ANALYSIS OF THE PERFORMANCE AND LIMITATIONS OF ICA-BASED RELATIVE IMPULSE RESPONSE IDENTIFICATION

Stefan Meier and Walter Kellermann

Multimedia Communications and Signal Processing
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
{stefan.a.meier, walter.kellermann}@fau.de

ABSTRACT

Estimating impulse responses for a single source is a crucial problem for many applications in audio signal processing, such as source extraction. Since absolute impulse responses are hard to identify, relative impulse responses or, equivalently, relative transfer functions are identified instead. Independent Component Analysis (ICA) for convolutive mixtures offers the possibility to determine relative impulse responses implicitly by separating the target source from interfering sources. In this paper, fundamental limitations of relative transfer function (RTF) estimation are analyzed by calculating least-squares (LS)-optimal estimates in adverse scenarios, where the influence of scatterers and reverberation on the performance must be accounted for. Hereupon, ICA-based RTF estimation in the TRINICON framework is compared with the LS-optimal estimates.

Index Terms— Independent component analysis, impulse response estimation, relative transfer functions

1. INTRODUCTION

A common problem in array signal processing is the extraction of a desired (point) source from a mixture of sources. In the past, numerous methods for multichannel signal extraction have been developed, the most popular ones being the well-known minimum variance distortionless constraint (MVDR) beamformer [1, 2] or its more general form, the linearly constraint minimum variance (LCMV) beamformer [3], and the multichannel Wiener filter (MWF) [4]. While the MWF requires an estimation of correlation matrices for the undesired signal components, which can, e.g., be performed while the desired source ('target') is inactive, LCMV beamforming requires the knowledge of room impulse responses from the target source to the sensors for extraction of the original source signal. In practice, estimating the absolute impulse responses from the source position to the microphones is a difficult task. A simpler, but still valuable alternative, however, is estimating the relative impulse responses between the microphones [5–7].

With LCMV beamforming as an application in mind, in this paper, Triple-N Independent Component Analysis for Convolutive Mixtures (TRINICON)-based convolutive ICA [8, 9] is investigated for estimating the relative impulse responses in a moderately reverberant acoustic environment. Extending the work in [8], where only free-field propagation is considered, the influence of the coherence of the target source components in the input signals and of scatterers between the sensors (such as the head in binaural hearing aids) is analyzed.

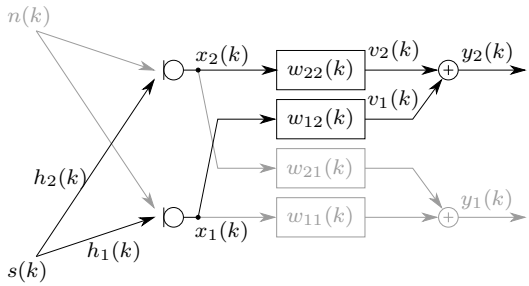


Fig. 1: Signal model for ICA-based source separation.

The paper is organized as follows: In Section 2, the signal model is presented. Hereupon, the TRINICON-based convolutive ICA method is recapitulated in Section 3 and the relation between ICA and relative room impulse response estimation is discussed in Section 4. In Section 5, the general limitations of RTF estimation are evaluated by considering LS-optimal estimates. Finally, experiments with TRINICON ICA are performed in Section 6 and compared with the optimum results.

2. SIGNAL MODEL

In Fig. 1, a typical scenario for ICA is depicted: The two microphones capture a mixture of the target source $s(k)$ filtered with absolute room impulse responses $h_1(k)$ and $h_2(k)$, and components of interferers or noise $n(k)$, which yields the input signals

$$x_p(k) = h_p(k) * s(k) + n_p(k), \qquad p \in \{1, 2\}, \quad (1)$$

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

where $*$ denotes (linear) convolution. In a determined two-source blind source separation (BSS) application, $n(k)$ is a point source and the components $n_p(k)$ can analogously to the target source be described by a convolution with impulse responses. Since we are only interested in the target source and for a more general representation, however, this relation is omitted in the signal model.

3. ICA-BASED BLIND SOURCE SEPARATION

ICA [10] aims at separating individual point sources from a mixture of signals by minimizing the mutual information between the output channels and, thus, maximizing statistical independence. We will consider the TRINICON-based ICA approach for convolutive mixtures as discussed in [11–13]. In this section, we briefly recapitulate the theory of TRINICON ICA.

