we propose a new method that allows for aligning up to several hundreds of consecutive bins into clusters. The depermutation of these clusters using some known techniques is then much easier than the original problem. The performance of the proposed method is evaluated on real-room recordings.

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

1. INTRODUCTION

In the case of linear and instantaneous mixtures of non-Gaussian signals, blind separation may be performed using the Independent Component Analysis (ICA). For this, numerous algorithms have been proposed [1, 2, 3]. The methods are called blind, as typically neither the sources nor the mixing system are known.

For real-world mixtures of acoustic signals, as for example speech, this simple approach fails. Due to the finite speed of sound and multiple reflections in closed rooms, the signals arrive at the microphones multiple times with different lags. This convolutive mixing process can be modeled using FIR filters. In typical scenarios, as for example office rooms, the length of these mixing filters can reach up to several thousand coefficients. Such mixtures can be separated using a set of unmixing FIR filters with at least the same length.

The unmixing filters can be calculated directly in the time domain [4, 5]. Unfortunately, this method suffers from high computational load and often poor convergence. Therefore, another approach is widely used: With the transformation to the time-frequency domain the convolution becomes a multiplication, and an instantaneous ICA algorithm can be used independently in each frequency bin. However, this simplification has a major disadvantage, as each bin can be arbitrarily scaled and permuted. These ambiguities need to be solved before the transformation to the time domain, as otherwise the separation process will fail.

The correction of scaling is needed, as otherwise only a filtered version is recovered. A typical solution is the minimal distortion principle [6] or inverse postfilters [7]. These methods accept the filtering done by the mixing system without adding new distortions. Other approaches solve the scaling ambiguity with the aim of filter shortening [8] or shaping [9, 10].

The correction of the random permutation of the discrete frequency bins is even more important as otherwise the whole separation process will fail. The depermutation algorithms can be organized in two major groups. The first group relies on the properties of the unmixing matrices. For example, they can be interpreted as beamformers, and the direction of arrival (DOA) information is used for a depermutation criterion [11]. Alternative formulations evaluate directivity patterns [12] or time differences of arrivals (TDOA) [13, 14, 15]. These approaches assume specific directions for the sources, and, in the case of reverberation, usually only a small part of the bins the TDOA can be estimated with high enough confidence for an effective depermutation. The remaining bins have to be aligned in a second stage using some other approach.

The second group of algorithms uses the alike time structure of the separated bins. The early approaches often exploited the assumption of high correlation between neighboring bins [7]. The dyadic sorting, as proposed in [16], also allows for a more robust depermutation scheme. The dyadic sorting has also been used in [17] with combination of a sparsity criterion. 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 [18].

In this work, we follow the argumentation from [11]. There the authors state that the direction of arrival, which is closely related to TDOA, is robust only for bins where this direction can be estimated with high confidence. Additional robustness is based on the fact that these bins are usually distributed over the whole range. After depermutation, the remaining bins are then aligned using the correlation approach. This is justified by high similarity of neighboring bins, which but is typically not true for distant bins.

In this paper, we propose to use the high similarity of neighboring bins for achieving a robust depermutation for some clusters of neighboring bins. The information of this depermutation is then used for estimation of an average TDOA for these clusters, which is then used for a calculation of a robust depermutation for all bins.

The proposed method is simple, as it does not require any confidence functions or harmonics analysis as in [11]. Additionally, the clustering procedure using a greedy algorithm is also simpler than the one proposed in [13]. The computational cost of the new approach is almost negligible compared to the ICA-stage. The robustness of the proposed method will be shown on real-world examples.

978-1-4799-0972-8/13/$31.00 ©2013IEEE

2. MODEL AND METHODS

The instantaneous mixing and unmixing processes form the basis for the convolutive case. Both methods will be described in the following.

2.1. BSS for instantaneous mixtures

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

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

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

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

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 $\boldsymbol{g}(\boldsymbol{y}) = (g_1(y_1), \ldots g_n(y_n))$ 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.

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:

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

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.

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

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

with K being 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. Using the minimal distortion principle [6] to resolve this ambiguity, the unmixing matrix reads

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

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

Without correction of the permutation, different signals will be restored at different frequencies and the whole separation process will fail. In the next section, we will propose a new scheme for calculation of the depermutation.

3. DEPERMUTATION ALGORITHMS

In this section, we describe the basic algorithms for depermutation. At first the basics of the correlation approach will be summarized and a new robust clustering method will be presented. Then the basics of TDOA will be shown and the needed extension for the average cluster TDOA will be discussed. The two new extensions then lead to a robust and fast depermutation algorithm.

3.1. Correlation approaches

Many depermutation algorithms are based on the statistics of the separated signals. For example, the assumption of high correlation of envelopes of neighboring bins yields a simple depermutation criterion [7]. 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 q-th element of $\boldsymbol{V}(\omega_k, \tau)$, and $\mathcal{T}$ is the number of frames. The alignment of the bins is made on the basis of the ratio

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

With $r_{kl} > 1$ the bins are assumed to be correctly aligned, and otherwise a permutation has occurred. The simple method, where consecutive bins are examined is not robust, as single wrong permutations lead to whole blocks of falsely permuted bins.

