

BSS for Improved Interference Estimation for Blind Speech Signal Extraction with Two Microphones

Yuanhang Zheng, Klaus Reindl, and Walter Kellermann

Multimedia Communications and Signal Processing

University of Erlangen-Nuremberg Cauerstr. 7, 91058 Erlangen, Germany

{zheng,reindl,wk}@LNT.de

Abstract—Blind Source Extraction (BSE) as desirable for acoustic cocktail party scenarios requires estimates for the target or interfering signals. Conventional single-channel approaches for obtaining the interference estimate rely on noise and interference estimates during absence of the target signal. For multichannel approaches using multiple microphone signals, a separation of simultaneously active target and interference signals becomes possible if the positions of the target and interfering sources are known. We propose a new method which exploits Directional BSS (Blind Source Separation with a geometric constraint) to estimate the interfering speech sources and diffuse background noise jointly and blindly. Herewith we can effectively deal with the underdetermined BSS scenario (more point sources than sensors) in reverberant environments and can even allow for additional babble noise in the background.

I. INTRODUCTION

Unlike Blind Source Separation (BSS), BSE aims at extracting only one target source from a mixture of signals. This task is widely encountered in hands-free communications and speech-driven human-machine interfaces. For this purpose, estimates for the target and/or interfering signals are required. Single-channel speech enhancement methods require target signal pauses to establish estimates for interference and noise which are difficult to detect and of limited value for nonstationary sources in the considered multi-speaker scenario. In multichannel systems spatial diversity is exploited for spatial filtering so that concurrent sources can still be separated. Beamforming is based on known sensor geometry and requires prior knowledge on the position of the desired source. Some adaptive interference cancelers which do not require exact knowledge of the desired source position, such as robust GSC structures [1], [2] are still not able to account for the reverberation of the target signal when determining interference estimates. As an alternative to beamforming, BSS promises to drastically reduce the number of sensors and to be independent of the array geometry [3], [4], [5]. However, the limitation of [3], [4] is that it completely relies on the determined BSS case where the number of the point sources cannot be larger than the number of the microphones. In [5], the proposed ICA-based noise estimation scheme is limited to a non-point-source noise condition.

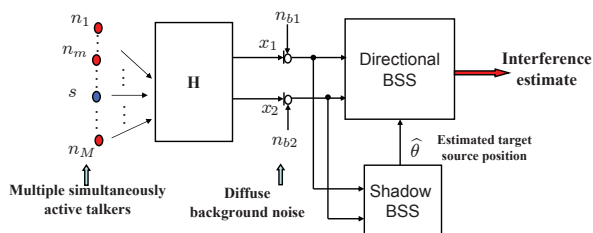


Fig. 1. Proposed scheme for interference estimation in a multi-speaker noisy environment

The proposed scheme shown in Fig. 1 covers determined and underdetermined cases. Here, Directional BSS [3] is exploited to act as a Blocking Matrix (BM) and to produce an estimate for interfering speech sources and diffuse background noise jointly and blindly while only using two microphones. In the following, the proposed algorithm will be described in Sec. II and experimental results will be presented in Sec. III.

II. DIRECTIONAL BSS

A. Basic BSS Model and Optimization Criteria

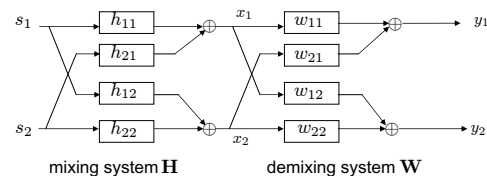


Fig. 2. Basic two-channel linear BSS signal model

Fig. 2 depicts the basic two-channel BSS signal model. Lower-case characters represent time-domain signals, whereas the frequency-domain representations are denoted by underlined lower-case characters. Matrices denoting Multiple-Input-Multiple-Output (MIMO) systems are represented by upper-case boldface characters (time domain) and underlined upper-case boldface characters (frequency domain), respectively. As illustrated in Fig. 2, two point sources s_1, s_2 are filtered by a linear MIMO system \mathbf{H} before they are picked up by two microphones. Thus, the microphone signals can be described in the discrete-time domain by:

$$x_p(k) = \sum_{m=1}^2 h_{mp}(k) * s_m(k), \quad (1)$$

where $*$ represents convolution, h_{mp} , $m, p \in \{1, 2\}$ denotes the FIR filter model from the m -th point source to the p -th microphone. For simplicity, the time index k given in (1) is omitted in the rest of this paper.

