MULTI-CHANNEL SOURCE SEPARATION PRESERVING SPATIAL INFORMATION

Robert Aichner [1]*, Herbert Buchner, Meray Zourub, and Walter Kellermann*

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
{aichner, buchner, zourub, wk}@LNT.de

ABSTRACT

In this paper we propose two novel methods for preserving the spatial information in source separation algorithms. Our approach is applicable to any source separation algorithm and is based on an additional supervised adaptive filtering with the reference signals generated by the source separation system. If a special constrained optimization scheme is applied to derive the source separation algorithm then the novel approach can be simplified. The quality of the spatial representation and the separation performance of both methods and two state-of-the-art approaches from the literature have been evaluated by a MUSHRA listening test according to the relevant ITU recommendation showing that the novel methods clearly outperform the state-of-the-art approaches.

Index Terms— Spatial Information, Spatialization, Source Separation, Hearing Aids, Spatial Auditory Displays

1. INTRODUCTION

In several applications as, e.g., hands-free communication interfaces, it is desirable to extract the clean source signals from multiple linear mixtures in complex acoustic environments. The ability of humans to understand speech in such complex scenarios has mainly been attributed to the binaural processing strategy. Thus, especially in applications with multi-channel reproduction systems it is important that the adaptive signal processing algorithms which aim at suppression of interfering sources preserve the spatial information. This allows the human auditory system to further improve the interference suppression by exploiting the spatial information on the remaining sources. Important examples of stereo reproduction are headset-based spatial auditory displays (e.g., for air traffic control) [1], or bilateral hearing aids. In a recent study [2] it has been shown that hearing impaired persons also benefit from binaural cues such as interaural time and level differences. Thus, it is important that the algorithms inside the hearing aid do not change these spatial cues. However, current state-of-the-art bilateral hearing aids have a negative impact on these binaural cues and interference suppression techniques applied independently at each ear can have an additional negative impact on localization performance [2]. Therefore, recently several binaural algorithms have been published which aim at interference suppression while maintaining the spatial cues. So far, mainly adaptive beamforming techniques [3, 4] or multi-channel Wiener filter methods [5] have been investigated.

In this paper, we deal with blind source separation (BSS) methods which are an attractive alternative to beamforming as no a-priori knowledge on source and sensor positions is required. BSS algorithms have the advantage that they are solely based on the fundamental assumption of mutual statistical independence of the different source signals. Thus, the separation is achieved by forcing the output signals to be mutually statistically decoupled up to joint moments of a certain order. Due to the insensitivity with respect to the sensor positions, BSS becomes especially attractive if the wireless link between the bilateral hearing aids, which is currently only used to transmit control information, would allow simultaneous processing of the acoustic signals from both devices. Moreover, in contrast to previous methods [5], no voice-activity detector is needed.

Conventional BSS approaches yield monaural estimates of the separated source signals. However, due to the reasons explained above it is desirable to preserve the spatial information of the acoustic environment. So far, the few publications on BSS algorithms preserve spatial cues either by post-processing of the estimated sources [6, 7] or by introducing a second cost function which constrains the set of possible solutions [8, 9]. In this paper, these existing approaches will be briefly reviewed and two novel approaches which maintain the spatial information will be presented. This will be followed by a formal experimental evaluation of the existing and proposed methods using a subjective listening test according to [10].

2. SUPPRESSING INTERFERERS BUT PRESERVING SPATIAL INFORMATION

2.1. Mixing and Demixing Model

The mixing of the original sources is modeled by finite impulse response (FIR) filters of length M as encountered, e.g., in acoustic environments, leading to the sensor signals

$$x_p(n) = \sum_{q=1}^{Q} \sum_{\kappa=0}^{M-1} h_{qp,\kappa} s_q(n-\kappa) \tag{1}$$

where $h_{qp,\kappa}$, $\kappa = 0, \ldots, M-1$ denote the coefficients of the FIR filter model from the q-th source s_q, $q = 1, \ldots, Q$ to the p-th sensor x_p, $p = 1, \ldots, P$. Conventional source separation algorithms aim at finding a corresponding demixing system, whose output signals $y_q(n)$ are described by

$$y_q(n) = \sum_{p=1}^{P} \sum_{\kappa=0}^{L-1} w_{pq,\kappa} x_p(n-\kappa), \tag{2}$$

where $w_{pq,\kappa}$, $\kappa = 0, \ldots, L-1$ denote the current weights of the multiple-input multiple-output (MIMO) filter taps from the p-th sensor channel $x_p(n)$ to the q-th output channel. The filter taps of the demixing system can be estimated by beamforming or BSS techniques.

