MULTIDIMENSIONAL LOCALIZATION OF MULTIPLE SOUND SOURCES USING AVERAGED DIRECTIVITY PATTERNS OF BLIND SOURCE SEPARATION SYSTEMS

Anthony Lombard [1], *Tobias Rosenkranz* [1]*, *Herbert Buchner* [2], *Walter Kellermann* [1]

[1] Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg, Cauerstr. 7, 91058 Erlangen, Germany,
{lombard,wk}@LNT.de, tobias.rosenkranz@siemens.com

[2] Deutsche Telekom Laboratories, Technical University Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany,
hb@buchner-net.com

ABSTRACT

In this paper, we propose a versatile acoustic source localization framework exploiting the self-steering capability of Blind Source Separation (BSS) algorithms. We provide a way to produce an acoustical map of the scene by computing the averaged directivity pattern of BSS demixing systems. Since BSS explicitly accounts for multiple sources in its signal propagation model, several simultaneously active sound sources can be located using this method. Moreover, the framework is suitable to any microphone array geometry, which allows application for multiple dimensions, in the near field as well as in the far field. Experiments demonstrate the efficiency of the proposed scheme in a reverberant environment for the localization of speech sources.

Index Terms— Acoustic source localization, blind source separation, microphone arrays

1. INTRODUCTION

Acoustic source localization aims at estimating the position of one or several sound sources by exploiting the spatial diversity offered by an array of microphones. It can serve in many applications as a preliminary step to other processes like, e.g., steering a beamformer or pointing a camera in the direction of a sound source. The localization procedure adopted in this paper is based on Blind Source Separation (BSS). Fig. 1 shows the general BSS setup. Because of the reverberation in the acoustical environment, Q source signals s_q ($q = 1 \dots Q$) are filtered by a Multiple-Input-Multiple-Output (MIMO) mixing system $\mathbf{H}$ modeled by M-tap Finite Impulse Response (FIR) filters h_{qp} between the q-th source and the p-th sensor. P signal mixtures x_p ($p = 1 \dots P$) are picked up by a microphone array, together with some background or sensor noise n_p. The source signals s_q are assumed to be mutually independent (which in general holds for speech and audio signals). To separate the source signals s_q without access to the acoustical mixing system $\mathbf{H}$, BSS algorithms aim at output signals y_q that are statistically independent by suitably adapting the weights of the BSS demixing system $\mathbf{W}$, which captures the L-tap FIR separating filters w_{pq} between the p-th sensor and the q-th output. In anechoic environments, BSS can be considered as a set of blind adaptive null-beamformers. Although it does not entirely hold under realistic conditions, this interpretation gives some useful insights into the self-steering capability of BSS techniques [1]. Actually, while accurate source location information is usually necessary to steer a beamformer, BSS offers the possibility to recover the original source signals from a (possibly reverberant) sound mixture without this prior knowledge. Intuitively, this self-steering capability should therefore imply that the BSS demixing filters contain some useful information on the location of each source.

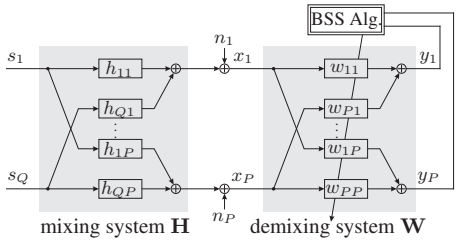

Fig. 1. BSS signal model.

Several methods have been proposed to extract the location information from the BSS filters. Most of them were originally developed to solve a permutation problem specific to narrowband BSS. They extract the location information in each frequency bin separately (see, e.g., [2, 3]). But since they rely on the phase information of the BSS demixing filters, they suffer from spatial aliasing occurring at high frequencies for large microphone spacings. This reduces the allowable size of the microphone array and consequently the spatial resolution of the localization scheme. A completely different approach has been described in [4], where BSS was used to perform blind system identification of the acoustical mixing system and extract Time-Differences-Of-Arrival (TDOAs). Operating directly in the time domain, this method does not suffer from spatial aliasing and explicitely accounts for multi-path sound propagation, contrary to the previous methods. It is easily applicable for $Q = 2$ sources but becomes more difficult for more sources. A simplified approximate solution for $Q > 2$ was therefore proposed in [1], where we exploited the directivity patterns of the BSS outputs, similar to the methods presented in [2]. But we applied an averaging procedure to combine every frequency bin and BSS output. The resulting BSS Averaged Directivity Pattern (BSS-ADP) allowed us to treat the general case of two or more sources, thereby gathering useful localization information even from frequency regions corrupted by spatial aliasing. Localization of up to six sources in a noisy and reverberant environment has been demonstrated.