3.1. TRINICON for blind source separation

As mentioned before, the goal of ICA is the minimization of the mutual information between the output channels $y_1(k)$ and $y_2(k)$. In the TRINICON framework, this minimization can be expressed in terms of a cost function [13]

$$\mathcal{J}_{\text{BSS}}(m) = \sum_{i=0}^{\infty} \beta(i,m) \frac{1}{N} \sum_{k=iL}^{iL+N-1} \log \frac{p_{\mathbf{y},2D}\left(\mathbf{y}(k)\right)}{\prod_{l=1}^{2} p_{\mathbf{y}_l,D}(\mathbf{y}_l(k))}, \tag{2}$$

where $\mathbf{y}_l(k)$ are $1 \times D$ vectors containing the D most recent samples of $y_l(k)$ and $\mathbf{y}(k) = [\mathbf{y}_1(k), \mathbf{y}_2(k)]$. $p_{\bullet,M}(\bullet)$ denotes an M-variate probability density function (pdf), L is the filter length, N is the block length, $\beta(i,m)$ is a window, and i and m are block indices. Note that in the case of independent outputs, the $2D$-variate pdf becomes equal to the product of D-variate pdfs and, hence, the whole term is minimized.

While broadband approaches do not suffer from the internal permutation problem inherent to narrowband approaches (i.e., the output channel order differs from subband to subband [14]), an external permutation ambiguity still exists, such that it cannot be controlled which source will appear in which output channel. In order to avoid this ambiguity, the authors in [15] proposed adding a geometric constraint to one output channel. Assuming that the target source $s(k)$ in Fig. 1 should be *suppressed* in output channel 2, the filters $w_{12}(k)$ and $w_{22}(k)$ are constrained by the cost function

$$\mathcal{J}_{\text{GC}} = \left|\mathbf{W}^{\dagger}(\mu)\mathbf{e}(\mu)\right|^2, \tag{3}$$

where $\mathbf{W}(\mu) = [W_{12}(\mu), W_{22}(\mu)]^{\mathsf{T}}$, $W_{12}(\mu)$ and $W_{22}(\mu)$ being the DFT-domain representations of $w_{12}(k)$ and $w_{22}(k)$, and $\mathbf{e}(\mu)$ is a steering vector of the form

$$\mathbf{e}(\mu) = \left[1, \exp\left(-\mathrm{j}\frac{2\pi\mu}{N}\frac{d}{c}\cos(\varphi_{\text{tar}})f_{\text{s}}\right)\right]^{\mathsf{T}}. \tag{4}$$

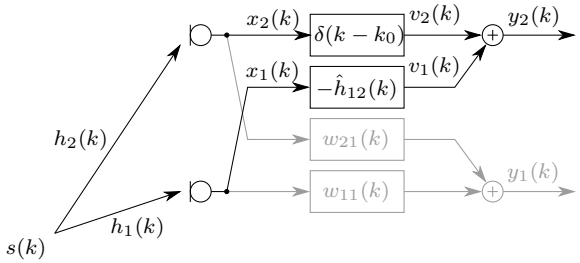


Fig. 2: Estimation of the RIR $\hat{h}_{12}(k)$.

In the steering vector, φ_{tar}, d, c and f_{s} denote the angle of the direction of arrival (DOA) of the target source relative to the array axis (0° being the left endfire position) which is assumed to be known, the microphone distance, the speed of sound and the sampling rate, respectively. By pushing adaptation towards a solution where $\mathbf{W}(\mu)$ is orthogonal to $\mathbf{e}(\mu)$, the geometric constraint tries to create a relative delay between the two filters that compensates for the time difference of arrival (TDOA) of the target source, hence, leading to a cancellation of the target source in Channel 2. Using a weighted sum of both cost functions, a solution where the target source appears in output signal $y_1(k)$ and the sum of all interfering sources appears in signal $y_2(k)$, is sought.

3.2. Ambiguity of the ICA solution

We assume that due to the geometric constraint in (3), the target source appears at the output $y_1(k)$ while being suppressed in output y_2. This implies that the target source components in signals $v_1(k)$ and $v_2(k)$ (see Fig. 1) are equal (with a different sign). As a result, the filters $w_{12}(k)$ and $w_{22}(k)$ have to satisfy the relation

$$s(k) * h_1(k) * w_{12}(k) + s(k) * h_2(k) * w_{22}(k) = 0 \tag{5}$$

and, thus, if this should hold for all possible signals $s(k)$,

$$h_1(k) * w_{12}(k) = -h_2(k) * w_{22}(k) \tag{6}$$

From (6), it is evident that the convolutive ICA problem does not have a unique solution. By restricting the adaptation of the filters, however, relative impulse responses can be estimated, which will be further discussed in the following section.

