Improved Wideband Blind Adaptive System Identification Using Decorrelation Filters for the Localization of Multiple Speakers

Anthony Lombard, Herbert Buchner, and Walter Kellermann

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

***Abstract*— This paper addresses the TDOA extraction problem for localizing multiple sources in noisy and reverberant environments with emphasis on speech excitation. TDOAs are estimated by performing blind adaptive MIMO system identification using a gradient-based BSS variant of the TRINICON framework. We present a novel method to improve the TDOA estimation for signals with lowpass-like spectral characteristics such as speech, for which the standard approach achieves only imperfect system identification. To this end, we propose to combine the BSS algorithm with decorrelation filters, thereby achieving a greatly improved wideband identification of the acoustical system. The approach is verified in a number of scenarios, where it provides more accurate TDOA estimates for the speaker localization at a negligible additional computational cost.**

I. Introduction

Acoustic Source Localization can be achieved with a two-step procedure consisting in, first, estimating relative temporal signal delays between one or several pairs of microphones and, in a second step, using the estimated delays (the Time Differences of Arrival (TDOA)) to calculate the position of each source. A novel TDOA estimation approach was proposed in [1] which accounts for multipath propagation in reverberant acoustical environments and covers scenarios with multiple sources. Using a gradient-based coefficient optimization variant of the TRINICON framework [2] exploiting Second-Order-Statistics (SOS), this method was originally developed as a Blind Source Separation (BSS) technique for convolutive mixtures [3]. It was then shown in [1] that this approach can also be used to perform the TDOA extraction via blind adaptive Multiple-Input-Multiple-Output (MIMO) system identification of the acoustical system. The method has already successfully been used in complex scenarios as described, e.g., in [4] for the multidimensional localization of multiple sound sources, or in [5], where it was combined with a particle filter for source tracking purposes.

BSS algorithms belong to the class of unsupervised adaptive algorithms as opposed to the supervised algorithms like the simple Normalized Least-Mean-Square (NLMS) algorithm [6], where a reference signal is available to drive the coefficient optimization. In the supervised case, it is well known that gradient-based methods suffer from convergence problems when applied to correlated signals like speech. Because of the lowpass-like spectral envelope of speech, and since the TDOA extraction [1] relies on the algorithm's ability to identify the acoustical system, it is of interest to evaluate how the spectral envelope of the excitation influences the (unsupervised) gradient-based BSS algorithm's TDOA extraction performance.

After a brief review on the BSS-based blind adaptive MIMO system identification and TDOA extraction in Sect. II, we will define in Sect. III a system identification performance measure adapted to the BSS context. In Sect. IV the simulation environment will be presented, and results showing how the excitation's spectral support influences the algorithm will be discussed in Sect. V. Finally, focusing on speech excitation signals, we will see in Sect. VI how suitable decorrelation filters can be advantageously applied to improve the robustness of the multiple speaker localization scheme. In this study, we particularly consider the illustrative case of fixed decorrelation filters.

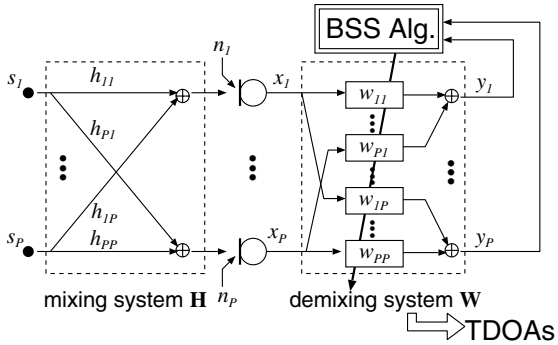

Fig. 1. MIMO model for the TDOA extraction using BSS.

II. TDOA estimation using blind adaptive MIMO system identification

Fig. 1 shows the general setup for the BSS-based TDOA extraction [1]. Because of the reverberation in the acoustical environment, the source signals $s_q(\kappa), q = 1, \ldots, P$, are filtered by a MIMO mixing system $\mathbf{H}$ modeled by Finite Impulse Response (FIR) filters $h_{qp}(\kappa)$ between the q-th source and the p-th sensor. The signal mixture is then picked up by the sensors $x_p(\kappa), p = 1, \ldots, P$, together with some background or sensor noise $n_p(\kappa)$. As the figure indicates, we assume in this study that the number of active sources is less or equal to the number of microphone signals (i.e., the P sources in the figure might or might not all be simultaneously active). Furthermore, the sources are assumed to be mutually independent, which in general holds for speech and audio signals.