BSS algorithms aim at determining a demixing system \mathbf{W} to extract the individual sources from the mixed signals. The output signals of the demixing system y_i , $i \in \{1, 2\}$ are described by:

$$y_i = w_{1i} * x_1 + w_{2i} * x_2, \quad (2)$$

where w_{pi} denotes the demixing filter from the p -th microphone to the i -th output channel.

There are different criteria for convolutive source separation proposed, which are all based on the assumption that sources are statistically independent. In our proposed scheme, the second-order statistics-based version of the TRINICON concept [6] is used for

BSS, where the mutual information between the output channels y_i is minimized. The cost function used to determine a demixing system \mathbf{W} is given by:

$$J_{\text{BSS}}(\mathbf{W}) = \hat{\mathbf{E}} \left\{ \log \frac{\hat{p}_{y,P}(\mathbf{y})}{\prod_{i=1}^P \hat{p}_{y_i}(y_i)} \right\} \quad (3)$$

where $\hat{\mathbf{E}}\{\cdot\}$ is the estimate of the statistical expectation. $\hat{p}_{y,P}$ denotes the estimate of the P -dimensional joint probability density function (pdf) of all channels and \hat{p}_{y_i} is the univariate pdf of the individual channel. Minimizing $J_{\text{BSS}}(\mathbf{W})$ indicates minimizing the Kullback-Leibler divergence (KLD) between $\hat{p}_{y,P}(\mathbf{y})$ and $\prod_{i=1}^P \hat{p}_{y_i}(y_i)$ which leads to maximization of the statistical independence of the outputs y_i .

For the determined 2×2 case, in each output channel one source can be suppressed by an exact spatial null. Nevertheless, for an underdetermined scenario when there are more sources than microphones, no determined solution can be achieved.

B. General underdetermined model and algorithm of Directional BSS

An underdetermined scenario for a two-microphone setup is already illustrated in Fig. 1. One target source s and M interfering point sources n_m are passed through the mixing matrix \mathbf{H} . Besides, the diffuse background noise signals n_{b_1} and n_{b_2} are also picked up by the microphones. Thus, the microphone signals are given by:

$$x_p = x_{p,s} + x_{p,n} = h_{0p} * s + \sum_{m=1}^M h_{mp} * n_m + n_{b_p}, \quad p \in \{1, 2\}, \quad (4)$$

where $x_{p,s}$ and $x_{p,n}$ represent the target component and interfering components contained in the microphone p .

As mentioned before, there is no determined solution for a demixing matrix to separate the individual sources. However, our aim in BSE is not to obtain a determined solution to separate all sources, but to produce an interference estimate by suppressing the target source. This can be done by using Directional BSS, where essentially a spatial null is forced to a certain direction to assure that the source arriving from this direction can be suppressed well. However, as BSS tries to produce independent outputs, correlated components arriving from other directions, i.e., reflections and reverberation will also be suppressed to the greatest extent possible.

The original frequency-domain algorithm for BSS with a geometric constraint was proposed in [3], where BSS was regarded as a set of beamformers whose response is constrained to a set of angles $\theta = [\theta_1, \dots, \theta_M]$ for recovering all M sources from the mixture. The main difference of our proposed algorithm for a two-microphone setup to the original algorithm [3] is that we apply a different geometric constraint. In [3] only the source m located at θ_m should be extracted, whereas all other sources should be suppressed. In contrast to [3], we aim at *suppressing* the target source located at θ and preserving all other sources.

The cost function for the geometric constraint in the frequency domain [3] can be expressed by:

$$J_{\mathbf{C}}(\mathbf{W}) = \|\mathbf{W}\mathbf{D}(\theta) - \mathbf{C}\|^2, \quad (5)$$

where $\mathbf{D}(\theta)$ is a steering vector pointing to the direction θ and \mathbf{C} is the geometric constraint (here \mathbf{C} is a zero vector which indicates that a spatial null will be forced in the direction of θ). Underlined characters denote frequency-domain representations. Based on the the cost function for the generic time-domain BSS algorithm (3), we propose the following filter update in the time domain:

