On Separation Performance Enhancement in Convolutive Blind Source Separation

Radoslaw Mazur and Alfred Mertins
Institute for Signal Processing
University of Lübeck
Ratzeburger Allee 160
23538 Lübeck, Germany

***Abstract*—In this paper we present a method for estimating the bin-wise separation performance in convolutive blind source separation. A common way to separate convolutive mixtures is the transformation to the time-frequency domain and the separation of the single bins using an instantaneous ICA algorithm. This approach reduces the complexity but leads to the so-called permutation problem which has been widely studied. Another problem arises when the single bins are only poorly separable or even not separable at all. These bins can significantly reduce the overall performance. In this paper we propose a method for detecting such bins based on properties of the unmixing matrices.**

I. Introduction

A typical way for separating convolutive mixtures is to transform the signals to the time-frequency domain and to use a bin-wise instantaneous separation in each frequency. For the instantaneous unmixing problem, many different approaches have been proposed [1], [2], [3].

The two arising difficulties of this approach are the so-called permutation and scaling problems, for which solutions have been proposed in [4], [5], [6], [7], [8]. A rather unaddressed problem, however, is the detection of bins where the separation is low or has even failed. A reason for separation failure can be a singular mixing matrix or the presence of only one source at a given frequency.

In this paper we propose a method for detecting these bins and derive an algorithm for maximizing the separation performance without adding new distortions.

II. Mixing and unmixing model

In the instantaneous case the mixing process of N sources into N observations can be modeled by an $N \times N$ matrix $\mathbf{A}$. Assuming 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 vector of source signals. To obtain the separated signals $\boldsymbol{y}(n) = [y_1(n), \ldots, y_N(n)]^T$ a multiplication with the unmixing matrix $\mathbf{B}$ has to be performed:

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

The unmixing matrix $\mathbf{B}$ is estimated only on the basis of the observed process $\boldsymbol{x}(n)$. The separation is considered successful when $\mathbf{BA} = \mathbf{D\Pi}$ with $\mathbf{\Pi}$ being a permutation matrix and $\mathbf{D}$ being an arbitrary diagonal matrix. The two matrices $\mathbf{\Pi}$ and $\mathbf{D}$ represent the two ambiguities of BSS. The order of the sources cannot be determined and any scaling of the signals yields a valid solution. For the separation we use the well-known gradient-based update rule according to [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 $\boldsymbol{g}(y) = (g_1(y_1), \ldots g_n(y_n))$ is a component-wise vector function of nonlinear score functions

$$g_i(s_i) = -\frac{p_i'(s_i)}{p_i(s_i)} \tag{5}$$

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 [9].

In a realistic acoustic scenario the model has to be extended as the reverberation has to be taken into account. The mixing system can be modeled by FIR filters of length L. Depending on the reverberation time and sampling rate, L can reach several thousands. The convolutive mixing model reads

$$\boldsymbol{x}(n) = \mathbf{H}(n) * \mathbf{s}(n) = \sum_{l=0}^{L-1} \mathbf{H}(l)\boldsymbol{s}(n-l) \tag{6}$$

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 $M \geq L-1$ and obtain

$$\boldsymbol{y}(n) = \mathbf{W}(n) * \boldsymbol{x}(n) = \sum_{l=0}^{M-1} \mathbf{W}(l)\boldsymbol{x}(n-l) \tag{7}$$

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

A common way to solve the convolutive BSS problem is the transformation to the time-frequency domain using the short-time Fourier transform (STFT) where the convolution becomes a multiplication [10]:

$$\boldsymbol{Y}(\omega_k, \tau) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k, \tau), \qquad k = 0, 1, \ldots, K-1, \tag{8}$$

with K being the FFT length. This approach simplifies the problem as in the time-frequency domain the separation task

is reduced to an instantaneous unmixing in each frequency bin. The downside of this approach is the possible permutation and scaling in each frequency bin:

$$\boldsymbol{Y}(\omega_k,\tau) = \boldsymbol{W}(\omega_k)\boldsymbol{X}(\omega_k,\tau) = \boldsymbol{D}(\omega_k)\boldsymbol{\Pi}(\omega_k)\boldsymbol{S}(\omega_k,\tau) \quad (9)$$

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

When the permutation is not corrected before reconstructing the time signals, the entire separation fails. Without this correction different signals will appear in every output at different frequencies. Several approaches for solving the permutation problem have been proposed [4], [5], [6], [7], [11], [12].