To separate the source signals $s_q(\kappa)$ without access to the acoustical mixing system $\mathbf{H}$, the BSS algorithm forces the output signals $y_q(\kappa)$ to be statistically decoupled by suitably adapting the weights of the BSS demixing system $\mathbf{W}$, which refers to the FIR demixing filters $w_{pq}(\kappa)$ between the p-th sensor and the q-th output. In the presence of broadband excitations, the BSS algorithm ideally converges to a solution where the overall system $\mathbf{C} = \mathbf{HW}$ with FIR filters $c_{pq}(\kappa)$ between the p-th source and the q-th output, is diagonal up to an unknown but uncritical permutation of the BSS outputs [3] (the uncritical permutation problem will be ignored in the rest of the paper). For $P = 2$, this solution is obtained when the two following conditions are fulfilled [1]:

$$h_{11}(\kappa) * w_{12}(\kappa) \quad = \quad -h_{12}(\kappa) * w_{22}(\kappa), \tag{1}$$

$$h_{21}(\kappa) * w_{11}(\kappa) \quad = \quad -h_{22}(\kappa) * w_{21}(\kappa), \tag{2}$$

("$*$" denotes convolution) such that the BSS algorithm performs a blind adaptive MIMO system identification of the acoustical mixing system $\mathbf{H}$. The TDOAs can then be calculated after each coefficient update by identifying the direct propagation path between the two sources and the two microphones [1]:

$$\hat{\tau}_1 \quad = \quad \arg\max_{\kappa} |w_{12}(\kappa)| - \arg\max_{\kappa} |w_{22}(\kappa)|, \tag{3}$$

$$\hat{\tau}_2 \quad = \quad \arg\max_{\kappa} |w_{11}(\kappa)| - \arg\max_{\kappa} |w_{21}(\kappa)|. \tag{4}$$

Note that an exact estimation of the mixing system $\mathbf{H}$ is not necessary for all filter taps to perform a successful TDOA estimation since we only need to identify the tap corresponding to the direct propagation path in each filter, as can be seen from (3) and (4). Moreover, to improve the spatial resolution of the localizer at a low computational cost, fractional delays can be obtained by performing an interpolation of the demixing filters in (3) and (4) before performing the TDOA estimation, without increasing the sampling rate for the BSS operations.

III. Performance measures

Traditionally, BSS algorithms are assessed by measuring the gain in Signal-to-Interference Ratio (SIR) obtained after applying the adaptive BSS demixing system $\mathbf{W}$. Following the notations introduced in Sect. II, but omitting the sample index κ like in the rest of the paper, the SIR gain is evaluated at each BSS output y_q as follows:

$$SIR_{\text{out}_q} = 10\log_{10} \frac{\| s_q * c_{qq} \|^2}{\| \sum_{\substack{p=1 \\ p\neq q}}^{P} s_p * c_{pq} \|^2}, \tag{5}$$

$$SIR_{\text{in}_q} = 10\log_{10} \frac{\| s_q * h_{qq} \|^2}{\| \sum_{\substack{p=1 \\ p\neq q}}^{P} s_p * h_{pq} \|^2}, \tag{6}$$

$$SIR_{\text{gain}_q} = SIR_{\text{out}_q} - SIR_{\text{in}_q}. \tag{7}$$

However, the performance of the BSS-based TDOA extraction does not depend on the SIR gain but on the ability of the algorithm to correctly identify the acoustical system, as described in Sect. II. For TDOA estimation purposes, a useful measure to express the effect of BSS is therefore the gain in System-Error-Norm (SEN) evaluated at each BSS output y_q as follows:

$$SEN_{\text{out}_q} = -10\log_{10} \frac{\| c_{qq} \|^2}{\| \sum_{\substack{p=1 \\ p\neq q}}^{P} c_{pq} \|^2}, \tag{8}$$

$$SEN_{\text{in}_q} = -10\log_{10} \frac{\| h_{qq} \|^2}{\| \sum_{\substack{p=1 \\ p\neq q}}^{P} h_{pq} \|^2}, \tag{9}$$

