OPTIMIZED DYADIC SORTING FOR SOLVING THE PERMUTATION AMBIGUITY IN ACOUSTIC BLIND SOURCE SEPARATION

Radoslaw Mazur, Jan Ole Jungmann, and Alfred Mertins

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

ABSTRACT

In this paper, we propose a modification to dyadic sorting scheme used for the permutation problem in convolutive blind source separation. In the frequency domain, the problem of separation of sources can be reduced to multiple instantaneous problems, which are easily solvable using independent component analysis. However, this simplified method leads to the problem of correctly aligning and scaling of the single frequency bins. These ambiguities need to be solved before the transformation to the time domain, as otherwise the separation process will fail. In this paper we combine dyadic sorting with an optimized way of calculation of correlation coefficients by using spectral summation. The improved performance will we shown on real world examples.

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

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 case, numerous algorithms have been proposed [1, 2, 3]. The methods are called blind, as typically neither the sources nor the mixing system is known.

When dealing with 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 multiple times with different lags. This mixing process is convolutive, and 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.

These unmixing filters can be calculated directly in the time domain [4, 5], but this method suffers form high computational load and often poor convergence. Therefore, an other 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. But this simplification has a major disadvantage, as each bin can be arbitrarily scaled and permuted. These ambiguities needs 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]. This method accepts 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 rely on the properties of the unmixing matrices. For example, they can be interpreted as beamformer, and the direction information is used for a depermutation criterion [11]. Alternative formulations evaluate directivity patterns [12, 13] or time difference of arrivals [14, 15] As these approaches assume specific directions for the sources, they usually fail in the presence of high reverberation and noise. In [16] the authors exploit the sparsity of the unmixing filters. However, in typical real world examples, only the first part of a filter exhibits this property.

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]. This method has been extended in [17, 18] to use activity patterns. For speech signals, which are sparse in the time-frequency domain, they usually yield a better depermutation criterion than the plain correlation technique. The dyadic sorting, as proposed in [19], also allows for a more robust depermutation scheme. The dyadic sorting has also been used in [20] 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 [21] and [22].

In this work we propose a new approach, which is based on the modified method from [20], where a full time domain representation of the single bins has been used. With the method of spectral summation only one coefficient needs to be calculated for each depermutation decision, which leads to a higher robustness. The performance will be even more improved by employing the activity patterns as in [17, 18] and will be shown on real world examples.

2. MODEL AND METHODS

The instantaneous mixing and unmixing process is the basis for the convolutive one. Both methods will be described in the following.

2.1. BSS for instantaneous mixtures

The mixing of N sources into N observations can modeled by an $N \times N$ matrix $\mathbf{A}$. With the assumption of negligible measurement noise, the observation signals $\boldsymbol{x}(n) = [x_1(n), \ldots, x_N(n)]^T$ are given by

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

with $\boldsymbol{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ being the source vector. The separation is again a multiplication with a matrix $\mathbf{B}$:

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

with $\boldsymbol{y}(n) = [y_1(n), \ldots, y_N(n)]^T$. The estimation of $\mathbf{B}$ is solely based on the observed process $\boldsymbol{x}(n)$. With $\mathbf{BA} = \mathbf{D\Pi}$ the separation is successful with $\mathbf{D}$ and $\mathbf{\Pi}$ standing for the two ambiguities of BSS: $\mathbf{\Pi}$ being a permutation matrix models the arbitrary order of the signals and $\mathbf{D}$ being a diagonal matrix stands for the unknown scaling of the outputs.

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 componentwise 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 [23].

2.2. Convolutive mixtures

In real-world acoustic scenarios it is necessary to consider reverberation. In this case, 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 taps. The convolutive mixing model reads

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

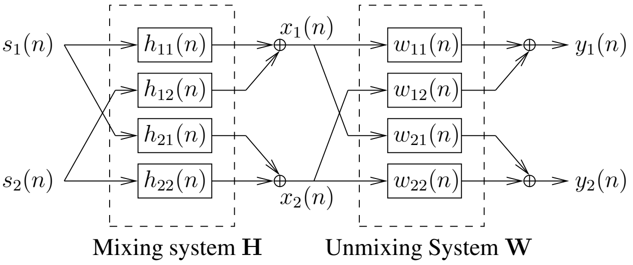

Fig. 1. BSS model with two sources and sensors.