4. APPLICATION OF ICA-BASED BLIND SOURCE SEPARATION TO RIR ESTIMATION

As discussed in [8], ICA can be exploited for estimating the relative impulse responses between the two microphones. In this paper, we will distinguish two versions:

1. By fixing $w_{22}(k)$ to an integer delay $\delta(k-k_0)$ as depicted in Fig. 2, the filter $w_{12}(k)$ is forced to the solution

$$w_{12}(k) = -\hat{h}_{12}(k) = -h_1^{-1}(k) * h_2(k) * \delta(k-k_0), \tag{7}$$

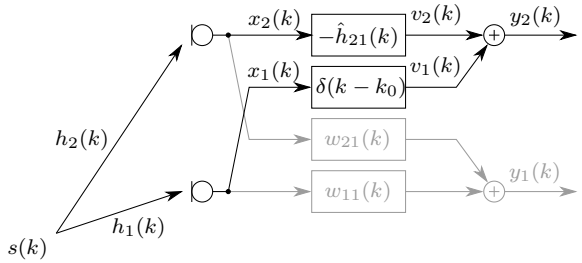


Fig. 3: Estimation of the RIR $\hat{h}_{21}(k)$.

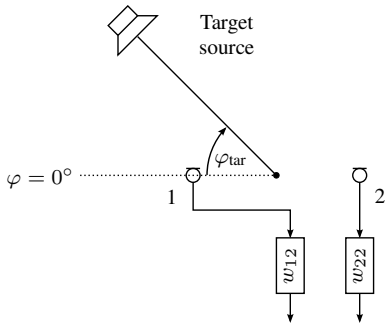


Fig. 4: Evaluation scenario for single-source scenarios.

where $\hat{h}_{12}(k)$ denotes the relative impulse response between microphone 1 and microphone 2. The delay $k_0 > 0$ needs to be specified in order to be able to identify non-causal impulse responses.

2. Equivalently, fixing $w_{12}(k)$ to $\delta(k - k_0)$ as depicted in Fig. 3 allows for identifying the relative impulse response $h_{21}(k)$ between microphone 2 and microphone 1.

$$w_{22}(k) = -\hat{h}_{21}(k) = -h_2^{-1}(k) * h_1(k) * \delta(k-k_0). \quad (8)$$

5. LIMITATIONS OF RTF ESTIMATION

In this section, structural limitations of RTF estimation in adverse scenarios are analyzed. To this end, we consider a scenario as shown in Fig. 4. A target source is active at an azimuth $\varphi_{\text{tar}} \in [0°, 360°)$. For the experiment, the microphone signals of the target source were generated by convolving a speech signal with impulse responses of a moderately reverberant room ($T_{60} \approx 400$ms) at a distance of 1m, captured with hearing aids mounted on a head as scatterer. All simulations were performed at a sampling rate $f_s = 16$kHz. Filter $w_{22}(k)$ was fixed to a unit impulse $\delta(k - k_0)$ such that only filter $w_{12}(k)$ was adapted, which should approach $-\hat{h}_{12}(k)$ as shown in (7). In order to analyze the limitations of RTF estimation, the filter was determined by calculating the LS estimate

$$\mathbf{w}_{12,\text{LS}} = \arg\min_{\mathbf{w}} \mathcal{E}\left\{\left(x_2(k-k_0) - \mathbf{w}^{\mathsf{T}}\mathbf{x}_1(k)\right)^2\right\}, \quad (9)$$

with $\mathbf{w}_{12,\text{LS}} = [w_{12,\text{LS}}(0), \ldots, w_{12,\text{LS}}(L-1)]^{\mathsf{T}}$ and $\mathbf{x}_1(k) = [x_1(k), \ldots, x_1(k-L+1)]^{\mathsf{T}}$. For the evaluation, the normalized RTF error (NRE) is used, which we define in the DFT domain as