However, the BSS-ADP method has only been formulated in [1] for linearly-aligned sensors and for the estimation of Directions-Of-Arrival (DOAs), assuming sources located far away from the sen-

* Now with Siemens Audiologische Technik, Erlangen

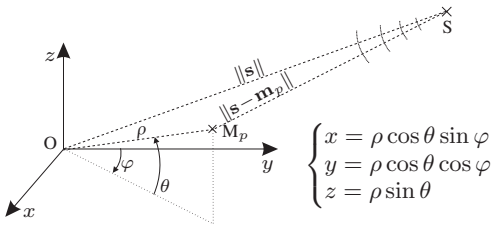

Fig. 2. Geometrical illustration.

sors. In this paper, we present a generalization of the BSS-ADP method to treat arbitrary sensor array geometries. The general formulation of the BSS-ADP is first described in Sect. 2. It allows to localize sound sources in the near field as well as in the far field, thereby considering the latter case as a special case of the first one, as discussed in Sect. 3. Experimental results are presented in Sect. 4 and concluding remarks are provided in Sect. 5.

2. LOCALIZATION USING THE BSS-ADP

2.1. The BSS directivity patterns

A directivity pattern is usually defined as the magnitude squared response of a Multiple-Input-Single-Output (MISO) system of filters (typically a beamformer) to monochromatic plane waves coming from all possible directions (see, e.g., [2]). This definition is therefore valid for sources located far away from the sensors. But it can be generalized to arbitrary source positions by considering spherical waves originating from isotropic sound sources. Considering an array of P microphones M_p, $p = 1 \dots P$, and the MIMO system of $P \times P$ BSS demixing filters depicted in Fig. 1, we define the directivity pattern of the q-th BSS output as follows:

$$B_{\mathbf{W}_{:q}}(\mathbf{s}, f) = \left| \sum_{p=1}^{P} w_{pq}(f) e^{-j2\pi f \frac{\Delta d(\mathbf{s}, \mathbf{m}_p)}{c}} \right|^2, \quad (1)$$

$$\Delta d(\mathbf{s}, \mathbf{m}_p) = \|\mathbf{s} - \mathbf{m}_p\| - \|\mathbf{s}\|, \quad (2)$$

where c denotes the velocity of sound, $\mathbf{s} = [x_\mathrm{s}, y_\mathrm{s}, z_\mathrm{s}]^T$ is a vector capturing the cartesian coordinates of the considered emitting location S, and $\mathbf{W}_{:q}$ captures the filters $w_{1q}, \dots, w_{Pq}$ contributing to the q-th BSS output. $\mathbf{m}_p = [x_{\mathrm{m}_p}, y_{\mathrm{m}_p}, z_{\mathrm{m}_p}]^T$ pertains to the position of the p-th microphone M_p, and $\|\cdot\|$ denotes the Euclidean norm of a vector. By convention, we assumed here the phase to be zero at the origin O of the coordinate system. $\Delta d(\mathbf{s}, \mathbf{m}_p)$ corresponds to a difference of radii and $\Delta d(\mathbf{s}, \mathbf{m}_p)/c$ can be interpreted as the TDOA of a spherical wave propagating from O to M_p and originating from S (see Fig. 2).

The definition (1) of the BSS directivity patterns ignores the presence of reflection paths. When the number Q of sources is equal to the number P of sensors, the ideal separation solution under the free-field (i.e., anechoic) assumption consists of a set of P null-beamformers (see [1]), each placing $P-1$ perfect spatial nulls (i.e., with infinite attenuation) in the direction of $P-1$ competing sources. The BSS directivity patterns reflect therefore the behavior of a BSS algorithm in an anechoic scenario. However, under realistic conditions, considering BSS as a set of blind beamformers is somehow misleading since, contrary to a beamformer, the ideal BSS solution still allows *perfect* interference rejection, even with multi-path propagation for each source. This corresponds to a joint diagonalization of the overall system $\mathbf{C}(f) = \mathbf{H}(f)\mathbf{W}(f)$ from the sources to the BSS outputs, in all frequencies and for an arbitrary $\mathbf{H}(f)$ [1]. The directivity patterns therefore do not truly reflect the behavior of a BSS algorithm under realistic conditions. But as long as the direct propagation paths are sufficiently strong compared to the reflection paths, they can be very useful for source localization since only the direct propagation paths carry some meaningful location information.