$$SEN_{\text{gain}_q} = SEN_{\text{in}_q} - SEN_{\text{out}_q}. \tag{10}$$

Comparing (5)-(7) and (8)-(10), we see that the SIR only differs from the SEN by a minus sign and by the fact that it includes the source signals in the performance measure. The SIR can actually be seen as a frequency-weighted version of the SEN, assigning weight to frequency regions according to the spectral support of the source signals. The SIR therefore measures how well the acoustical system could be identified at frequencies of importance for the source separation task. It is also of interest to note the analogy existing between, SIR (5)-(7) and SEN (8)-(10) defined for the BSS problem on the one hand, and the well-known Echo Return Loss Enhancement (ERLE) and the coefficient misalignment used for Acoustic Echo Cancellation (AEC) in the context of supervised adaptive system identification problems on the other hand. The misalignment in the supervised case, also sometimes called SEN, should not be confused with the SEN defined in (8)-(10) which is valid in the context of BSS and does not directly provide a measurement of the misalignment between desired and estimated systems but evaluates how far the overall system $\mathbf{C}$ deviates from the desired BSS solutions, i.e., any diagonal overall systems.

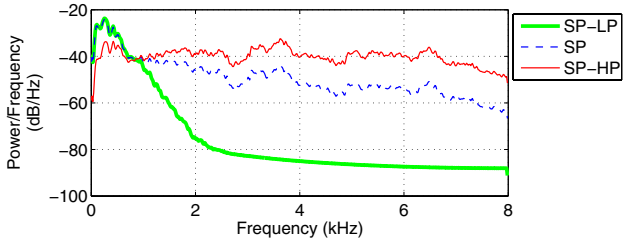

Fig. 2. Power spectral density of the first source signal estimated via the Welch method (window size 40ms, overlap 20ms, total observation time 21s).

IV. Simulation environment

For the simulations considered in the rest of the paper, excitation signals s_q of duration 21s were used. Room impulse responses were measured in three different environments labeled R1, R2 and R3 in the following, with reverberation times T_{60} approximately equal to 250ms, 400ms and 1000ms, respectively, and sensor spacing 21cm. The microphone signals x_q were obtained by convolving the source signals s_q with the measured impulse responses and by adding some noise n_q when required. The TDOA extraction was performed at a sampling rate of 16kHz for $P = 2$ sources and fractional time delay estimates were obtained using an interpolation factor of three to realize a spatio-temporal resolution corresponding to 48kHz sampling rate (Sect. II).

A total of 33 scenarios were considered, with varying source signals, SNR and room reverberation. In each scenario, 77 trials were performed to simulate different source positions. The results presented in the next sections refer to the performance obtained with the BSS algorithm averaged in each scenario over the 77 trials and over the two BSS outputs, in terms of SEN gain, absolute TDOA estimation error and success rate (successful TDOA extraction assumed for an absolute estimation error smaller than $1/3$ sample).

V. Influence of the excitation's spectral support on the BSS localization performance

To assess the influence of the excitations' spectral support on the algorithm, simulations were performed with three types of excitation, each with a different spectral support. Since the SOS-based BSS algorithm [1] needs non-stationary signals, we started from two speech signals denoted in the rest of the paper as signals $s_{\text{SP},1}$ and $s_{\text{SP},2}$. Signals $s_{\text{SP-LP},1}$ and $s_{\text{SP-LP},2}$ were then generated by applying a fifth-order lowpass Butterworth filter f_{LP} (cut-off frequency 1kHz) on the two speech signals, while signals $s_{\text{SP-HP},1}$ and $s_{\text{SP-HP},2}$ were obtained by filtering $s_{\text{SP},1}$ and $s_{\text{SP},2}$ using a first-order highpass FIR filter $f_{\text{HP}}(z) = (1 - 0.95z^{-1})/1.95$. Fig. 2 shows the power spectral density of the three types of excitation (here exemplarily for the first source). We see that signals $s_{\text{SP-LP},q}$ have a pronounced lowpass-like spectral characteristic whereas signals $s_{\text{SP-HP},q}$ present a much flatter spectrum. Though clearly lowpass-like, the original speech signals $s_{\text{SP},q}$ offer a compromise between the two previous excitation types.