[1] Now with Microsoft Corporation, Redmond, WA, USA.

1-4244-0728-1/07/$20.00 ©2007 IEEE

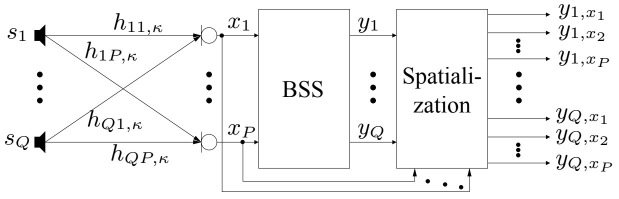

Fig. 1. Block diagram of BSS preserving spatial information.

In (2) and Fig. 1 it can be seen that conventional source separation algorithms estimate only monaural representations y_q of the separated sources. As a consequence the spatial information about the sources obtained by the multiple sensors is lost in y_q. In contrast to this, the preservation of spatial information requires estimation of the contribution of each desired source signal s_q to each sensor signal x_p which is given as $s_q(n) * h_{qp,n}$. Thus, for each source s_q we have to estimate P output signal components $y_{q,x_1}, \ldots, y_{q,x_P}$ leading to a total of PQ output signals (Fig. 1). In most applications it is sufficient to extract only one desired source and thus, only P outputs $y_{q,x_1}, \ldots, y_{q,x_P}$ have to be estimated. The decision as to which of the Q output signals y_q corresponds the desired source could, e.g., be solved by using the spatial information [11]. However, this topic exceeds the scope of this paper. Moreover, in the remainder of this paper we assume that the number of *active source signals* Q is less or equal to the number of microphones P.

2.2. Existing Approaches

So far, only a few approaches in literature are able to at least partially maintain spatial information. They can be classified into post-processing schemes [6, 7] and methods which add a second cost function to constrain the set of possible solutions [8, 9].

2.2.1. Post-Processing Methods

The post-processing approach proposed in [7] is motivated by the relation between blind system identification and blind source separation [12] for a certain class of broadband BSS algorithms (e.g., [13]). For instance, for $P = 2$ and the filter length $L = M$, a demixing system that ensures perfect separation is given by the mixing system with the channel-wise indices reordered [12] yielding:

$$\begin{array}{ll} \mathbf{w}_{11} = \alpha_1 \mathbf{h}_{22}, & \mathbf{w}_{12} = -\alpha_2 \mathbf{h}_{12}, \\ \mathbf{w}_{21} = -\alpha_1 \mathbf{h}_{21}, & \mathbf{w}_{22} = \alpha_2 \mathbf{h}_{11}. \end{array} \tag{3}$$

Hence, the demixing FIR filters $\mathbf{w}_{pq} = [w_{pq,0}, \ldots, w_{pq,L}]^T$ are scaled estimates of the mixing FIR filters $\mathbf{h}_{qp} = [h_{qp,0}, \ldots, h_{qp,M}]^T$ with the scaling factors denoted as α_1, α_2. Spatial information is mainly contained in the time delays and level differences between the different mixing filters $\mathbf{h}_{qp}$. Thus, based on (3) it was suggested in [7] that the spatial cues can be recovered by convolving the monaural BSS output signals y_q with the appropriate demixing filters to generate a spatialized version of the signal. For the case $P = 2$ the spatialized outputs are given as:

$$\begin{array}{ll} y_{1,x_1}(n) = y_1(n) * w_{11,n}, & y_{2,x_1}(n) = y_2(n) * w_{21,n}, \\ y_{1,x_2}(n) = y_1(n) * w_{12,n}, & y_{2,x_2}(n) = y_2(n) * w_{22,n}. \end{array} \tag{4}$$

In [7] also an equation for obtaining y_{q,x_p} for arbitrary P was given. It should be pointed out that by using (4) the separated source in y_q and the remainder of the interfering sources which could not be suppressed completely will be projected to the same spatial position.

Another related post-processing method was described earlier in [6] and is based on the inversion of the demixing system in the discrete Fourier transform (DFT) domain. In [7] both approaches were compared, both theoretically and experimentally, and it was shown that the method based on (4) yields better results. Therefore, post-processing based on (4) is included as a reference algorithm in the experiments in Sect. 3.