2.2. Computation of the averaged directivity pattern

Instead of considering each frequency bin and each output separately like in [2], we apply an averaging procedure before extracting the source locations. It consists of summing the BSS directivity patterns over the frequencies, and over the $P-1$ "best" BSS outputs, i.e., discarding for each frequency point and each (discrete) look-up location S, the output with maximum array response:

$$q^*(\mathbf{s}, f) = \arg\max_q B_{\mathbf{W}_{:q}}(\mathbf{s}, f), \quad (3)$$

$$\bar{B}_{\mathbf{W}}(\mathbf{s}) = \int_{f_{\min}}^{f_{\max}} \sum_{\substack{q=1 \\ q \neq q^*(\mathbf{s},f)}}^{P} B_{\mathbf{W}_{:q}}(\mathbf{s}, f) \; df. \quad (4)$$

Approximating a BSS system as a set of null-beamformers cancelling $P-1$ sources in each output (see the discussion in Sect. 2.1), we can expect $\bar{B}_{\mathbf{W}}(\mathbf{s})$ to show local minima pointing at the source positions. In practice, the integral is replaced by a summation over a finite number of frequency points. The boundaries $f_{\min}$ and $f_{\max}$ are intended to reduce the impact of low resolution at very low frequencies for small microphone apertures, and the effect of spatial aliasing at high frequencies for large microphone spacings. Restricting the analysis bandwith is also an easy way to reduce the computational complexity of the scheme. But note that $f_{\max}$ does not need to be chosen small enough to completely avoid spatial aliasing because only the "true" spatial nulls (as opposed to the unwanted grating lobes) add up coherently when summing over the BSS outputs and frequencies, which automatically attenuates the impact of spatial aliasing at high frequencies. This averaging procedure therefore allows to gather useful localization information from a large range of frequencies, including (at least some of) the higher frequency regions, even with large microphone spacings.

Additionally, another very efficient way to attenuate unwanted side lobes caused by spatial aliasing or reflections, is to apply a non-linear transformation $g(\cdot)$:

$$\tilde{B}_{\mathbf{W}}(\mathbf{s}) = g\left(\frac{\bar{B}_{\mathbf{W}}(\mathbf{s}) - \min_{\mathbf{s}} \bar{B}_{\mathbf{W}}(\mathbf{s})}{\max_{\mathbf{s}} \left\{ \bar{B}_{\mathbf{W}}(\mathbf{s}) - \min_{\mathbf{s}} \bar{B}_{\mathbf{W}}(\mathbf{s}) \right\}} \right). \quad (5)$$

As can be seen from (5), we first normalize $\bar{B}_{\mathbf{W}}(\mathbf{s})$ before applying $g(\cdot)$, so that it is spread between 0 and 1. $g(\cdot)$ should be a monotonically increasing function with monotonically decreasing derivative in the interval $[0\ 1]$. It should also satisfy the condition $g(0) = 0$ and quickly converge towards 1 to ensure $g(1) \approx 1$. In this paper, we choose $g(x) = \tanh(4x)$, following the suggestion made in [5], where this function was heuristically selected (but omitting the normalization step described above) and proved to be effective for similar peak enhancement purposes.

3. NEAR-FIELD AND FAR-FIELD MODELS

We saw in Sect. 2.2 that the BSS-ADP can be used for acoustic source localization. When evaluated for all possible source locations S, $\tilde{B}_{\mathbf{W}}(\mathbf{s})$ provides an acoustical map of the scene with local minima indicating the source positions. To obtain such a map, we need to calculate the distance $\Delta d(\mathbf{s}, \mathbf{m}_p)$ defined in (2), for all possible positions S and microphones M_p, $p = 1 \dots P$. In practice, we consider only a finite set of look-up positions.