The dyadic sorting from [16] depermutes pairs of bins using (11). 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 usually do not outbalance the majority.

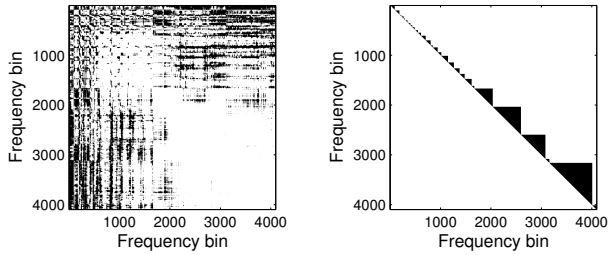
3.2. Clustering using correlation

The method of dyadic sorting accepts some wrongly permuted bins. In order to increase robustness, we present a simple greedy clustering procedure, which assures that a cluster contains only correctly aligned bins:

Step 1. Initialize the bin counter to the first bin: $k = 1$.

Step 2. Start a new cluster at k.

Step 3. Calculate the correlation and the alignment coefficient $r_{k+1,l}$ according to (11) of bin $k+1$ to all bins in the current cluster.



(a) (b)

Figure 1: Visualization of the proposed clustering method. a) Alignment coefficients for all bins from (11). White points ($r_{kl} > 1$) indicate a correct and black ($r_{kl} < 1$) a false decision for a perfectly depermuted case. b) The detected clusters with nonambiguous coefficients.

Step 4. If all $r_{k+1,l} > 1$ then add the k-th bin to the current cluster. $k \leftarrow k + 1$. Continue at Step 3.

Step 5. If all $r_{k+1,l} < 1$ then the k-th bin can be added to the current cluster after depermutation. $k \leftarrow k + 1$. Continue at Step 3.

Step 6. Otherwise, as the depermutation information is ambiguous, start a new cluster, $k \leftarrow k + 1$, and continue at Step 3.

In Figure 1 the result of this procedure for the dataset form [19] is presented. In Figure 1(a) the alignment coefficients for all bins, calculated using (11) are shown. White points ($r_{kl} > 1$) indicate a correct and black points ($r_{kl} < 1$) a false decision for a perfectly depermuted case. This matrix is symmetric, and the high number of black points in the upper right corner indicate a very low similarity between the low and high frequencies for the signals. In Figure 1(b) the result of the greedy clustering procedure is shown. The black areas indicate the bins, where all alignment coefficients are unambiguous and can be joined to a cluster. In this case there are only 36 clusters. This is a major simplification compared to the original problem with 4097 discrete bins.

3.3. DOA and TDOA

For the case of no spatial aliasing, the authors of [11] calculate the direction of arrival for the 2×2-case for a single bin as

$$\Theta_i(\omega_k) = \arccos \frac{\arg\left(\frac{[H(\omega_k)]_{1i}}{[H(\omega_k)]_{2i}}\right)}{2\pi f c^{-1} \Delta_d} \tag{12}$$

with $\Theta_i(\omega_k)$ being the angle relative to the microphone array, $[H(\omega_k)]_{li}, l \in \{1, 2\}$ the coefficients of the mixing matrix corresponding to the i-th source, f the frequency, c the speed of sound, and Δ_d the distance between the microphones.

By knowing the array dimensions and using an additional confidence function, which indicates the quality of the estimation of the direction, a clustering of bins could be achieved. The remaining bins could be depermuted using the correlation approach with additional analysis of the harmonics structure.

The TDOA approach, as for example in [13], does not need the knowledge of the exact geometry of the recording array. The time difference can be calculated as

$$\text{TDOA}_i(\omega_k) = \frac{1}{2\pi f} \arg\left(\frac{[H(\omega_k)]_{1i}}{[H(\omega_k)]_{2i}}\right). \tag{13}$$

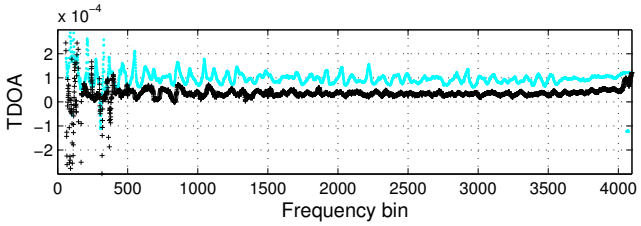


Figure 2: TDOAs of single frequency bins for a 2×2 case.