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.

In order to simplify the separation process, the transformation to the time-frequency domain is often used. Using the short-time Fourier transform (STFT), the convolution approximately becomes a multiplication [24]:

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

where K is the FFT length. This approach allows for an independent estimation of an unmixing matrix in each frequency bin by an instantaneous ICA. The drawback is the possible permutation and arbitrary 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. One solution is the minimal distortion principle [6], which does not add any new distortion while accepting the filtering done by the mixing system. 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 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 ALGORITHMS

In this section we describe some basic depermutation algorithms and their extensions. Then, a new combination with an improved robustness will be derived.

3.1. Correlation approach

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.

3.2. Dyadic sorting

An improvement to the simple sequential depermutation which uses dyadic sorting has been proposed in [19]. This method compares at the first stage neighboring bins using (11). After this initial pairwise alignment, those pairs are again compared in the next stage. This approach is repeated in the next stage with the quadruples and so on, until all frequency bins have been sorted. An example of this scheme is shown in Fig. 2, where eight bins are sorted. The main assumption of this method is that single wrong permutations, which may occur at lower stages, do not interfere at higher stages.

Still, in the work of [19], the depermutation at higher stages is essentially based on the correlation of the single bins within the larger blocks. With the correlation assumption not always holding between more distant frequency bins and by simple averaging of r_{kl} for multiple bins, the depermutation is not robust.

In [20] a modification of the dyadic sorting has been introduced. Beside the use of a sparsity criterion, an improvement to the calculation of the ratio r_{kl} has been proposed. Here, the coefficients are calculated on the time representation $z(\omega_{ab}, n)$ of the bins in the frequency range $[a, b]$ of $Y(\omega,\tau)$. Using this time representation allows for calculation of a single coefficient r_{kl} for a whole range of frequency bins. This approach avoids the problematic averaging as used in [19]. By consecutive depermutation and summation of pairs which represent whole blocks of frequency bins, the procedure yields, besides the permutation information, also the reconstructed signals.

3.3. Activity patterns

In [17, 18] an alternative method to the correlation of the envelopes has been proposed. Here, the authors exploit the sparsity of speech signals, and compute the dominance of the i-th single separated signal as

$$\text{powRatio}_i(\omega_k,\tau) = \frac{\|\mathbf{w}_i(\omega_k)y_i(\omega_k,\tau)\|^2}{\sum_{k=1}^{N}\|\mathbf{w}_k(\omega_k)y_k(\omega_k,\tau)\|^2} \tag{12}$$

The values of these activity patterns are normalized to $0 \leq \text{powRatio}_i \leq 1$. A value of approximately one indicates a dominance of the given signal, while low values denote the dominance of some other signals. The comparison of activity patterns instead of envelopes by (10) and (11) is usually more robust. However, this assumption is violated when one signal is dominant the whole time. This can be problematic for speech signals, which usually have no energy below the fundamental frequency. An example will be given in the experiments section.

In [17, 18] the depermutation algorithm is based on calculation of centroids which roughly represent the average activity pattern for a signal. Alternatively, the authors also propose a two stage algorithm with multiple centroids which are optimized to represent parts of the signals.

3.4. New algorithm

Here, we propose to combine the above presentend methods in a new way. The main idea is to use the dyadic sorting, but with some modifications in order to achieve more robustness.

The first modification is the calculation of r_{kl} using time domain representations of the frequency bins in a similar way as in [20]. These representations can be computed using a non-decimating DFT-filter bank. By using the method of spectral summation, the time representation of a block of bins can be calculated by a simple summation

$$z(\omega_{ab},\tau) = \sum_{k=a}^{b} Y(\omega_k,\tau) \tag{13}$$

and the envelope is given by $v(\omega_{ab},\tau) = \|z(\omega_{ab},\tau)\|$. Depending on the analysis window of the DFT-filter bank, some previous modulations of the single bins may be needed.