$$\Delta d(\mathbf{s},\mathbf{m}_p) \quad = \quad \sqrt{(x_\mathrm{s}-x_{\mathrm{m}_p})^2+(y_\mathrm{s}-y_{\mathrm{m}_p})^2+(z_\mathrm{s}-z_{\mathrm{m}_p})^2} - \sqrt{{x_\mathrm{s}}^2+{y_\mathrm{s}}^2+{z_\mathrm{s}}^2} \tag{6}$$

$$= \quad \rho_\mathrm{s}\sqrt{1-2\frac{\rho_{\mathrm{m}_p}}{\rho_\mathrm{s}}\left[\cos\theta_\mathrm{s}\cos\theta_{\mathrm{m}_p}\cos(\varphi_\mathrm{s}-\varphi_{\mathrm{m}_p})+\sin\theta_\mathrm{s}\sin\theta_{\mathrm{m}_p}\right]+\left(\frac{\rho_{\mathrm{m}_p}}{\rho_\mathrm{s}}\right)^2} - \rho_\mathrm{s} \tag{7}$$

$$\underset{|\rho_{\mathrm{m}_p}/\rho_\mathrm{s}|\to 0}{\approx} \quad -\rho_{\mathrm{m}_p}\left[\cos\theta_\mathrm{s}\cos\theta_{\mathrm{m}_p}\cos(\varphi_\mathrm{s}-\varphi_{\mathrm{m}_p})+\sin\theta_\mathrm{s}\sin\theta_{\mathrm{m}_p}\right] \tag{8}$$

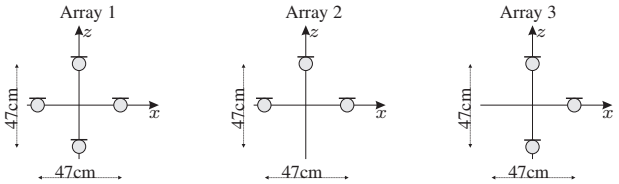

Fig. 3. Microphone array geometries.

3.1. The near-field propagation model

The distance $\Delta d(\mathbf{s},\mathbf{m}_p)$ can be easily computed using the cartesian coordinate system. The resulting expression (6) is an exact distance calculation, conforming with the near-field propagation model of spherical waves (although we consider here the dispersion-free case). The near-field model requires computation of the BSS-ADP for all possible positions within the three-dimensional space and may therefore be computationally very demanding. In many applications, it might be sufficient to work under the far-field assumption and localize the sources by their DOA only.

3.2. The far-field assumption

The computation of $\Delta d(\mathbf{s},\mathbf{m}_p)$ may be greatly simplified by considering plane waves for sources far away from the sensors. For arrays of linearly-aligned sensors, this approximation allows to compute $\Delta d(\mathbf{s},\mathbf{m}_p)$ very easily, using basic notions of trigonometry. But non-linear array geometries are not as straightforward to handle. To treat the general case, we first need to reformulate (6) using the spherical coordinates (ρ,φ,θ) depicted in Fig. 2. We then obtain (7). If we now choose the origin O of the coordinate system close to the sensor array, we may apply the far-field approximation $|\rho_\mathrm{s}| \gg |\rho_{\mathrm{m}_p}|, \forall p = 1\ldots P$ to neglect the quadratic term under the square root and use a Taylor series expansion up to the first order for the root. This leads to the far-field approximation (8) of $\Delta d(\mathbf{s},\mathbf{m}_p) = \Delta d(\varphi_\mathrm{s},\theta_\mathrm{s},\mathbf{m}_p)$, which turns out to be independent of the range coordinate ρ_s, as expected. Note that the special case of linearly-aligned microphones placed along the z-axis (Fig. 2) can be easily obtained by inserting $\theta_{\mathrm{m}_p} = \frac{\pi}{2}, \forall p = 1\ldots P$ into (8). In this case, the distance $\Delta d(\mathbf{s},\mathbf{m}_p) = \Delta d(\theta_\mathrm{s},\mathbf{m}_p) = -\rho_{\mathrm{m}_p}\sin\theta_\mathrm{s}$ becomes also independent of φ_s and exhibits therefore a cylindrical symmetry around the array axis, as expected.