For restoring the original sources as closely as possible, the correction of the scaling ambiguity is needed. Otherwise a filtered version is recovered. A method using the inverse postfilter has been proposed in [13]. The idea is to recover the signals as they have been recorded at the microphones. This approach accepts the filtering done by the mixing system without adding new distortions. In [14] a similar technique, known as the minimal distortion principle, has been proposed. This method uses the following unmixing matrix

$$\boldsymbol{W}'(\omega) = \mathrm{dg}(\boldsymbol{W}^{-1}(\omega)) \cdot \boldsymbol{W}(\omega) \quad (10)$$

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

This approach, however, does not take into account the problem of poorly separated bins. If there is strong crosstalk at some frequencies, properly chosen small scaling factors for the affected frequencies could lead to an enhanced separation ratio. A method for estimating such scaling factors will be proposed in the next section.

III. The proposed Algorithm

The proposed algorithm consists of two parts. In the first part, we describe a model for blind estimation of the bin-wise separation performance. The second part consists of the transformation of this information to a set of scaling factors. The algorithm will be described for two sources, but can be extended to more than two.

A. Blind estimation of the bin-wise separation performance

The main idea of the proposed algorithm can be described as follows. If the mixing matrix at a given frequency ω_k is close to being singular, its vectors are usually almost parallel. This means the unmixing vectors are also almost parallel. In this case, small errors in the estimated unmixing vectors may lead to strong crosstalk in this frequency bin.

The ideal case where the unmixing vectors $\boldsymbol{w}_1(\omega_k)$ and $\boldsymbol{w}_2(\omega_k)$ are biorthogonal to the mixing vectors $\boldsymbol{h}_1(\omega_k)$ and $\boldsymbol{h}_2(\omega_k)$ (i.e., the columns of the mixing matrix $\boldsymbol{H}(\omega_k)$) is shown in Fig. 1 for the case where no overall permutation occurs. Because of $\boldsymbol{w}_1(\omega_k) \perp \boldsymbol{h}_2(\omega_k)$ and $\boldsymbol{w}_2(\omega_k) \perp \boldsymbol{h}_1(\omega_k)$ we have for the overall gains $g_{ij}(\omega_k)$ from source j to output i

$$g_{11}(\omega_k) = \boldsymbol{w}_1^H(\omega_k)\boldsymbol{h}_1(\omega_k) \quad (11)$$

$$g_{22}(\omega_k) = \boldsymbol{w}_2^H(\omega_k)\boldsymbol{h}_2(\omega_k) \quad (12)$$

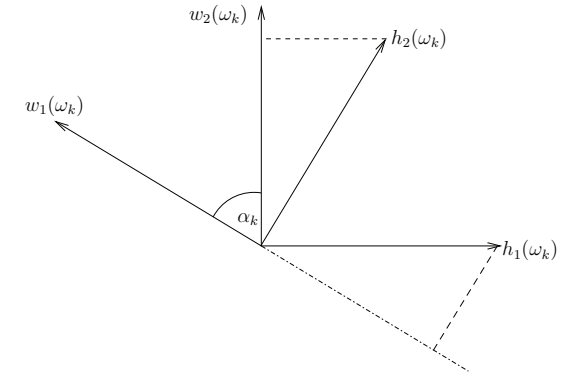

Fig. 1. The projection of the components $\boldsymbol{h}_1(\omega_k)$ and $\boldsymbol{h}_2(\omega_k)$ on $\boldsymbol{w}_1(\omega_k)$ in the ideal case.

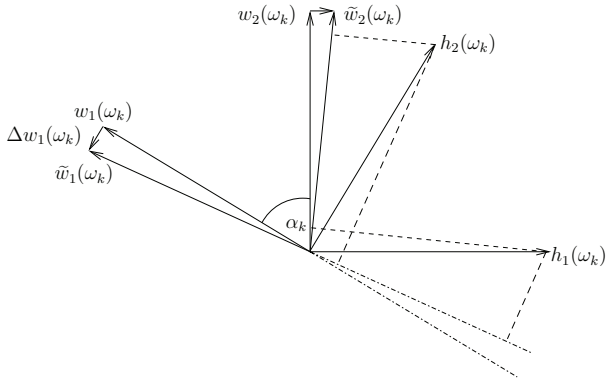

Fig. 2. The projection of the components $\boldsymbol{h}_1(\omega_k)$ and $\boldsymbol{h}_2(\omega_k)$ on $\widetilde{\boldsymbol{w}}_1(\omega_k)$ in a realistic case where the projection vector is not orthogonal to one of the components.

and $g_{12}(\omega_k) = g_{21}(\omega_k) = 0$ regardless of the actual angle α_k. When using the minimal distortion principle, the gain differences are properly corrected between the frequency bins.

