Combined LCMV-TRINICON beamforming for separating multiple speech sources in noisy and reverberant environments

Shmulik Markovich-Golan[1,2], *Member, IEEE,* Sharon Gannot[1], *Senior Member, IEEE,* and Walter Kellermann[2], *Fellow, IEEE*

***Abstract*—The problem of source separation using an array of microphones in reverberant and noisy conditions is addressed. We consider applying the well-known linearly constrained minimum variance (LCMV) beamformer (BF) for extracting individual speakers. Constraints are defined using relative transfer functions (RTFs) for the sources, which are acoustic transfer functions (ATFs) ratios between any microphone and a reference microphone. The latter are usually estimated by methods which rely on single-talk time segments where only a single source is active and on reliable knowledge of the source activity. Two novel algorithms for estimation of RTFs using the TRINICON (Triple-N ICA for convolutive mixtures) framework are proposed, not resorting to the usually unavailable source activity pattern. The first algorithm estimates the RTFs of the sources by applying multiple two-channel geometrically constrained (GC) TRINICON units, where approximate direction of arrival (DOA) information for the sources is utilized for ensuring convergence to the desired solution. The GC-TRINICON is applied to all microphone pairs using a common reference microphone. In the second algorithm, we propose to estimate RTFs iteratively using GC-TRINICON, where instead of using a fixed reference microphone as before, we suggest to use the output signals of LCMV-BFs from the previous iteration as spatially processed references (SPRs) with improved signal-to-interference-and-noise ratio (SINR). For both algorithms, a simple detection of noise-only time segments is required for estimating the covariance matrix of noise and interference. We conduct an experimental study in which the performance of the proposed methods is confirmed and compared to corresponding supervised methods.**

***Index Terms*—Blind source separation, LCMV, relative transfer function, voice activity**

I. Introduction

Speech enhancement problems have attracted the attention of both industry and research community for several decades. Applications and use cases are vast and diverse, such as telecommunications, entertainment and multimedia systems, human-machine interfaces, hearing aids and many more. Technological advances enabled the use of microphone arrays, allowing to utilize spatial properties of the signals and extending the spectral-temporal filtering methods to spatial-temporal methods.

In this contribution, we consider noise reduction and speech enhancement problems for multiple speakers in determined and over-determined scenarios where the number of sources is less or equal to the number of microphones. Most algorithms for this problem can be classified as either *supervised*, which usually optimize a second-order statistics (SOS)-based criterion given a priori information about the spatial configuration or activity-patterns of the sources, or *unsupervised*, leading to so-called *blind* algorithms, which utilize statistical independence and often higher-order statistics (HOS). Algorithms of both classes are based on applying linear filters to the microphone signals followed by a summation, also known as *beamforming*.

Supervised algorithms for speech enhancement are surveyed in [1]–[4]. The multichannel Wiener filter (MWF) is a BF which obtains the minimum mean squared error (MMSE) for the estimated desired speech signal. The speech-distortion-weighted (SDW)-MWF [5] extends the MWF and enables control over the tradeoff between desired signal distortion and noise suppression by introducing a relative weighting of the residual noise component in the optimization. In the multiple speakers scenario a mean squared error (MSE) expression with multiple weighted distortion components and individual desired responses can be optimized, denoted multiple speech distortion weighted (MSDW)-MWF [6].

The LCMV beamformer [7] optimizes the noise variance at the output while maintaining a set of linear constraints. A special case of the LCMV which satisfies a single constraint of keeping a desired speech signal undistorted at the output is denoted as the minimum variance distortionless response (MVDR) beamformer. The linearly constrained BFs have an equivalent and efficient form, denoted as generalized sidelobe canceler (GSC) [8], [9]. A common approach is to design and apply the BF in the short-time Fourier transform (STFT) domain, e.g., the transfer function GSC (TF-GSC) [10]. Note that in [6] the MVDR and LCMV can be obtained as extreme cases of the SDW-MWF and MSDW-MWF, respectively.