4. EXPERIMENTAL EVALUATIONS

4.1. The experimental setup

To evaluate the localization performance of the presented scheme, speech signals of duration 10 seconds each were played by a loudspeaker at different positions in a living-room-like environment ($T_{60} \approx 300$ms) and recorded using the cross-shaped array 1 depicted in Fig. 3. The triangular configurations 2 and 3 in Fig. 3 were simply obtained by omitting one of the microphones used for the first configuration. The distance between the sources and the center of the sensor array was about 1.5 meters. Microphone signal mixtures were then generated by summing up the contributions coming from different sources placed at different locations. Demixing filters of length 128 samples were obtained from a broadband second-order-statistics BSS realization of the TRINICON framework [6], operating at the sampling rate 16kHz. Since the considered array geometries were all planar and vertical, the sources were localized by their DOA, under the far-field assumption (Sect. 3.2). The directivity patterns (1) of the demixing filters were computed for all possible azimuth and elevation angles φ_s and θ_s using (8), before applying the averaging procedure described in Sect. 2.2.

The localization results are depicted in the following by displaying BSS-ADP maps, i.e., the value of $\tilde{B}_\mathbf{W}(\theta_\mathrm{s},\varphi_\mathrm{s})$ in all considered directions. Black shades denote low BSS-ADP levels, the circles show the results of the BSS-ADP search and the squares show the reference source positions. References were obtained in each scenario by geometrical measurements made during the recordings.

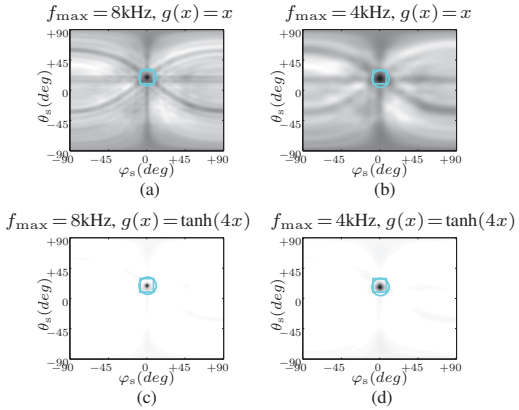

Fig. 4. BSS-ADP maps for one source using the cross-shaped array 1, with different limits $f_{\max}$ and transformations $g(\cdot)$.

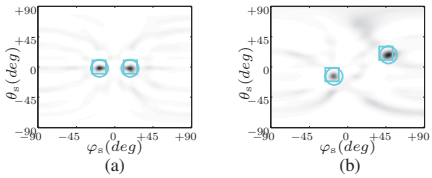

Fig. 5. BSS-ADP maps for two sources using the cross-shaped array 1, for different source positions.

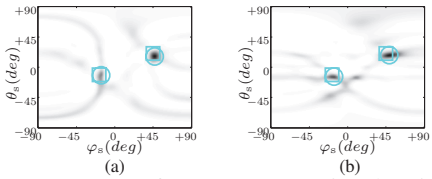

Fig. 6. BSS-ADP maps for two sources using the triangular arrays 2 (left plot) and 3 (right plot).

4.2. Results

The effects of the boundary $f_{\max}$ and of the non-linear mapping $g(\cdot)$ (see Sect. 2.2) are depicted in Fig. 4. The lower boundary $f_{\min}$ in (4) was set to 0 in all cases, like in the rest of the paper. Since we considered here large microphone spacings (47cm), the presence of side lobes due to spatial aliasing is well visible in Fig. 4a. The impact of spatial aliasing can be slightly attenuated by choosing $f_{\max} =$ 4kHz (although 4kHz is still much too high to avoid spatial aliasing completely), as can be seen from Fig. 4b. But the effect of the non-linear mapping $g(\cdot)$ is much more significant, as shown by Fig. 4c and Fig. 4d. Note however that the dominant spatial null could be clearly identified at the expected position in all cases.

Finally, Fig. 5 and Fig. 6 show the BSS-ADP maps obtained for two active speech sources, for the 4-sensor cross-shaped array and for the two triangular geometries, respectively. We chose $f_{\max} =$ 4kHz and $g(x) = \tanh(4x)$ like in Fig. 4d. The source positions were chosen identically for Fig. 5b and Fig. 6 to allow a direct comparison. Two clear peaks appear when using the 4-sensor array. The maps obtained with the 3-sensor array show slightly more side lobes, which makes the localization more ambiguous. This is because reducing the number of sensors reduces the averaging effect involved in the BSS-ADP calculation (4). Nevertheless, the correct source locations could be found in all considered scenarios.