2.2.2. Constrained Optimization Methods

In conventional BSS the mutual information between the ouput channels y_q is minimized. A measure of statistical independence which is often used as optimization criterion for non-Gaussian processes is the Kullback-Leibler divergence (KLD) between the estimate of the P-dimensional joint probability density function (pdf) $\hat{p}_{y,P}$ of all channels and the univariate pdfs $\hat{p}_{y,1}$ of the individual channels which is given as

$$\mathcal{J}_{\text{MCBD}}(n) = \hat{\mathrm{E}} \left\{ \log \frac{\hat{p}_{y,P}(\mathbf{y}(n))}{\prod_{q=1}^{P} \hat{p}_{y,1}(y_q(n))} \right\}, \tag{5}$$

where $\mathbf{y}(n) = [y_1(n), \ldots, y_P(n)]^T$, and $\hat{\mathrm{E}}\{\cdot\}$ is the estimate of the statistical expectation. As the temporal dependencies of the source signals (e.g., speech signals) are not modeled in (5) the output signals y_q will become temporally whitened and thus the algorithms derived from $\mathcal{J}_{\text{MCBD}}$ perform multi-channel blind deconvolution (MCBD). To avoid this whitening effect and also to preserve the spatial information, the conventional criterion $\mathcal{J}_{\text{MCBD}}$ originally developed for i.i.d. data signals was complemented in [8] by a second update equation. For binaural signals, i.e., $P = 2$ as addressed in [8], the structure shown in Fig. 2 has been proposed. It can be shown that the update equation proposed in [8] is based on an optimization criterion combining two KLDs leading to

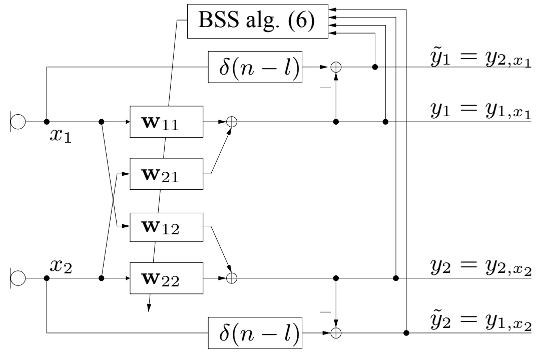

Fig. 2. Block diagram of the constrained optimization method and optimum solution exemplarily shown for $P = 2$.

$$\mathcal{J}(n) = \mathcal{J}_{\text{MCBD}}(n) + \beta\, \hat{\mathrm{E}} \left\{ \log \frac{\hat{p}_{\tilde{y},P}(\tilde{\mathbf{y}}(n))}{\prod_{q=1}^{P} \hat{p}_{\tilde{y},1}(\tilde{y}_q(n))} \right\} \tag{6}$$

with $\tilde{y}_q = x_q(n-l) - y_q(n)$ being the BSS output signal y_q subtracted from the sensor signal x_q which is delayed by l samples, i.e., convolved by $\delta(n-l)$ (see Fig. 2). The combination of the channels is given as $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2]^T$. The parameter β (which is set equal to one in [8]) allows a trade-off between both cost functions. The optimum solution of the simultaneous minimization of both KLDs combined in $\mathcal{J}$ is given as [8]

$$\tilde{y}_1 = x_1(n-l) - y_1(n) \quad = \quad y_{2,x_1}(n) \tag{7}$$

$$\tilde{y}_2 = x_2(n-l) - y_2(n) \quad = \quad y_{1,x_2}(n) \tag{8}$$

$$y_1(n) = y_{1,x_1}(n) \quad (9)$$

$$y_2(n) = y_{2,x_2}(n) \quad (10)$$

and thus yields separated signals with their spatial information preserved. An extension to $P > 2$ can be found in [9]. This is to the authors' knowledge the only approach where a criterion to preserve spatial information is incorporated in the BSS optimization criterion and thus this method is evaluated experimentally in Sect. 3.

2.3. Novel Approaches

As the first of the two novel approaches an adaptive post-processing method will be explained which is applicable to any BSS algorithm. Based on this method we discuss a constrained optimization scheme where the supervised adaptive filters of the previous concept may be omitted by introducing the minimal distortion principle (MDP) [14].

2.3.1. Post-Processing by an Adaptive Multi-Channel Interference Canceller

The structure of this approach is depicted in Fig. 3 for simplicity for the case $P = 2$. The demixing system $\mathbf{w}_{pq}$ yielding the separated sources y_q can be determined by any BSS algorithm, e.g., the algorithm proposed in [13]. For many applications the *extraction of a single source* is sufficient so that the output channel containing the desired source has to be identified. Due to the blind identification property of the demixing system [12] obtained by the BSS algorithm in [13], e.g., the spatial information contained in the demixing filters $\mathbf{w}_{pq}$ could be used for this decision [11]. In the following it is assumed that the channel y_1 contains the monaural representation of the desired source s_1 and all other output channels $y_2, \ldots, y_P$ contain estimates of the interfering sources.