Designing the above-mentioned BFs requires knowledge of SOS or ATFs of the various sources as well as of the noise. In [10] it was suggested to use the ratios between source-microphones ATFs and the ATF of a reference microphone, denoted RTFs, in the construction of the MVDR-BF. Several procedures exist for estimating the RTF: noise covariance subtraction (CS) [11]–[14]; noise covariance whitening (CW) [14]–[16] or methods based on speech non-stationarity [10], [17].

Blind source separation (BSS) algorithms, as implied by their name, do not require any information about the sources, and are solely based on their statistical properties, such as mutual statistical independence or non-Gaussian distribution. For a survey on BSS methods please refer to [18]–[23]. Early contributions rely on criteria derived from SOS whereas in more recent contributions HOS is also considered, also known as independent component analysis (ICA). The permutation ambiguity, which may arise from applying scalar BSS individually for each frequency bin in the STFT domain, can be mitigated by introducing a smoothing operating over nearby frequencies [24], [25], or by soft or hard geometrical constraints [26], [27]. Alternatively, a broadband criterion, defined in the time domain, can be optimized in the frequency domain for efficiency and completely avoiding the internal permutation ambiguity [28].

In [29], [30], a framework denoted TRINICON for blind processing of multiple input/multiple output (MIMO) systems based on minimizing mutual information is proposed. Specifically in BSS, Triple-N ICA for convolutive mixtures (TRINICON) is attractive as it can simultaneously exploit properties of non-stationarity, non-

[1] The Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel.
[2] The Multimedia Comm. and Sig. Proc., Univ. Erlangen-Nuremberg, Cauerstr. 7, Erlangen, Germany.

whiteness, and non-Gaussianity of the involved signals. The concept of introducing a geometrical constraint into the BSS criterion was incorporated into the TRINICON algorithm in [31], denoted GC-TRINICON. It was used for *source extraction*, i.e., extracting a desired source out of multiple sources, in the under-determined scenario. This concept was then used to extend the TRINICON criterion in [32] by introducing a set of linear constraints, denoted linearly-constrained minimum mutual information (LCMMI). The latter constraints allow to designate the target speakers based on their DOA and to improve the convergence by effectively reducing the number of degrees of freedom in the BF optimization. This work was further generalized in [33] for the estimation of RTFs using multiple TRINICON procedures applied to microphone pairs. For the initial estimates of DOAs of the sources, blind ICA-based algorithms such as [34], [35] can be used. Both of these methods are applicable to the multiple-speakers convolutive mixtures case. Real-time and reduced complexity versions of the TRINICON algorithm have been suggested in [36] and [37]. In [38] a real-time implementation to a GPU platform is described.

In the current contribution, we extend [33] by incorporating the estimation method for source RTFs into a set of LCMV-BFs designed for extracting individual sources. We propose two algorithms in which GC-TRINICON is used for RTF estimation which is later used for constructing LCMV-BFs. The first algorithm uses multiple two-channel GC-TRINICON units whose input signals are given by a chosen reference signal and one of the other sensor signals. The second proposed algorithm also applies the GC-TRINICON procedure to pairs of signals. However, instead of using one of the microphone signals as reference we propose to use the outputs of LCMV-BFs as spatially processed references (SPRs). This procedure can be applied iteratively resulting in incrementally improved SPRs contributing to an improved RTF estimate.

Various available algorithms which work under idealized conditions (such as using perfect source activity knowledge and perfect noise knowledge by the LCMV [15], or as using training data by the multichannel non-negative matrix factorization (MCNMF) [39]) obtain satisfactory results in source-separation problems. The main motivation of the proposed algorithm is that it offers real-time capability and requires no training, but a simple detector for identifying noise-only time segments and coarse estimates of the sources' DOAs. Yet, its performance gets close to the optimum one of algorithms operating under idealized conditions.