We assumed here that the number of sources was known a priori. It determined the number of peaks to be detected but it did not influence the BSS-ADP computation. Note however that the number of sources may be estimated by counting the number of significant local minima in the acoustical map.

5. CONCLUDING REMARKS

We provided a versatile framework for acoustic source localization using Blind Source Separation. The method can be applied to localize one or several simultaneously active sound sources under reverberant conditions, possibly in multiple dimensions and using an arbitrary microphone array geometry. In [7], we proposed another method to solve a similar problem of localizing two simultaneously active sound sources in two dimensions. The localization procedure adopted there was based on the estimation of TDOAs, where relative temporal signal delays (i.e., the TDOAs) had to be first estimated, before calculating the source positions in a second step. By estimating one TDOA for each source and for each dimension, this two-step procedure can be directly applied for the localization of one source in several dimensions or for the localization of several sources in one dimension. But the generalization to the simultaneous localization of multiple sources in several dimensions is not straightforward since an intermediate correlation-based step is necessary to associate each TDOA to the correct source. This pairing problem arises when each dimension is treated separately, which is not the case with the method proposed in this paper since we perform BSS only once, using the entire (possibly multi-dimensional) sensor array.

The BSS-ADP search provides an acoustical map where multiple local minima pointing at the source locations arise. Other algorithms exist which compute similar maps for localization purposes. Among them, the most popular method is the Steered-Response Power with PHAse Transform (SRP-PHAT), where peaks arise due to the constructive summation of direct propagation paths impinging on multiple microphone pairs [8, 9]. In contrast with the multiple-step procedure [4, 7] based on TDOA estimation, the BSS-ADP and SRP-PHAT approaches therefore belong to the class of direct methods. They use however different criteria since the BSS-ADP method relies on BSS techniques like [4, 7]. Moreover, the BSS-ADP method explicitly accounts for the presence of multiple sources, while SRP-PHAT may encounter some problems in multiple-source scenarios [10]. Fig. 7 depicts acoustical maps obtained with the BSS-ADP and the SRP-PHAT methods in the scenario considered in Fig. 5b. The normalization and non-linear peak enhancement (5) with $g(x) = \tanh(4x)$ was used for the BSS-ADP, and no bandwith reduction was applied (i.e., $f_{\max} = 8$kHz) to assure a fair comparison between both methods. Since the SRP-PHAT produces positive peaks, the non-linear function had to be slightly modified, applying $g(x) = \tanh(4(1-x))$ on the normalized SRP-PHAT map. As can be seen from the picture, in this two-source scenario, the criterion used by the BSS-ADP method produced a sharp peak for each source, contrary to the SRP-PHAT method, which produced only a weak peak for the second source.

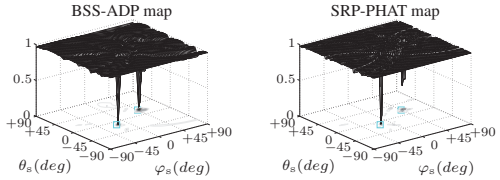

Fig. 7. Acoustical maps obtained by the BSS-ADP and SRP-PHAT methods using array 1 in the presence of two sources.

6. REFERENCES

[1] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Exploiting the self-steering capability of blind source separation to localize two or more sound sources in adverse environments," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, 2008.

[2] M. Ikram and D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2002, vol. 1, pp. 881–884.

[3] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Multiple source localization using independent component analysis," in *Proc. APSURI*, 2005.

[4] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON-based blind system identification with application to multiple-source localization and separation," in *Blind Speech Separation*, S. Makino T.-W. Lee and S. Sawada, Eds. Springer-Verlag, Berlin, 2007.

[5] F. Nesta, M. Omologo, and P. Svaizer, "A novel robust solution to the permutation problem based on joint multiple tdoa estimation," in *Int. Workshop for Acoustic Echo and Noise Control (IWAENC)*, Seattle, Washington, USA, 2008.

[6] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy environments," *Signal Processing*, vol. 86, no. 6, pp. 1260–1277, Jun. 2006.

[7] A. Lombard, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using blind adaptive MIMO system identification," in *Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Heidelberg, Germany, 2006.

[8] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 288–292, 1997.

[9] J. Dmochowski, J. Benesty, and S. Affes, "Fast steered response power source localization using inverse mapping of relative delays," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008.

[10] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two-step acoustic map analysis," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008.