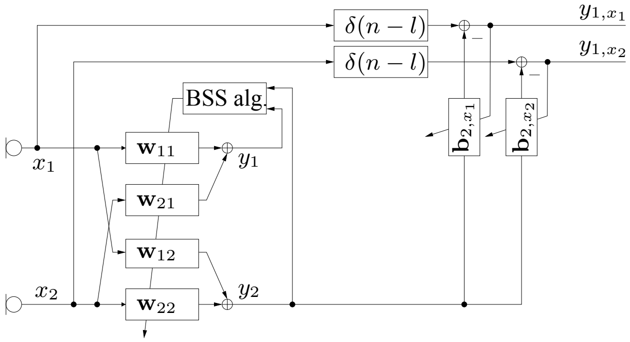

Fig. 3. Preservation of spatial information for $P = 2$ and for y_1 containing the monaural representation of the desired source.

To obtain a spatially correct representation of the desired source and – in contrast to the method in Sect. 2.2.1 – also of the suppressed interfering sources, we propose to apply a set of adaptive filters. As shown in Fig. 3 these supervised adaptive filters use the estimates of the interfering sources y_q, $q = 2, \ldots, P$ as a reference signal to perform interference cancellation. If the desired source s_1 is silent then the adaptation of the filter $\mathbf{b}_{q,x_p}$ $(q = 2, \ldots, P, p = 1, \ldots, P)$ will yield the contribution of the q-th interferer to the p-th sensor signal at the output of $\mathbf{b}_{q,x_p}$. Interference cancellation is achieved by subtracting the outputs of the adaptive filters from the delayed sensor signals x_p. If the desired signal s_1 is active then the adaptation of $\mathbf{b}_{q,x_p}$ has to be stopped. Thus, to obtain reliable estimates of the interferers an adaptation control is needed which adapts the filters $\mathbf{b}_{q,x_p}$ only in the case when the *desired source is inactive*. A reliable decision on desired source activity can be made if the adaptation control is implemented for each DFT bin independently. A sophisticated DFT-based method was presented in [15] and is used in our experiments.

Ideally, the interferers contained in the sensor signals are completely cancelled and thus, the resulting signals $y_{1,x_1}, \ldots, y_{1,x_P}$ contain the spatially correct desired source. In practice, the interfering sources can only be suppressed to a certain degree so that there is usually some residual of the interfering signals. Thus, $y_{1,x_1}, \ldots, y_{1,x_P}$ contain the desired source *and* the remaining suppressed interference as picked up by the multiple sensors x_p and therefore preserve the spatial information of the *whole acoustic environment* allowing further *binaural* processing by the human auditory system. To estimate the adaptive filters $\mathbf{b}_{q,x_p}$, a DFT-domain algorithm based on the following optimization criterion

$$\mathcal{J}_{\mathbf{b}_{q,x_p}}(n) = (1-\lambda) \sum_{i=0}^{n} \lambda^{n-i} \sum_{\nu=0}^{R-1} |y_{q,x_p}^{(\nu)}(i)|^2 \quad (11)$$

can be used. The frequency-domain representation $y_{q,x_p}^{(\nu)}$ of y_{q,x_p} for the ν-th bin of a length-R DFT is obtained from a signal block of the variables x_p, y_q, and the filters $\mathbf{b}_{q,x_p}$, respectively. The forgetting factor λ is chosen to $0 < \lambda < 1$. In our experiments we use an algorithm derived in [16] incorporating robust statistics [17] in the optimization criterion (11). This makes the estimation more robust against outliers.