In Fig. 2 a more realistic case is shown, in which the biorthogonality condition between the estimated unmixing vectors and the true mixing vectors does not hold perfectly. As the unmixing vectors are estimated via an ICA-Algorithm using a finite data set there are estimation errors $\Delta\boldsymbol{w}_i(\omega_k)$ such that unmixing vectors $\widetilde{\boldsymbol{w}}_i(\omega_k) = \boldsymbol{w}_i(\omega_k) + \Delta\boldsymbol{w}_i(\omega_k)$ are used instead of the ideal ones. The overall gains $g_{ij}(\omega_k) = \widetilde{\boldsymbol{w}}_i^H(\omega_k)\boldsymbol{h}_j(\omega_k)$ now become

$$\begin{aligned} g_{11}(\omega_k) &= \boldsymbol{w}_1^H(\omega_k)\boldsymbol{h_1}(\omega_k) + \Delta\boldsymbol{w}_1^H(\omega_k)\boldsymbol{h_1}(\omega_k) \\ g_{12}(\omega_k) &= \Delta\boldsymbol{w}_1^H(\omega_k)\boldsymbol{h_2}(\omega_k) \\ g_{21}(\omega_k) &= \Delta\boldsymbol{w}_2^H(\omega_k)\boldsymbol{h_1}(\omega_k) \\ g_{22}(\omega_k) &= \boldsymbol{w}_2^H(\omega_k)\boldsymbol{h_2}(\omega_k) + \Delta\boldsymbol{w}_2^H(\omega_k)\boldsymbol{h_2}(\omega_k) \end{aligned} \quad (13)$$

It is clear that because of $g_{12}(\omega_k) \neq 0$ and $g_{21}(\omega_k) \neq 0$ there is crosstalk. The signal-to-interference ratio (SIR) at frequency ω_k for output one can be written as $SIR_1(\omega_k) = |S_1(\omega_k)|^2$ with

$$S_1(\omega_k) = \frac{\boldsymbol{w}_1^H(\omega_k)\boldsymbol{h_1}(\omega_k) + \Delta\boldsymbol{w}_1^H(\omega_k)\boldsymbol{h_1}(\omega_k)}{\Delta\boldsymbol{w}_1^H(\omega_k)\boldsymbol{h_2}(\omega_k)} \quad (14)$$

Assuming $\|\boldsymbol{h_1}(\omega_k)\| = \|\boldsymbol{h_2}(\omega_k)\|$ and $\Delta\boldsymbol{w}_1(\omega_k) \perp \boldsymbol{w}_1(\omega_k)$

with $\|\Delta \boldsymbol{w}_1(\omega_k)\| = \|\boldsymbol{w}_1(\omega_k)\| \sin\gamma_k$ as well as $|\Delta \boldsymbol{w}_1^H(\omega_k)\boldsymbol{h_1}(\omega_k)| \approx \|\Delta \boldsymbol{w}_1(\omega_k)\| \cdot \|\boldsymbol{h_1}(\omega_k)\|$ we obtain

$$|S_1(\omega_k)| \approx \frac{\cos(\frac{\pi}{2} - \alpha_k)}{\sin\gamma_k} + \cos(\alpha_k) \quad (15)$$

with α_k being the angle between the mixing vectors, which is assumed to be the same as the one between the unmixing vectors. Given two complex vectors $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, we compute the angle between them as

$$\alpha = \arccos\left(\frac{|\boldsymbol{h}_1^H \boldsymbol{h}_2|}{\|\boldsymbol{h}_1\| \, \|\boldsymbol{h}_2\|}\right) \quad (16)$$

This formulation maps negative angles to positive ones, and it maps angles α larger than $\pi/2$ to $\pi - \alpha$. However, this is not a problem in our context, as these changes are similar to the ones caused by the ambiguities of the ICA method. Note that, under the assumptions made, the same SIR would be obtained for the second output. γ_k, which is the angle between $\boldsymbol{w}_1(\omega_k)$ and $\widetilde{\boldsymbol{w}}_1(\omega_k)$, is not known, but experiments showed that an assumption of 0.001π for all frequencies is quite realistic.

In Fig. 3 a result of estimating the SIR for every frequency bin is given. For comparison, in Fig. 4 the real SIR is shown. One can observe that there is a high similarity between the estimated and true SIR.

B. Calculation of scaling factors

Using equation (15) an estimated SIR for every frequency bin can be calculated. This information is then used to calculate a set of scaling coefficients $c(\omega)$ that maximize the overall SIR.