$$\Delta \mathbf{W} = \frac{\partial J_{\text{BSS}}(\mathbf{W})}{\partial \mathbf{W}} + \eta_{\mathbf{C}} \text{DF}T^{-1} \left\{ \frac{\partial J_{\mathbf{C}}(\mathbf{W})}{\partial \mathbf{W}^*} \right\}, \quad (6)$$

where $\text{DF}T^{-1}\{\cdot\}$ denotes the inverse discrete Fourier transform yielding a nonzero update contribution of the same length as the demixing filter length N , and where the superscript $\{\cdot\}^*$ denotes complex conjugation. The weight $\eta_{\mathbf{C}}$, typically in the range $0.4 < \eta_{\mathbf{C}} < 0.6$, indicates the importance of the Directional BSS constraint.

Since in our proposed scheme only one output channel of BSS should be controlled by the constraint, we choose channel 1 to be controlled without loss of generality. Thus, the output signal y_1 of the proposed scheme can be approximated by:

$$y_1 = w_{11} * x_1 + w_{21} * x_2 \approx \sum_{m=1}^M \hat{n}_m + \sum_{p=1}^2 \hat{n}_{b_p}, \quad (7)$$

which is the joint estimate for the interfering point sources and diffuse background noise (\hat{n}_m denotes an estimate of n_m , same for \hat{n}_{b_p}). It should be noted that in the originally proposed Directional BSS [3], a frequency-domain BSS algorithm was used for merging the geometric beamforming. Thus, not only the TRINICON approach [6], but also other BSS algorithms could be associated with the geometric constraint.

C. Estimating the target angular position

In the previous section, the angular position θ of the target source is assumed to be known a priori. But in practice this information is unknown. In order to obtain the Direction of Arrival (DOA) information of the target source, a method of 'peak' detection is used to estimate the target source position.

BSS demixing filters are considered as a set of 'blind adaptive beamformers' which steer the beam automatically to the sources without any prior information. Usually, the BSS adaptation enhances one spatial null in each BSS channel such that the direct acoustic path of one source is suppressed by exactly this spatial null. Given prior initialization of the filters w_{11} and w_{22} with a positive unit impulse, the spatial null is represented by a minimum (negative peak) of the impulse response of the demixing filter w_{21} or w_{12} and the position of the peak can be used for source localization [7]. Based on this observation, if a source in a predefined angular range is active, a negative peak will appear in the corresponding range of the demixing filter impulse responses. Hence, supposing that only one possibly active source exists in the expected angular range of the target, we can estimate the source direction θ by searching the peak in the relevant range and convert the position of this peak to be $\hat{\theta}$ as an estimate of the true θ .

However, once the BSS is controlled by the geometric constraint, the peak will always be forced into the position corresponding to $\hat{\theta}$, even if the target source moves from θ to another position. In order to estimate the actual source location fast and reliably, a Shadow BSS without a geometric constraint running in parallel to the main Directional BSS is introduced, which is designed to react fast to variations of the source position by virtue of its short filter length and periodical reinitialization [8]. As illustrated in Fig.1, the Shadow BSS detects the movement of the target source and delivers $\hat{\theta}$ as the estimate of the true current θ to the Directional BSS. In this way, Directional BSS can apply the geometric constraint according to the current $\hat{\theta}$ and follows the target source movement.

III. EXPERIMENTAL RESULTS

A. Comparison of Directional BSS with a Delay&Subtract beamformer

In this section, we compare Directional BSS as BM to a frequency-independent Delay&Subtract beamformer furnished with perfect lo-

calization information with respect to the frequency response and performance of target speech suppression in different acoustic environments.

In the following, the overall system behavior is studied. The BM (Delay&Subtract beamformer or Directional BSS) is steered to 0° . To compare both BM concepts, the transfer function for different source positions $-90^\circ \leq \phi < 90^\circ$ is evaluated as depicted in Fig. 3. s and ϕ represent the point source and the corresponding position. h_{0i} , $i \in \{1, 2\}$ denotes the i -th Room Impulse Response (RIR) from the source to the corresponding microphone. The RIRs were measured in a living-room-like environment with a reverberation time of approximately $T_{60} \approx 300$ ms. The measurements were performed for two different microphone spacings, $d_{\text{mic}} \in \{9, 15\}$ cm. The distance between the source and the microphones was 1.1m. The output signal of the BM is denoted by y_1 and the filter coefficients are represented by w_{11} and w_{21} . For an idealized frequency-independent Delay&Subtract beamformer steered to 0° , w_{11} and w_{21} are given by 1 and -1 for all frequencies, respectively. For Directional BSS, the coefficients are a set of converged BSS demixing filters of length 1024 adapted with only one source located at 0° . For the following evaluation, the sampling frequency was set to 16kHz.