Tab. I shows the steady-state performance (averaged over the last six seconds of the simulations) obtained with the BSS algorithm in the noiseless case as well as in the noisy case with additive spatio-spectrally white noise at 5dB and 15dB SNR. To obtain

TABLE I

STEADY-STATE EVALUATION OF THE MULTIPLE SOURCE LOCALIZATION SCHEME FOR EXCITATION SIGNALS WITH VARYING SPECTRAL SUPPORTS

	Microphone input SNR (White Noise)	SEN_{gain} in dB			Absolute TDOA estimation error in samples			TDOA estimation success rate in %		
		R1	R2	R3	R1	R2	R3	R1	R2	R3
SP-LP	$+\infty$	**0.5**	0.4	0.2	0.522	0.719	0.885	42.07	18.23	22.31
SP		**3.0**	2.0	0.8	0.103	0.120	0.136	100.00	97.30	94.48
SP-HP		**14.8**	9.6	4.6	0.106	0.096	0.095	100.00	99.70	100.00
SP-LP	15dB	0.4	0.4	0.2	0.522	0.736	1.020	40.65	20.63	24.58
SP		1.9	1.4	0.7	0.107	0.125	0.191	99.93	96.37	93.58
SP-HP		10.6	7.0	3.8	0.102	0.096	0.091	100.00	99.73	100.00
SP-LP	5dB	0.2	0.2	0.1	0.653	0.880	**1.694**	39.16	21.47	**19.79**
SP		0.7	0.6	0.3	0.114	0.144	**0.624**	99.42	91.39	**81.18**
SP-HP		4.3	2.5	1.8	0.101	0.286	**0.153**	100.00	97.63	**99.27**

TABLE II

STEADY-STATE EVALUATION OF THE MULTIPLE SPEAKER LOCALIZATION SCHEME WITH (SP,f_{HP}) OR WITHOUT (SP) DECORRELATION FILTERS

	Noise type and microphone input SNR		SEN_{gain} in dB			Absolute TDOA estimation error in samples			TDOA estimation success rate in %		
			R1	R2	R3	R1	R2	R3	R1	R2	R3
SP	$+\infty$		**3.0**	2.0	0.8	0.103	0.120	**0.136**	100.00	97.30	**94.48**
SP,f_{HP}			**14.8**	9.6	4.6	0.106	0.096	**0.095**	100.00	99.70	**100.00**
SP	White	15dB	1.9	1.4	0.7	0.107	0.125	0.191	99.93	96.37	93.58
SP,f_{HP}			4.3	2.6	1.7	0.101	0.326	0.302	100.00	96.67	97.53
SP		5dB	0.7	0.6	0.3	0.114	0.144	0.624	99.42	91.39	81.18
SP,f_{HP}			0.5	0.3	0.3	2.254	3.136	3.219	78.00	65.20	65.31
SP	Ventilation	15dB	2.7	2.5	1.0	0.257	0.204	0.691	97.76	97.45	92.27
SP,f_{HP}			14.1	9.2	4.5	0.106	0.096	0.095	100.00	99.71	100.00
SP		5dB	**0.8**	0.6	0.5	2.419	2.121	**1.375**	76.75	76.47	**86.69**
SP,f_{HP}			**12.0**	7.7	3.9	0.104	0.099	**0.093**	100.00	99.91	**100.00**

a comparable convergence time for each type of excitation and allow a fair comparison, the BSS adaptation step-size was chosen differently for each type. A preliminary study actually showed that source signals with a flatter spectrum require a larger step-size than signals with, e.g., a lowpass-like spectrum because the adaptation energy is distributed among a broader range of frequencies. As expected, Tab. I clearly shows that the performance offered by the BSS algorithm depends highly on the nature of the excitation signal under consideration. In particular, the lowpass-like signals $s_{\text{SP-LP},q}$ suffer from a lack of spectral diversity compared with, e.g., the more broadband signals $s_{\text{SP-HP},q}$. This results also in very poor (because concentrated to the low-frequency region) system identification of the acoustical system (the SEN gain is much worst in all scenarios) and a very low TDOA estimation success rate. Results obtained with the speech signals $s_{\text{SP},q}$ also show much lower SEN gain values than those observed for signals $s_{\text{SP-HP},q}$. Yet the impact on the localization performance is limited for rooms R1 and R2, as can be seen from the low absolute TDOA errors and the high success rates obtained with signals $s_{\text{SP},q}$. Tab. I shows therefore that the localization of speakers using the BSS algorithm is feasible and accurate in scenarios with low to moderate reverberation and noise level. In more adverse environments however (room R3 or lower SNR), we notice a significant improvement of the localization accuracy when the system is excited with the highpass-filtered signals $s_{\text{SP-HP},q}$. This is due to the spectral flattening introduced by the highpass filter f_{HP} which allows the BSS algorithm to correctly identify the acoustical system over the entire frequency range.