The overall SIR is maximal when the crosstalk in every frequency bin is the same. With the estimated energy $E(\omega_k) = \sum_\tau |\boldsymbol{Y}(\omega_k, \tau)|^2$ in every bin the scaling coefficients are calculated as

$$c(\omega_k) = \frac{E(\omega_k) \cdot \tilde{S}(\omega_k)}{\frac{1}{K} \cdot \sum_k E(\omega_k)} \quad (17)$$

with

$$\tilde{S}(\omega_k) = \frac{|S_1(\omega_k)|}{\frac{1}{K}\sum_{k=0}^{K-1} |S_1(\omega_k)|} \quad (18)$$

Coefficients calculated for a real-world example using (17) are shown in Fig. 5(a). Here the lower frequencies are attenuated while the higher ones are emphasized. The reason for this is that at lower frequencies there is poor separation and high energy at the same time. This behavior leads to a high coloration of the signals.

A less invasive method would be to assume an equal energy in all bins and to use $c(\omega_k) = \tilde{S}(\omega_k)$. Compared to the original solution based on the minimal distortion principle, this approach still enhances the overall SIR but the coloration is not as strong as with the coefficients from (17). The coefficients $c(\omega_k) = \tilde{S}(\omega_k)$ are shown in Fig. 5(b).

An even less invasive method is to keep coefficients that are above a certain level untouched, which means that for such bins, the scaling is the same as that calculated by the minimal distortion principle. With $c(\omega_k) = \min\{\tilde{S}(\omega_k), 1\}$ the poorly separated bins are attenuated while no special emphasis is put on the good ones.

TABLE I
COMPARISON OF SEPARATION PERFORMANCE FOR THE DIFFERENT ALGORITHMS

	Left	Right	Overall	SFM
Default	16.8	17.5	17.2	0.61
$c(\omega_k)$ from (17)	29.9	27.2	28.5	0.07
$c(\omega_k) = \tilde{S}(\omega_k)$	20.3	23.8	22.3	0.51
$c(\omega_k) = \min\{\tilde{S}(\omega_k), 1\}$	20.2	23.8	22.2	0.53

IV. SIMULATIONS

The validity of the proposed approach has been tested on real-world data available at [15]. The permutation problem was not addressed as the perfect depermutation is known. For the comparison of the algorithms, the values of the overall SIR and the spectral flatness measure (SFM)

$$SFM = \frac{\sqrt[P]{\prod_{p=0}^{P-1} |G(k)|^2}}{\frac{1}{P}\sum_{p=0}^{P-1} |G(k)|^2} \quad (19)$$

for the overall system $\boldsymbol{G}(n) = \boldsymbol{H}(n) * \boldsymbol{W}(n)$ consisting of the mixing and unmixing system will be used [16], [17]. A higher SIR means better separation and a higher SFM means less linear distortion of the separated signals.

The results of the minimal distortion principle are shown in the first line in Table I. As one can see, the signals have been separated quite well. The distortions of the signals are only the result of the mixing system and are quite small.

The next line shows the results for scaling coefficients calculated using (17). With this method, the overall SIR has been be boosted by over 11dB. This improvement is achieved at the cost of strong coloration, witch is reflected in the poor SFM.

The choice $c(\omega_k) = SIR(\omega_k)$ still leads to an overall enhancement of the SIR by 5dB. The distortions are much less invasive, and informal listening tests showed that the perceived quality is almost as good as with the minimal distortion principle, but with reduced crosstalk.

The last modification with a limit of one for $c(\omega_k)$ is a small one. It leads to negligible decline in SIR and a small improvement in the SFM. The differences are barely audible.

V. SUMMARY

In this paper we presented a way for estimating the bin-wise separation performance in convolutive blind source separation. Based on the estimate, we derived an algorithm for increasing the overall separation ratio. The validity of the approach has been shown on a real-world example.

REFERENCES

[1] S. 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. 1996, vol. 8, pp. 757–763, The MIT Press.

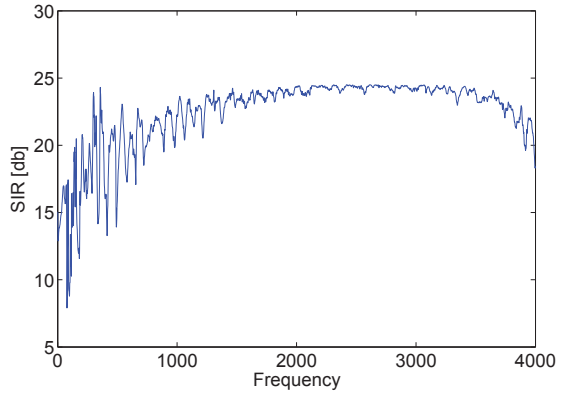

Fig. 3. The estimated separation ratio for single bins.