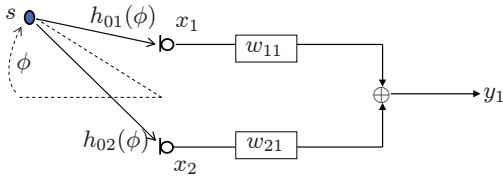


Fig. 3. System to evaluate the frequency response of different BMs

In general, the spatiotemporal frequency response associated with the BM is given by:

$$\underline{h}_{\text{trans}}(\omega, \phi) = \frac{y_1}{s} = \underline{h}_{01}(\omega, \phi)w_{11}(\omega) + \underline{h}_{02}(\omega, \phi)w_{21}(\omega). \quad (8)$$

In Figs. 4 and 5 the magnitude responses for an idealized frequency-independent Delay&Subtract beamformer and Directional BSS are depicted for the array of $d_{\text{mic}} = 9$ cm and $d_{\text{mic}} = 15$ cm respectively. Please note that these are no beampatterns in the usual sense where it is assumed that the acoustic waves propagate in the free field and no scatterers are considered. Instead, (8) also considers the acoustic environment by taking into account the transfer functions from the source position to the microphones. Thereby, $\underline{h}_{\text{trans}}$ also captures reflections of source signals originating from given source positions. Thus, if (8) exhibits a minimum for a certain angle it indicates that all signal components originating from this angle, including possible reflections at surfaces in the acoustic environment, are suppressed. Comparing (a) and (b) in Figs. 4, 5, both BMs have a similar magnitude response. For both BMs, spatial aliasing is unavoidable at $f > 6$ kHz ($d_{\text{mic}} = 9$ cm) and at $f > 2.6$ kHz ($d_{\text{mic}} = 15$ cm). Besides, both BMs do not have a significant spatial selectivity for low frequencies (lower than 300Hz) and it is observed that the frequency range with no spatial selectivity is larger for the case with $d_{\text{mic}} = 9$ cm than for $d_{\text{mic}} = 15$ cm. In this frequency range, not only the target source but also interferers located at positions differing from 0° are suppressed to a large extent and consequently, no noise estimate can be obtained. Despite similar behavior of the two BMs in the range of low frequencies, it is clearly noticeable that Directional BSS achieves a more pronounced spatial null than the Delay&Subtract beamformer which reflects a much better suppression

performance of the Directional BSS compared to a Delay&Subtract beamformer.

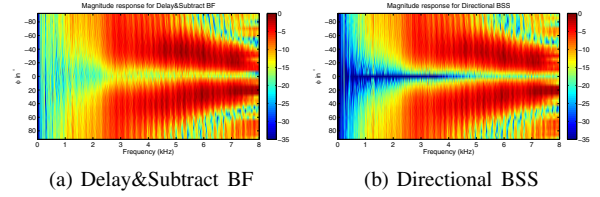


Fig. 4. Magnitude responses [dB] for the two BMs steered to 0° ($d_{\text{mic}} = 9$ cm, $T_{60} \approx 300$ ms)

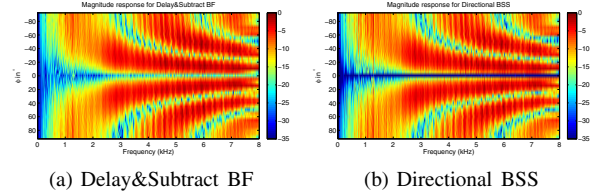


Fig. 5. Magnitude responses [dB] for the two BMs steered to 0° ($d_{\text{mic}} = 15$ cm, $T_{60} \approx 300$ ms)

Since the magnitude response depicts the system behavior for white signals and not for speech signals, we also investigate the speech suppression performance of both BMs. Thus, we define the target speech suppression gain for each channel p as follows:

$$\text{Gain}_{\text{sup}_p} = 10 \log_{10} \left(\frac{\sum_{\omega} |\underline{h}_{0p}|^2 \hat{P}_{ss}}{\sum_{\omega} |\underline{h}_{\text{trans}}|^2 \hat{P}_{ss}} \right), p \in \{1, 2\}, \quad (9)$$

where \hat{P}_{ss} denotes the estimated (long-time) auto Power Spectral Density (PSD) of the clean target signal s , which is calculated here by averaging PSD estimates of 14 different speech signals, each of which lasts 10s. The obtained \hat{P}_{ss} is illustrated in Fig. 6.