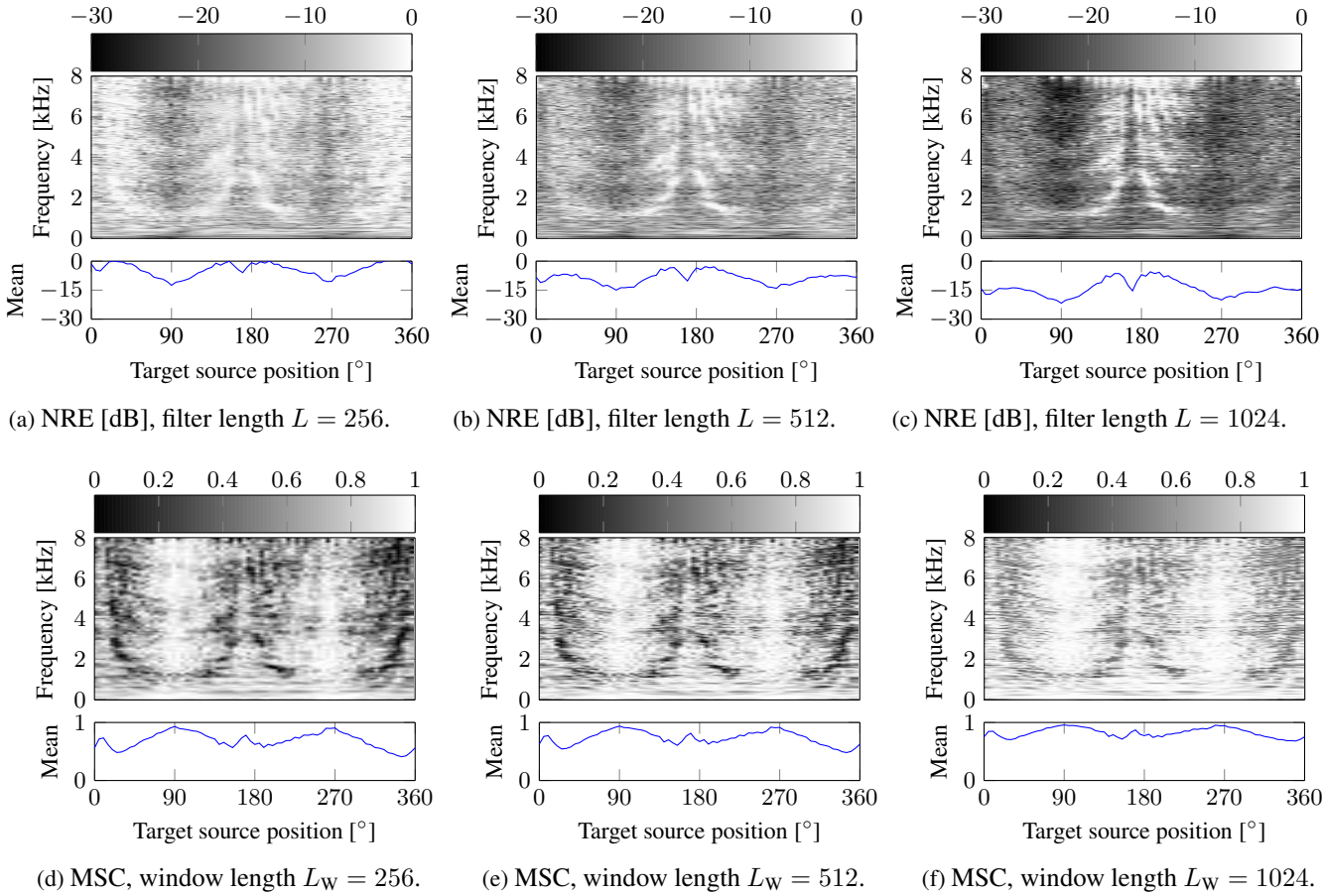
$$\text{NRE}_{ij}(\mu) = 20\log_{10}\left[\frac{|H_{ij,\text{est}}(\mu) - H_{ij,\text{true}}(\mu)|}{|H_{ij,\text{true}}(\mu)|}\right]. \quad (10)$$

The true RTFs are calculated by dividing the DFTs of $h_1(k)$ and $h_2(k)$, while the estimated relative transfer function is calculated by transforming the LS estimate according to (9) into the DFT domain. Irrespective of the filter length, a DFT length of 2048 is used.