The second modification is the use of activity patterns as proposed in [17] instead of the envelopes of the signals. Here, we calculate the activity of the restored bandpass signals as

$$\text{powRatio}_i(\omega_{ab},\tau) = \frac{\|z_i(\omega_{ab},\tau)\|^2}{\sum_{k=1}^{N}\|z_k(\omega_{ab},\tau)\|^2} \tag{14}$$

Using the modified dyadic sorting, every depermutation decision is based on one coefficient, which makes it more robust than the averaging in the original one. Additionally, the calculation of centroids and the clustering of the frequency bins as in [17] can be avoided.

After the calculation of all stages of the dyadic sorting scheme, the separated signals can be obtained from the real part of the last summation. Additionally, the track keeping of the single decisions yields the discrete permutation information, which can be used for comparison with other algorithms.

4. EXPERIMENTS

The experiments using the proposed algorithm have been performed using real world data available at [25] and [26]. The setup was chosen to be similar to that in [20] 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.

The first dataset contains two speech signals (one male, one female) in a low reverberant room. The separation in the ICA stage is successful and non-blind depermutation algorithm results in a very good separation ratio of 18.4 dB as shown in Table 1. With the low reverberation, the direction of arrival approach and the $\alpha\beta$-algorithm from [21] are both able to depermute almost all bins, and the separation performance is almost as good as in the non-blind case. The sparsity dyadic depermutation from [20] is able to depermute most of the bins, and has a separation performance of 15.4 dB. The plain dyadic sorting based on correlation fails. The proposed algorithm is able to depermute almost all bins correctly. Close inspection shows, that merely some bins in the lowest rage are mixed up. Below 110 Hz, where both signals have no meaningful energy, the wrong permutations have no impact. In the range between 110 and 175 Hz only the male signal has some significant energy and therefore is dominant the whole time. Unfortunately, as stated in section 3.3, this contradicts the assumptions needed for the sorting using activity pattern. The result is a permutation in these frequencies and a quite low performance of 8.3 dB as shown in first row of Table 1. By simply removing these frequencies from the results, the overall performance increases to 19.1 dB. Although this is not a completely fair comparison, it shows, that the sorting of the other frequency bins is very good.

The second dataset is recorded in a higher reverberant room. In this case the direction of arrival approach fails, as the assumption of the single direction for every source is not valid. The dyadic sorting scheme from [19] based on correlation also fails. The same is true for the $\alpha\beta$-Algorithm.

The sparsity based approach form [20] performs quite well with 8.1 dB separation performance. The new proposed algorithm is also able to depermute most of the bins and has a similar performance with 8.2 dB. As both signals have similar energy profile in the lower frequencies, the above mentioned modification do not yield any improvements.

5. CONCLUSIONS

In this paper we proposed a modification to the dyadic sorting scheme used for the permutation problem in the convolutive blind source separation. The new criterion uses activity patterns instead of correlation of envelopes and a modified dyadic sorting method, where only one coefficient is used for a depermutation decision. The performance has been shown on real world examples.

Table 1. Comparison of the results for different depermutation algorithms in terms of separation performance in dB. Dataset 1 is taken from [25]. Dataset 2 is recorded in higher reverberant room [26].

Algorithm	Dataset 1	Dataset 2
Proposed (plain)	8.3	8.2
Proposed (modified)	19.1	8.3
Sparsity approach [20]	15.4	8.1
Dyadic sorting [19]	2.7	3.0
DOA-Approach [11]	17.3	3.4
$\alpha\beta$-Algorithm [21]	18.4	0.3
Non blind	18.4	9.4

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

[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] 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.