2.3.2. Constrained Optimization Method based on the MDP

In [14] the minimal distortion principle (MDP) was proposed which constrains the demixing filters $\mathbf{w}_{pq}$ of the BSS algorithm to *avoid distortion* of the separated signals. The MDP is equivalent to constraining the output signals to $y_q \overset{!}{=} y_{q,x_q}$. If a BSS algorithm together with the MDP is applied, then the structure of the post-processing algorithm (Fig. 3) can be simplified. We again assume that the desired source is obtained in y_1. Due to the MDP this channel already represents the contribution of the desired source at x_1, i.e., y_{1,x_1} (see Fig. 4). Thus, the filter $\mathbf{b}_{2,x_1}$ may be omitted. Additionally, the estimate of the interfering soure at output y_2 corresponds to y_{2,x_2}. Hence, the contribution of the desired source at the sensor x_2 can easily be obtained by subtracting the interfering source y_2 from the second sensor x_2 without the need of the adaptive filter $\mathbf{b}_{2,x_2}$ (Fig. 4). The delay $\delta(n-l)$ accounts for a possible delay introduced by the BSS filters $\mathbf{w}_{pq}$. A cost function which achieves separation under the MDP constraint is given as

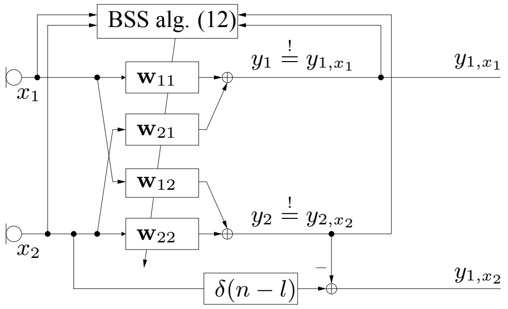

Fig. 4. Constrained optimization method based on the MDP.

$$\mathcal{J}_{\mathrm{MDP}}(n) = \mathcal{J}_{\mathrm{MCBD}}(n) + \gamma\, \hat{\mathrm{E}} \left\{ \|\mathbf{x}(n-l) - \mathbf{y}(n)\|^2 \right\}, \quad (12)$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_P(n)]^T$, $\mathbf{y}(n) = [y_1(n), \ldots, y_P(n)]^T$ and $\|\cdot\|$ denotes the 2-norm.

3. EXPERIMENTS

The algorithms were tested in a living room environment with a reverberation time of $T_{60} = 320\,\mathrm{ms}$. $P = 2$ was chosen and speech signals were used as desired source s_1 and interfering source s_2. As the proposed technique is very attractive for hearing aids, the two microphones were placed at the left and right human ear. In such applications it is often assumed that the desired source comes approximately from the front, i.e., $\theta_1 \approx 0^\circ$. Thus, two scenarios were tested: a) $\theta_1 = 0^\circ$, $\theta_2 = 60^\circ$ and b) $\theta_1 = -30^\circ$, $\theta_2 = 60^\circ$. The experimental results of the two proposed and the two state-of-the-art algorithms have been evaluated using a listening test called MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) according to [10]. The test was performed with ten experienced listeners and the presented stereo test signals were of length 10 s. To evaluate only the *spatial* impression, the stereo hidden reference signal was chosen as the two microphone signals. Thereby the segmental signal-to-interference ratio ($\mathrm{SIR}_{\mathrm{seg}}$) was adjusted to the same suppression of the interfering source as achieved at the monaural output of the BSS algorithm where an improvement of $\Delta\mathrm{SIR}_{\mathrm{seg}} = 12\,\mathrm{dB}$ was achieved. Thus, for the *hidden reference* a separated desired signal is generated with *perfect spatial impression* of both, desired source and residual of the interfering source. The anchor signal is the monaural BSS output of the desired source. The listeners were instructed to rate the test signals with respect to the reference signal. Thus, in general the results are influenced by the spatial impression and the separation performance. However, all algorithms achieved comparable $\Delta\mathrm{SIR}_{\mathrm{seg}}$ without any signal distortions such as musical noise and differed mainly by the ability to preserve spatial information. The experimental results include the initial convergence phase of the algorithms and are shown in Fig. 5 with the indices of the algorithms being given by

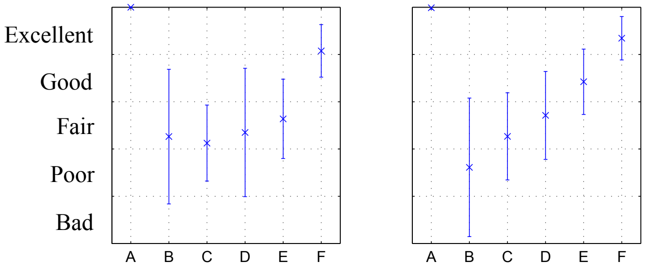

Fig. 5. Results of the MUSHRA test for scenario (a) (left plot) and scenario (b) (right plot).