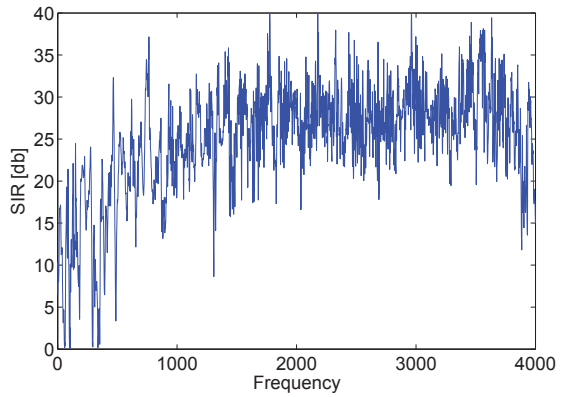

Fig. 4. The separation ratio for single bins.

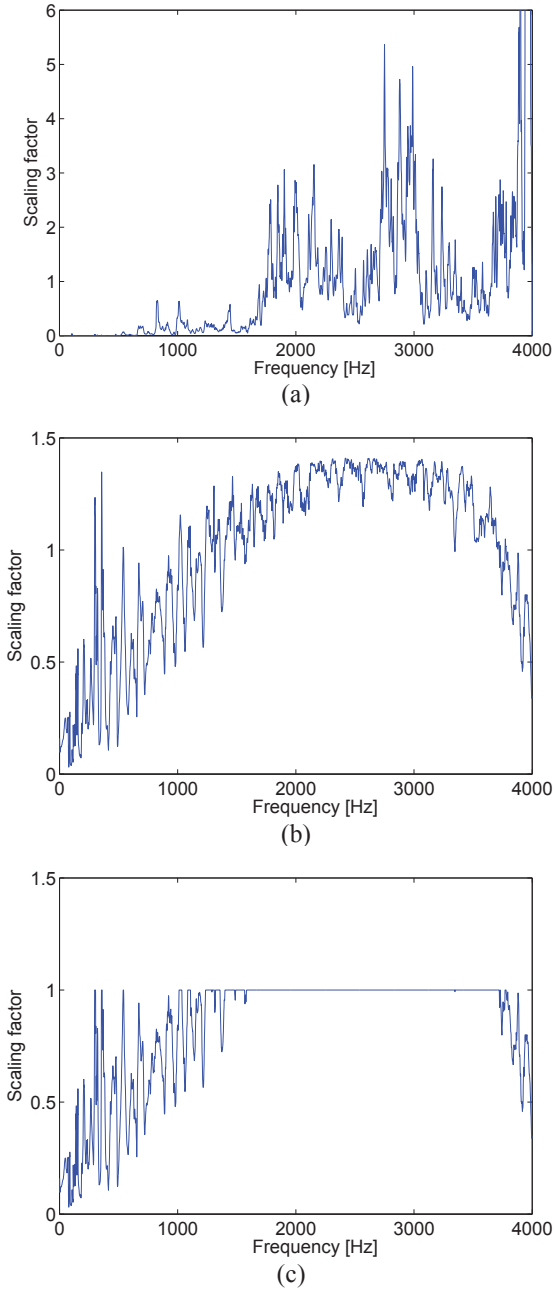

Fig. 5. Three variants of scaling coefficients. (a) $c(\omega_k)$ from (17). (b) $c(\omega_k) = \tilde{S}(\omega_k)$. (c) $c(\omega_k) = \min\{\tilde{S}(\omega_k), 1\}$.

[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] 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.

[5] J. Anemüller 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] R. Mazur and A. Mertins, "Solving the permutation problem in convolutive blind source separation," in *Independent Component Analysis and Signal Separation*, M. E. Davies, Ch. J. James, S. A. Abdallah, and M. D. Plumbley, Eds. 2007, vol. 4666 of *LNCS*, pp. 512–519, Springer.

[12] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation of 3-d located many speech signals," in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2005, pp. 9–12.

[13] 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.

[14] 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.

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

[16] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. Int. Workshop Independent Component Analysis and Blind Signal Separation*, Aussois, France, Jan. 1999.

[17] James D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communication*, vol. 6, no. 2, pp. 314–232, Feb 1988.