In Fig. 5a–c, the results obtained for filter length $L \in \{256, 512, 1024\}$ and $k_0 = \frac{L}{8}$ (which turned out to be a suitable choice for all filter lengths) are shown dependent on target source position and frequency. Below the color plot, the average over all frequencies is plotted. By comparing the three subfigures, we can draw the following conclusions: First of all, the RTF estimation yields the best results for source positions in the left half plane (especially for filter length $L = 1024$). This fact can be explained by considering (7): In order to obtain the optimum solution, the impulse response $h_1(k)$ needs to be inverted. In the left half plane ($0° \leq \varphi_{\text{tar}} \leq 90°$ and $270° \leq \varphi_{\text{tar}} < 360°$), a direct path between the target source and microphone 1, whereas in the right half plane ($90° \leq \varphi_{\text{tar}} \leq 270°$), $h_1(k)$ contains strong scattering effects. Obviously, inverting $h_1(k)$ would require much more taps than inverting $h_2(k)$ does. Moreover, all three figures exhibit a characteristic pattern, where the worst results in each half plane are obtained at angles of approximately $\pm 20°$ relative to endfire position. In order to further analyze this issue, the coherence of the microphone signals calculated by means of the Welch method with window lengths equal to the LS filter lengths is plotted in Fig. 5d–f. By comparing Fig. 5a–c and Fig. 5d–f, we can see a strong relation between the coherence of the input signal and the achievable RTF estimation accuracy.

6. EVALUATION OF ICA-BASED RTF ESTIMATION

In Section 5, the limitations of the RTF identification in the scenario at hand were evaluated by deriving optimum LS estimates from single-source signals. In this section, the performance achievable with TRINICON ICA will be compared with the optimum performance. For the ICA method as explained in [12], a filter length of $L = 1024$, a block size for calculating the correlation matrices of $K = 4096$ and a weight of $\eta = 0.5$ for the directional constraint were chosen. Speech signals of length 30s were used as source signals and 500 iterations were applied to the ICA algorithm. In order to prevent the directional constraint (which assumes free-field propagation) from unnecessarily disturbing the adaptation, the directional constraint was only applied if the relative



(a) NRE [dB], filter length $L = 256$. (b) NRE [dB], filter length $L = 512$. (c) NRE [dB], filter length $L = 1024$.

(d) MSC, window length $L_{\mathrm{W}} = 256$. (e) MSC, window length $L_{\mathrm{W}} = 512$. (f) MSC, window length $L_{\mathrm{W}} = 1024$.

Fig. 5: NRE in [dB] (a–c) and magnitude squared coherence (d–f) of the target source components dependent on target source position, frequency and filter / window length.

delay between the filters w_{12} and w_{22} did not match the desired delay. Due to the symmetry of the scenario, we only consider the range from 0° to 180°.

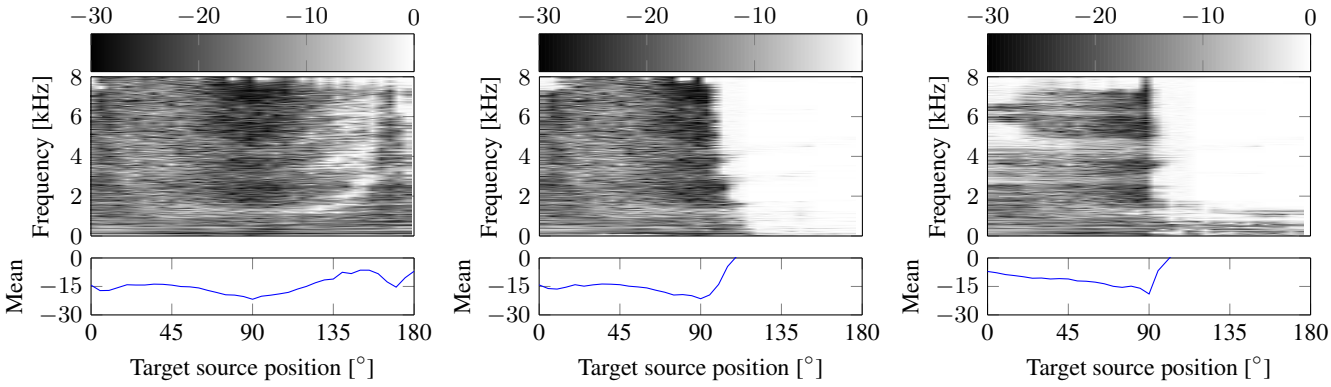
The simulations are organized as follows: As a first step, the performance achieved in the absence of an interferer (and additive noise) is compared with that achieved with the LS-optimal filters. After that, an interferer will be added at the opposite endfire position relative to the target source (i.e., 0° for $\varphi_{\text{tar}} \in [90^\circ, 180^\circ]$ and 180° for $\varphi_{\text{tar}} \in [0^\circ, 90^\circ)$) in order to analyze the effect of interference on the RTF estimate.