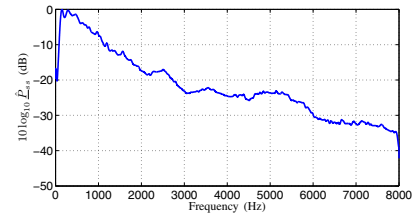


Fig. 6. Averaged PSD of speech signals

In Fig. 7 we compare the target speech suppression performance of the two BMs averaged over both channels for the array of $d_{\text{mic}} = 9$ cm (Fig. 7a) and $d_{\text{mic}} = 15$ cm (Fig. 7b). Since we define a target range of $-20^\circ \leq \theta \leq 20^\circ$, the results are presented for different target locations ranging from -20° to 20° in steps of 5° . Correspondingly, the involved BMs are also steered to the actual target location. Fig. 7 illustrates that Directional BSS has a significantly better target source suppression performance than Delay&Subtract beamformer which is attributed to its dereverberating effect.

B. Tracking capability of the proposed scheme

In order to test how fast the proposed scheme can react to the movement of the target source, we simulated a worst-case scenario with three speakers talking simultaneously at equal level and additional

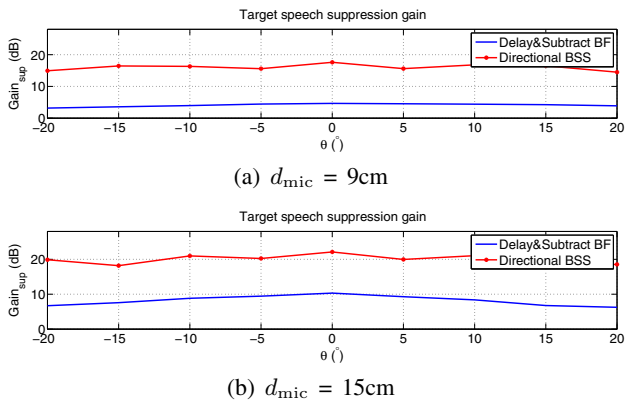


Fig. 7. Target speech suppression gain of the two BMs ($T_{60} \approx 300\text{ms}$)

background babble noise (Fig. 8a). Fig. 8b prescribes the movement of the target source. At $t = 10\text{s}$ the target source moves instantaneously from 0° to 15° and at $t = 20\text{s}$ it moves instantaneously from 15° to -15° .

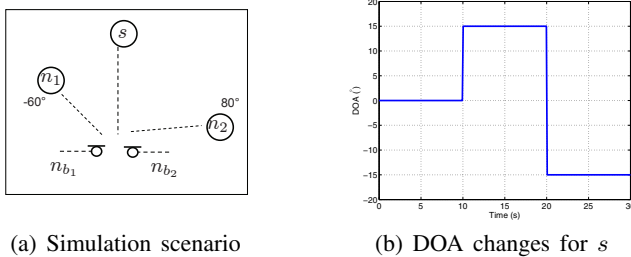


Fig. 8. Simulation scenario for testing tracking ability

Fig. 9 depicts the temporal target speech suppression gain (Fig. 9b) and the estimated DOA of the target source (Fig. 9c) for the array of $d_{\text{mic}} = 15\text{cm}$. The upper plot (Fig. 9a) illustrates the temporal SIR of the input signals. It can be seen that the Shadow BSS can reliably estimate the true DOA within 1s after the target source moves (Fig. 9c). Moreover, as long as the estimated DOA is correct, Directional BSS can quickly converge to a stable target speech suppression of 10-15dB (Fig. 9b) (except in the two speech pauses between 12-14s and 26-28s), which indicates that the proposed scheme is suitable for a real-world application like teleconferencing with personalized computers or laptops. In fact, a computer-based real-time demonstrator with a two-element microphone array is already implemented and verifies the tracking capability and robustness of this scheme.