[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] R. Mazur and A. Mertins, "A method for filter shaping in convolutive blind source separation," in *Independent Component*

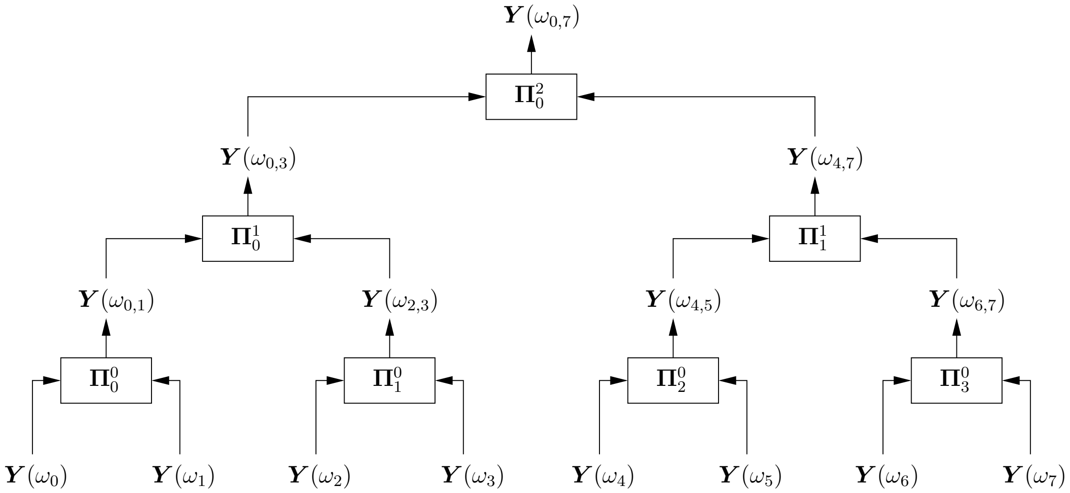
$\boldsymbol{Y}(\omega_{0,7})$
$\boldsymbol{\Pi}_0^2$
$\boldsymbol{Y}(\omega_{0,3})$
$\boldsymbol{Y}(\omega_{4,7})$
$\boldsymbol{\Pi}_0^1$
$\boldsymbol{\Pi}_1^1$
$\boldsymbol{Y}(\omega_{0,1})$
$\boldsymbol{Y}(\omega_{2,3})$
$\boldsymbol{Y}(\omega_{4,5})$
$\boldsymbol{Y}(\omega_{6,7})$
$\boldsymbol{\Pi}_0^0$
$\boldsymbol{\Pi}_1^0$
$\boldsymbol{\Pi}_2^0$
$\boldsymbol{\Pi}_3^0$
$\boldsymbol{Y}(\omega_0)$
$\boldsymbol{Y}(\omega_1)$
$\boldsymbol{Y}(\omega_2)$
$\boldsymbol{Y}(\omega_3)$
$\boldsymbol{Y}(\omega_4)$
$\boldsymbol{Y}(\omega_5)$
$\boldsymbol{Y}(\omega_6)$
$\boldsymbol{Y}(\omega_7)$

Fig. 2. Dyadic permutation sorting scheme for the case when the total number of frequency bins is $K = 8$.

Analysis and Signal Separation (ICA2009). 2009, vol. 5441 of *LNCS*, pp. 282–289, Springer.

[10] Radoslaw Mazur and Alfred Mertins, "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] 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] 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] F. Nesta and M. Omologo, "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] Prasad Sudhakar and Rémi Gribonval, "A sparsity-based method to solve permutation indeterminacy in frequency-domain convolutive blind source separation," in *ICA '09: Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, Berlin, Heidelberg, 2009, pp. 338–345, Springer-Verlag.

[17] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *IEEE International Symposium on Circuits and Systems (ISCAS 2007)*, May 2007, pp. 3247 –3250.

[18] H. Sawada, S. Araki, and S. Makino, "Mlsp 2007 data analysis competition: Frequency-domain blind source separation for convolutive mixtures of speech/audio signals," in *IEEE Workshop on Machine Learning for Signal Processing*, aug. 2007, pp. 45 –50.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

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