VI. USING BSS WITH DECORRELATION FILTERS FOR ROBUST SPEAKER LOCALIZATION

For speaker localization purposes, one way to overcome the problems introduced by the lowpass-like spectral characteristics of speech is to partially pre-whiten the microphone signals using a decorrelation filter on each sensor signal before applying the BSS demixing system $\mathbf{W}$. This should result in forcing the BSS algorithm to perform a wideband system identification of the acoustical system $\mathbf{H}$. Since speech signals have a lowpass-like spectral envelope, the decorrelation filters should be highpass filters.

Because the BSS algorithm described in Sect. II performs its coefficient update only based on the output signals y_q, as illustrated in Fig. 1, and because in some applications the BSS outputs y_q are transmitted together with the estimated TDOAs to a far-end user or are used for further processing (see, e.g., [4]), a better solution consists in moving the (linear) decorrelation filters to the BSS outputs, as depicted in Fig. 3 where the decorrelation filters $f_{\text{D},q}$ are not in the signal path of z_q anymore. This way, the signals y_q can serve as spectrally flattened signals for the BSS adaptation while the signals z_q can be directly transmitted to a far-end user. This approach is similar to the technique used to improve the performance of the NLMS algorithm in the supervised adaptive filtering case (see, e.g., [7] for AEC applications) but here neither inverse decorrelation filters nor inverse modeling by the adaptive filters are required.

The results presented in Sect. V indicate that the first-order highpass filter f_{HP} used to generate the signals $s_{\text{SP-HP},q}$ can effectively serve as a fixed linear decorrelation filter, common to each BSS output. Comparing the signals y_q transmitted to the algorithm depicted in Fig. 1 when the highpass-filtered excitation signals $s_{\text{SP-HP},q} = s_{\text{SP},q} * f_{\mathrm{HP}}$ are applied

$$\begin{aligned} y_q &= \sum_{i=1}^{P}\left(\sum_{j=1}^{P} s_{\text{SP-HP},j} * h_{ji} + n_i\right) * w_{iq} \\ &= f_{\mathrm{HP}} * \left(\sum_{i=1}^{P}\sum_{j=1}^{P} s_{\text{SP},j} * h_{ji} * w_{iq}\right) + \sum_{i=1}^{P} n_i * w_{iq}, \quad (11) \end{aligned}$$

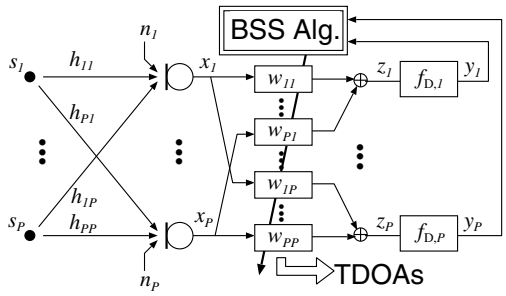

Fig. 3. MIMO model for the TDOA extraction using BSS in combination with decorrelation filters.

and those transmitted to the algorithm depicted in Fig. 3 for the regular speech excitation signals $s_{\mathrm{SP},q}$ and $f_{\mathrm{D},q}=f_{\mathrm{HP}}, q=1,\ldots,P$,

$$y_q \;=\; f_{\mathrm{HP}} * \left(\sum_{i=1}^{P} \sum_{j=1}^{P} s_{\mathrm{SP},j} * h_{ji} * w_{iq} + \sum_{i=1}^{P} n_i * w_{iq} \right), \quad (12)$$

we see that both cases are actually equivalent in the noiseless case ($n_i = 0$). In this case, and using simply f_{HP} as a decorrelation filter, we can therefore expect the same convincing results as those obtained in Sect. V for signals $s_{\mathrm{SP\text{-}HP},q}$. However, in noisy environments, applying the decorrelation filters in (12) acts on both speech and noise terms, which is not the case in (11). Since speech has most of its energy in the low frequency regions, it follows for a noise with some energy at high frequencies (e.g., a white noise) that the (highpass) decorrelation filters can deteriorate drastically the SNR in the signals y_q transmitted to the BSS algorithm, which is however not the case for a noise with a lowpass-like spectrum.

