A Modified Clustering Approach for Solving the Permutation Problem in Convolutive Blind Source Separation

Radoslaw Mazur, Jan Ole Jungmann, and Alfred Mertins
Universität zu Lübeck, Institute for Signal Processing, Ratzeburger Allee 160, D-23562 Lübeck
Email: {mazur, jungmann, mertins}@isip.uni-luebeck.de

Abstract

In this paper we propose a modification to 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 convolutively mixed sources can be reduced to multiple instantaneous problems, which may be solved using independent component analysis. The drawbacks of this approach are the so called permutation and scaling problems, which have to be corrected before the transformation to the time domain. Here, we use a new method that allows for aligning up to several hundreds of consecutive bins into clusters and propose a modification which allows for an even more effective clustering. The depermutation of these clusters using some known techniques is then much easier than the original problem.

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].

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. In this case, the separation is only possible when the unmixing system is again a set of FIR filters.

As the direct calculation of the unmixing filters in time domain is very demanding, time-frequency approaches are often used. Here, the convolution becomes a multiplication and each frequency bin can be separated using an instantaneous method. 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 [3]. This method accepts the filtering done by the mixing system without adding new distortions.

The random permutation of the single frequency bins has an even bigger impact. Without a correct alignment, different signals appear in the single outputs and the whole separation process fails.

Many different approaches for solving this problem have been proposed. Often, the time structure of the separated bins is used and the assumption of high correlation between neighboring bins is utilized. This has been used for example in [2]. 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 [5].

The second type of approach 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 [8]. Alternatively, time differences of arrivals (TDOA) have been used in [7, 6].

In [8, 7] the approach is based on estimating DOAs and TDOAs. Frequency bins, whose TDOAs are estimated with high confidence, are used for aligning the remaining bins using the correlation method. In [4] this approach has been reversed: First calculate clusters of correctly depermuted bins using the correlation and then use the average TDOAs of these bins for final arrangement. The first stage of the algorithm is based on the fact, that large portions of the frequency bins can be depermuted with very high confidence when creating clusters with unambiguous correlations.

In this work we modify the calculation of these clusters, as the requirement of unambiguous correlations is too hard and some easing on this criterion allows for bigger clusters.

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 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.

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:

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

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

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

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:

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

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:

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

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

The scaling can be solved the minimal distortion principle [3]. A modified approach for the permutation problem will be shown in the next section.

Depermutation Algorithm

In this work, we follow the two stage approach from [4]. The first stage uses the time structure of the separated bins to calculate correlation coefficients

$$r_{kl} = \frac{\rho_{pp}(\omega_k, \omega_l) + \rho_{qq}(\omega_k, \omega_l)}{\rho_{pq}(\omega_k, \omega_l) + \rho_{qp}(\omega_k, \omega_l)}, \qquad (9)$$

with $\rho_{pq}(\omega_k, \omega_l)$ being the correlation of the outputs p and q at frequencies ω_k and ω_l. In [4] a simple greedy algorithm for unambiguous clustering has been used: With C_m being a set of indices of the m-th cluster add the following bin k only if

$$r_{kl} > 1, \text{ for all } l \in C_m \qquad (10)$$

and start a new cluster otherwise. This criterion assures that all bins in a cluster are alike correlated. Large clusters usually indicate, that their the bins are correctly depermuted.

Since the correlation assumption is not valid for distant bins, this method can not be used for depermuting the clusters. Therefore, the second stage in [4] uses the TDOAs

$$\text{TDOA}_i(\omega_k) = \frac{1}{2\pi f} \arg\left(\frac{[H(\omega_k)]_{1i}}{[H(\omega_k)]_{2i}}\right) \qquad (11)$$

for calculation of the average cluster TDOA

$$\text{acTDOA}_i(C_m) = \text{mean}(\text{TDOA}_i(\omega_k)), \quad k \in C_m. \quad (12)$$

The aligning of the $\text{acTDOA}_i(C_m)$ on $\text{acTDOA}_i(C_M)$, with C_M being the largest cluster finally yields the depermutation algorithm for all frequencies.

Table 1: Comparison of the cluster sizes for the different Algorithms

Clustering	# Clusters	Average Cluster Size
None	4097	1
Old (10)	36	113
New (13)	27	151

Here, we propose to ease the requirement (10) in order to create larger clusters and make the depermutation method more robust. Based on the assumption, that frequency bins, which are further away, do not necessarily correlate positively and the observation, and that with bigger clusters the criterion (10) is violated only for a very few bins, we use the criterion

$$r_{kl} > 1, \text{ for all } l \in C_m, |k-l| < \tfrac{1}{2}|C_m| \qquad (13)$$

with $|C_m|$ being the number of bins in C_m. Here we assure, that every bin is positively correlated with all bins in the cluster, which are not further away than half the cluster size.

The results of the new clustering method are compared on the same dataset as in [4]. The number of clusters could be reduced and the average cluster size increased. The separation performance did not change.

Conclusions

In this work, we presented a modification of an algorithm for solving the permutation ambiguity in convolutive blind source separation. The new algorithm allows for creating of bigger clusters of correctly depermuted bins, which make the overall procedure more robust.

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*, volume 8, MIT Press, Cambridge, MA, 1996.

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

[3] K. Matsuoka. Minimal distortion principle for blind source separation. In *Proceedings of the 41st SICE Annual Conference*, volume 4, pages 2138–2143, 5-7 Aug. 2002.

[4] R. Mazur, J. O. Jungmann, and A. Mertins. A new clustering approach for solving the permutation problem in convolutive blind source separation. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, Oct. 2013.

[5] 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*, 17(1):117–126, Jan. 2009.

[6] F. Nesta and M. Omologo. Generalized state coherence transform for multidimensional tdoa estimation of multiple sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):246 –260, Jan 2012.

[7] 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.15, no.5:1592–1604, July 2007.

[8] 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*, 12(5):530–538, Sept. 2004.