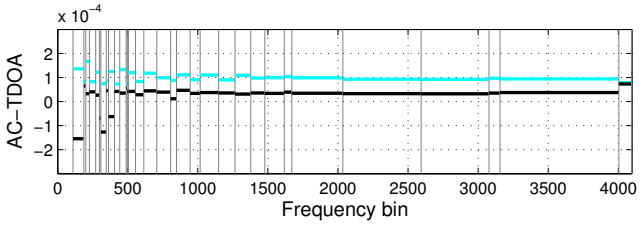


Figure 3: Average cluster TDOAs. Vertical lines indicate cluster boundaries.

In Figure 2 the results of the TDOA calculation for the data set from [19] are shown. In this case, it is quite easy to calculate the depermutation for almost all bins above 520. But below bin 520, there are numerous bins whose TDOA is ambiguous. No simple clustering is possible as the TDOAs scatter beyond the averages of the other source.

3.4. Average cluster TDOA

Using the information form the clustering allows for a much more robust calculation of the average cluster TDOA:

$$\text{acTDOA}_i(C_m) = \text{mean}(\text{TDOA}_i(\omega_k)), \quad k \in C_m \tag{14}$$

with C_m being a set of indices of the bins of the m-th cluster. In Figure 3 the result of this calculation is shown. The average cluster TDOA is much more consistent and the scattering of values is significantly reduced. This allows for a simple final clustering procedure that aligns $\text{acTDOA}_j(C_m)$ on $\text{acTDOA}_i(C_M)$, with C_M being the largest cluster, by minimizing

$$\sum_{i=1}^{2} (\text{acTDOA}_j(C_m) - \text{acTDOA}_i(C_M))^2, \quad j \in \{1, 2\}. \tag{15}$$

The overall procedure is much simpler than [11]. The calculation of the clustering and average cluster TDOAs is easy and does not require any confidence functions or harmonics analysis. The computational cost is almost negligible compared to the ICA-stage.

4. SIMULATIONS

The experiments using the proposed algorithm have been performed using real-world data available at [19]. The setup was chosen to be similar to that in [17] and [11]. With a sampling rate of 8 kHz, the FFT length was chosen to be 8192, and a 2048 point Hann analysis window has been used.

This dataset contains two speech signals (one male, one female) in a low reverberant room. As the signals do not have meaningful energy below 110 Hz, only bins above this frequency are taken into consideration. The separation in the ICA stage is successful for almost all bins, and a non-blind depermutation algorithm results in a very good separation ratio of 17.6 dB, as shown in Table 1.

Table 1: Comparison of the results for different depermutation algorithms in terms of separation performance in dB.

Algorithm	SIR in dB
Proposed	17.3
DOA-Approach [11]	17.3
Sparsity approach [17]	15.4
Dyadic sorting [16]	2.7
Non blind	17.6

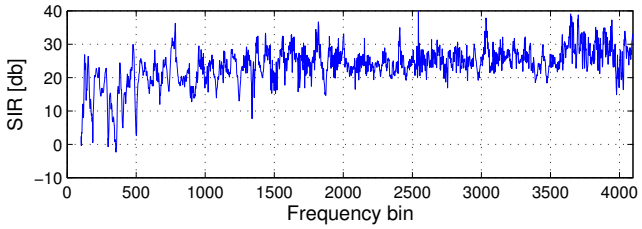


Figure 4: Separation performance over frequency after final alignment for the proposed algorithm. Only few poorly separated bins around 350 are wrongly permuted. These errors do not have any significant impact on the overall SIR.

The new clustering procedure was able to group the 4097 frequency bins into 36 clusters with only a few poorly separated bins being wrong. After calculation of the average cluster TDOAs the remaining grouping using (15) could depermute all clusters correctly. The overall separation performance is 17.3 dB, which is the same, as in [11] and only slightly less than the ideal nonblind depermutation with 17.6 dB. In Figure 4 the final result of the separation performance per bin is shown. Here, the few wrongly permuted bins around 350 are visible, as they have a negative SIR. The separation of these bins with less than 5 dB failed at the ICA-stage, and both permutations are almost equally bad.

5. CONCLUSIONS

In this paper we propose a new clustering approach for solving the permutation ambiguity in convolutive blind source separation. The new approach is to group frequency bins using the time structure into nonambiguous clusters. The calculation of the average cluster TDOA allows then for a simple and fast depermutation of all bins. The performance of the approach has been shown on real-world example.

6. 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*, vol. 8, MIT Press, Cambridge, MA, 1996.

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

[7] 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.

[8] 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.

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

[10] ——, "A method for filter equalization in convolutive blind source separation," in *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation*, St. Malo, France, Sept. 2010.

[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] 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.

[13] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components with estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. vol.15, no.5, pp. 1592–1604, July 2007.

[14] F. Nesta and M. Omologo, "Approximated kernel density estimation for multiple tdoa detection," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 149 –152.

[15] ——, "Generalized state coherence transform for multidimensional tdoa estimation of multiple sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 246 –260, Jan 2012.

[16] 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.

[17] R. Mazur and A. Mertins, "A sparsity based criterion for solving the permutation ambiguity in convolutive blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 1996–1999.

[18] ——, "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.

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