(A) Hidden reference
(B) Monaural BSS output
(C) Postfiltering by the demixing system (Sect. 2.2.1)
(D) Constrained optimization using (6) (Sect. 2.2.2)
(E) Proposed method based on MDP (Sect. 2.3.2)
(F) Proposed method based on adaptive filtering (Sect. 2.3.1)

The BSS algorithm used in (B), (C), and (F) is explained in detail in [13] and β, γ in (D), (E) are chosen to $\beta = 1, \gamma = 0.1$. In Fig. 5 the average grades and the standard deviations are depicted.

For scenario (a) the monaural BSS output and the algorithms (C)-(E) are rated very similar. The reason is that the desired source is at 0° which corresponds to the perceived position in a monaural representation. As mentioned before (Sect. 2.2.1), for (C) all sources contained in y_1 will be projected to the same position so that there is no difference to the monaural representation. Methods (D) and (E) do not succeed completely in preserving the spatial information of the suppressed interfering source. The residual interferer sounds diffuse. This is in contrast to (F) which succeeds in preserving spatial information of both, desired and interfering source. We attribute this to the efficient decoupling of the optimization criteria.

In scenario (b) the desired source position deviates from 0° so that already the approach (C) improves the spatial impression even if the residual interferer is perceived to be also located at -30°. Method (D) obtains an average rating as the interferer sounds again diffuse and due to the complicated cost function (6) the convergence of the algorithm is slow. The proposed approach (E) shows improved performance, even if the interfering source is not clearly localizable due to its diffuse character. Again (F) shows excellent performance as all sources can be well localized. Also note that in both scenarios the listeners were always able to detect the hidden reference.

4. CONCLUSIONS

Two novel methods for maintaining the spatial information in source separation algorithms have been presented. The experimental evaluation by a MUSHRA listening test showed the improved performance compared to state-of-the-art algorithms.

5. ACKNOWLEDGEMENTS

We thank Dr.-Ing. Jürgen Herre from the Fraunhofer IIS for providing the software tool to perform the MUSHRA listening test.

6. REFERENCES

[1] W.T. Nelson et al., “Spatial audio displays for speech communications,” in *Proc. of the Human Factors and Ergonomics Society*, 1999, pp. 1202–1205.

[2] T. van den Bogaert et al., “Horizontal localization with bilateral hearing aids: Without is better than with,” *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 515–526, January 2006.

[3] T. Lotter and P. Vary, “Dual-channel speech enhancement by superdirective beamforming,” *EURASIP Journal on Applied Signal Processing*, pp. 1–14, 2006.

[4] T. Hoya et al., “Stereophonic noise reduction using a combined sliding subspace projection and adaptive signal enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, pp. 309–230, May 2005.

[5] T.J. Klasen et al., “Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2005, vol. 3, pp. 29–32.

[6] S. Ikeda and N. Murata, “An approach of blind source separation of speech signals,” in *Proc.8th International Conference on Artificial Neural Networks*, Skovde, Sweden, September 1998, pp. 761–767.

[7] S. Wehr et al., “Post-processing for BSS algorithms to recover spatial cues,” in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, Sept. 2006.

[8] T. Takatani et al., “Evaluation of SIMO separation methods for blind decomposition of binaural mixed signals,” in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, Sept. 2005, pp. 233–236.

[9] T. Takatani et al., “High-fidelity blind separation of acoustic signals using SIMO-model-based ICA with information-geometric learning,” in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, Sept. 2003, pp. 251–254.

[10] ITU-R, “Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems,” January 2003.

[11] H. Buchner et al., “Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering,” in *Proc. ICASSP*, 2005, vol. 3, pp. 97–100.

[12] H. Buchner, R. Aichner, and W. Kellermann, “Relation between blind system identification and convolutive blind source separation,” in *Proc. Joint Workshop on Hands-Free Communication and Microphone Arrays*, March 2005.

[13] R. Aichner, H. Buchner, and W. Kellermann, “A novel normalization and regularization scheme for broadband convolutive blind source separation,” in *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, March 2006, pp. 527–535.

[14] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source separation,” in *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, Dec. 2001, pp. 722–727.

[15] W. Herbordt, T. Trini, and W. Kellermann, “Robust spatial estimation of the signal-to-interference ratio for non-stationary mixtures,” in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2003, pp. 247–250.

[16] W. Herbordt et al., “Application of a double-talk resilient DFT-domain adaptive filter for bin-wise stepsize controls to adaptive beamforming,” in *Proc. Int. Workshop on Nonlinear Signal and Image Processing*, May 2005.

[17] P.J. Huber, *Robust Statistics*, Wiley, New York, 1981.