The above analysis is confirmed by the experimental results presented in Tab. II which shows the steady-state performance obtained for the speech excitation signals $s_{\mathrm{SP},q}$ with the BSS algorithm described in Sect. II (labeled SP in the table) and with the BSS algorithm combined with decorrelation filters f_{HP} as described in Fig. 3 (labeled SP,f_{HP} in the table) for the case of $P = 2$ sources. With spatio-spectrally white noise, the decorrelation filters cause a substantial deterioration of the SEN gain and introduce an unacceptable increase in TDOA estimation error, confirming that the approach was unadapted in this case. However in the noiseless case as well as in the presence of a ventilation noise with lowÃijass-like spectrum, the gain offered by the simple use of a fixed (highpass) decorrelation filter was striking, allowing a remarkable TDOA estimation success rate approaching or reaching 100% even in highly reverberant environments and with a ventilation noise at 5dB SNR.

All the results presented until now concerned the steady-state behavior of the algorithm. To provide a better idea on the positive impact of the decorrelation filters, Fig. 4 shows the evolution over time of the algorithms' performance with the 5dB SNR ventilation noise. The results have been averaged over the three rooms R1, R2 and R3 and over all positions and BSS outputs. The SIR gain is also depicted to show that the improved system identification obtained using the decorrelation filters also helped improving the separation performance of the BSS algorithm in such adverse environments.

VII. Conclusion

We studied the influence of the excitation's spectral support on a previously presented TRINICON-based BSS algorithm capable of providing TDOA estimates for several sources via blind adaptive MIMO system identification of the acoustical system. Focusing on

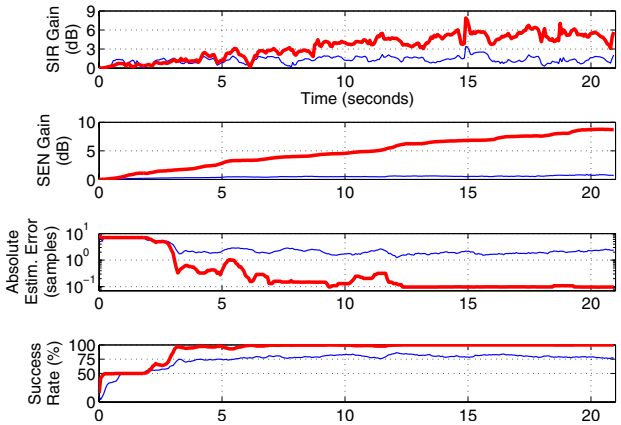

Fig. 4. Performance of the BSS-based multiple speaker localization scheme with (thick lines) or without (thin lines) decorrelation filters in a 5dB SNR ventilation noise environment.

the speaker localization task, experimental results showed that the algorithm's performance was affected by an imperfect system identification, due to the lowpass-like spectral characteristics of speech. The observed performance degradation is limited in moderately reverberant and noisy environments but is significant in more adverse scenarios. Conclusions drawn from this experimental study lead us to consider a new approach combing the BSS algorithm with decorrelation filters to spectrally flatten the BSS output signals. The case of a fixed first-order highpass decorrelation filter at each output was then studied. Further experimental evaluations showed that the approach highly improves the robustness of the multiple speaker localization scheme in a clean speech scenario or in the presence a lowpass-like background noise. Since most natural noise types show some lowpass-like spectral enveopes (e.g., ventilation or traffic noises), the approach can be advantageously applied in many realistic scenarios, without requiring extra computations other than the decorrelation filtering.

References

[1] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann. Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering. In *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Philadelphia, PA, USA, Mar. 2005.

[2] H. Buchner, R. Aichner, and W. Kellermann. TRINICON: A versatile framework for multichannel blind signal processing. In *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 3, pages 889–892, Montreal, Canada, May 2004.

[3] H. Buchner, R. Aichner, and W. Kellermann. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Trans. Speech Audio Processing*, 13(1):120–134, Jan. 2005.

[4] 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, Sep. 2006.

[5] F. Antonacci, D. Riva, D. Saiu, A. Sarti, M. Tagliasacchi, and S. Tubaro. Tracking multiple acoustic sources using particle filtering. In *Eur. Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sep. 2006.

[6] S. Haykin. *Adaptive Filter Theory*. Prentice Hall Information and System sciences series. Prentice Hall, Inc., Upper Saddle River, New Jersey, 4th edition, 2002.

[7] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp. Acoustic echo control - an application of very-high-order adaptive filters. *IEEE Signal Processing Mag.*, 16(4):42–69, July 1999.