6.1. Single-source scenario

In Fig. 6b, the performance of ICA-based RTF estimation is shown in terms of NRE. A comparison with the LS-optimal results in Fig. 6a reveals two conclusions: 1.) For $\varphi_{\text{tar}} \in [0^\circ, 90^\circ]$, ICA achieves results that are very close to the LS estimate, and only slight degradations at frequencies close to 8kHz are observable. 2.) On the other hand, for $\varphi_{\text{tar}} \in [90^\circ, 180^\circ]$, the performance deteriorates strongly. This effect can be explained as follows: Since our desired solution in (7) implicitly requires an inversion of the impulse response $h_1(k)$, the filter w_{12} requires less samples for $\varphi_{\text{tar}} \in [0^\circ, 90^\circ]$ (where the direct path is dominant) than for $\varphi_{\text{tar}} \in (90^\circ, 180^\circ]$ (where the direct path is influenced by the scatterer), which was already observable from the LS results. Since we only constrain output channel 2, ICA still has the possibility to suppress the target source in output channel 1 (instead of channel 2 as enforced) and, hence, model the simpler inversion of $h_2(k)$ in filter $w_{21}(k)$. Although the geometric constraint creates a peak in the filter w_{12} to model the desired delay, the ICA adaptation creates other peaks at different positions, such that in total, the desired RTF cannot be modeled. A simple remedy, however, consists in estimating the impulse responses specified in (8) and inverting this solution.

6.2. Two-source scenario

When an additional interferer is active, a degradation becomes observable (average over all frequencies 2.5–4.2dB for $20^\circ \leq \varphi_{\text{tar}} \leq 90^\circ$ and up to 8.3dB for $\varphi_{\text{tar}} = 5^\circ$). For low frequencies ($\leq$ 4kHz), however, which are dominant



(a) NRE [dB] with LS. (b) NRE [dB] with ICA, 1 source. (c) NRE [dB] with ICA, 2 sources.

Fig. 6: Comparison of the NRE achieved with LS and ICA for $0° \leq \varphi_{\text{tar}} \leq 180°$ and a filter length $L = 1024$.

in speech signals and attenuated less by the scatterer, the mean NRE over all positions only increases by 1.2dB from -15.0dB to -13.8dB.

7. CONCLUSIONS

In this paper, the influence of scatterers and limited filter lengths on relative transfer function estimation was investigated in a reverberant scenario ($T_{60} \approx 400$ms). It was shown that there exists a strong relation between the coherence of the target source components and the achievable identification error (even with the optimum LS estimate). Since relative transfer function estimation implicitly involves inverting one of the two absolute impulse responses, a scatterer can lead to pronounced differences in terms of alignment error for different choices of the reference signal. This effect can be expected from considering the LS-optimal solutions (where the reference signal is explicitly defined) and becomes crucial in the semiblind ICA approach, where the desired solution is only enforced by an additional constraint. With a suitable choice of the reference signal, however, ICA achieves results that are similar to the LS estimates in single-source scenarios and degrade only slightly in the presence of an interferer.

REFERENCES

[1] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[2] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[3] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[4] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, 2002.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, vol. 1, Springer, Berlin/Heidelberg, 2008.

[6] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[7] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.

[8] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2013.

[9] K. Reindl, S. Meier, H. Barfuss, and W. Kellermann, "Minimum mutual information-based linearly constrained broadband signal extraction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1096–1108, June 2014.

[10] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, New York, 2001.

[11] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Montreal, Canada, 2004, pp. 889–892.

[12] 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 Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[13] H. Buchner, R. Aichner, and W. Kellermann, *Blind Source Separation for Convolutive Mixtures: A Unified Treatment*, Kluwer Academic Publishers, Boston, Feb. 2004.

[14] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*, Springer, Berlin/Heidelberg, 2005.

[15] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Proc. Int. Workshop on Comp. Advances in Multi-Sensor Adapt. Proc. (CAMSAP)*, Aruba, Dutch Antilles, Dec. 2009.