IV. CONCLUSION

In this contribution, a novel interference estimation using Directional BSS for blind source extraction has been presented. The proposed concept takes advantage of BSS with a geometric constraint to deal with the underdetermined BSS scenario such that a meaningful joint estimate for both interfering speech signals and diffuse background noise can be obtained using only two microphones. Experimental results confirm that the target speech suppression performance of the proposed method is superior to the performance of a common Delay&Subtract beamformer even if the latter is furnished with ground-truth source location information. Besides, both simulations and a real-time demonstrator verify the robustness of the proposed

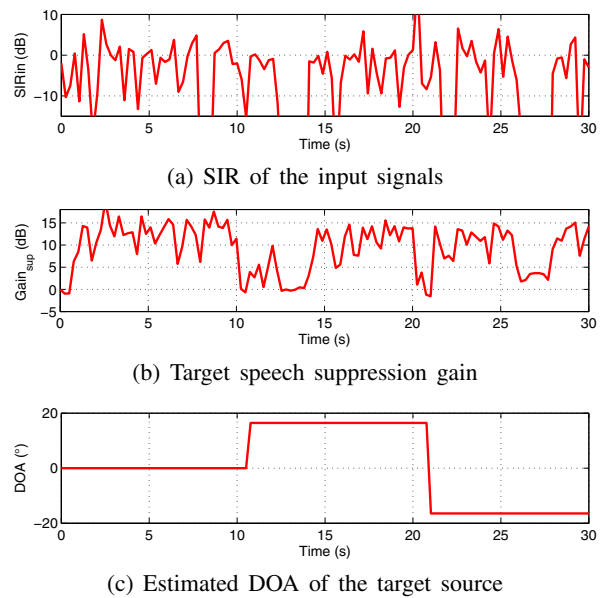


Fig. 9. Performance of the proposed scheme: Temporal target speech suppression gain and estimated DOA of the target source ($d_{\text{mic}} = 15\text{cm}$, $T_{60} \approx 300\text{ms}$)

method. It can be expected that such a noise and interference estimate will lead to reduced speech distortion in Wiener-filtering techniques due to the reduced target speech residual in the noise estimate. Complementary research for possible applications focusses on the exploitation of this noise estimate by Wiener-type filters to yield an enhanced target signal.

REFERENCES

- [1] O. Hoshuyama and A. Sugiyama, *A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters*, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Atlanta, USA, May, 1996.
- [2] W. Herboldt, H. Buchner, S. Nakamura, and W. Kellermann, *Application of a Double-Talk Resilient DFT- Domain Adaptive Filter for Bin-wise Stepsize Controls to Adaptive Beamforming*, Int. Workshop on Nonlinear Signal and Image Processing (NSIP), Sapporo, Japan, May, 2005.
- [3] L. C. Parra and C. V. Alvino, *Geometric Source Separation: Merging Convolutional Source Separation With Geometric Beamforming*, IEEE Transactions on Speech and Audio Processing, vol. 10, pp. 352-362, Sep. 2002.
- [4] E. Visser and T. W. Lee, *Speech Enhancement using Blind Source Separation and Two-Channel Energy Based Speaker Detection*, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Hong Kong, China, April, 2003.
- [5] Y. Takahashi, K. Osaka, H. Saruwatari, and K. Shikano, *Blind Source Extraction for Hands-Free Speech Recognition Based on Wiener Filtering and ICA-Based Noise Estimation*, Hands-Free Speech Communication and Microphone Arrays (HSCMA), Trento, Italy, May, 2008.
- [6] H. Buchner, R. Aichner, and W. Kellermann, *A Generalization of Blind Source Separation Algorithms for Convolutional Mixtures Based on Second-Order Statistics*, IEEE Transactions on Speech and Audio Signal Processing, vol. 13, no. 1, pp. 120-134, Jan. 2005.
- [7] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann, *Simultaneous Localization of Multiple Sound Sources using blind adaptive MIMO Filtering*, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Philadelphia, USA, March, 2005.
- [8] S. Wehr, A. Lombard, H. Buchner, and W. Kellermann, *Shadow BSS for Blind Source Separation in Rapidly Time-Varying Acoustic Scenes*, IEEE Int. Symp. Independent Component Analysis and Blind Separation (ICA), London, UK, Sep. 2007.