The paper is structured as follows. The problem is formulated in Sec. II. Sections III and IV give necessary background on LCMV beamforming and the TRINICON framework, respectively. In Sec. V we propose a combined GC-TRINICON based RTF estimation [33] and a LCMV-BF. The second proposed algorithm, which combines GC-TRINICON and LCMV beamforming using SPRs in an iterative estimation procedure, is derived in Sec. VI. Results of an experimental study comparing the proposed methods with ideal and estimation-based supervised methods as well with the state-of-the-art MCNMF method [39] are given in Sec. VII. Finally, the contribution is summarized and main conclusions are drawn in Sec. VIII.

II. Problem formulation

Consider a scenario with P speakers in a reverberant enclosure received by a microphone array with M microphones. Let $s_p(n)$ for $p = 1, \ldots, P$ denote the P speech signals received by a reference microphone, here identified with $m = 1$. The received signal at the mth microphone is given by:

$$x_m(n) = \sum_{p=1}^{P} h_{pm}(n) * s_p(n) + v_m(n), \tag{1}$$

where $h_{pm}(n)$ denotes the relative impulse response (RIR) RIR between the the reverberant components of the p-th speaker, also denoted for brevity as the components of the p-th speaker, at the m-th microphone and at the reference microphone, and $v_m(n)$ comprises all noise and interference components at the m-th microphone ($m = 1, \ldots, M$). Note that $h_{p1}(n) \triangleq \delta(n)$, where $\delta(n)$ is the unit impulse. Transforming the problem to the STFT domain with a window length of K and overlap of κ yields:

$$\underline{x}_m(\ell, k) \triangleq \sum_{p=1}^{P} \underline{h}_{pm}(k)\underline{s}_p(\ell, k) + \underline{v}_m(\ell, k) \tag{2}$$

where $\underline{h}_{pm}(k)$ is the discrete Fourier transform (DFT) of $h_{pm}(n)$, also known as the RTF between the p-th source components at the m-th and at the reference microphones. Note that underlined expressions, i.e., $\underline{\bullet}$, denote terms in the DFT or STFT domains. The indices ℓ and k, corresponding to the STFT domain, stand for the time-frame index and frequency-bin, respectively, whereas the index n indicates sample time. It is assumed that the length of the RIRs is shorter than the window length K, such that convolution in the time domain is transformed into multiplication in the STFT domain. For brevity, and since the algorithms presented and proposed in this paper estimate the required parameters using a batch of samples, we confine ourselves to the static scenario. Thus, for our analysis we assume that speakers are static and that the noise is stationary. In many practically relevant scenarios, the time-variance of the systems to be identified can be assumed to be sufficiently slow, so that time-invariance for the duration of the signal segments can be assumed. The extension of the proposed algorithms to rapidly changing acoustic scenarios is left for future work. Please refer to [17] and [40] for tracking the RTF in dynamic scenarios. Actually, the stationarity of the noise is not required by the TRINICON method, as long as the signal to noise ratio (SNR) is sufficiently high. However, we adopt the stationarity assumption for a consistent formulation.

Concatenating the microphone components at each frequency into a single $M \times 1$ vector, the following vector notation can be defined:

$$\underline{\boldsymbol{x}}(\ell, k) \triangleq \underline{\boldsymbol{H}}(k)\underline{\boldsymbol{s}}(\ell, k) + \underline{\boldsymbol{v}}(\ell, k) \tag{3}$$

where

$$\underline{\boldsymbol{x}}(\ell, k) \triangleq \left[\begin{array}{ccc} \underline{x}_1(\ell, k) & \cdots & \underline{x}_M(\ell, k) \end{array}\right]^T \tag{4a}$$

$$\underline{\boldsymbol{s}}(\ell, k) \triangleq \left[\begin{array}{ccc} \underline{s}_1(\ell, k) & \cdots & \underline{s}_P(\ell, k) \end{array}\right]^T \tag{4b}$$

$$\underline{\boldsymbol{v}}(\ell, k) \triangleq \left[\begin{array}{ccc} \underline{v}_1(\ell, k) & \cdots & \underline{v}_M(\ell, k) \end{array}\right]^T \tag{4c}$$

$$\underline{\boldsymbol{h}}_p(k) \triangleq \left[\begin{array}{ccc} \underline{h}_{p1}(k) & \cdots & \underline{h}_{pM}(k) \end{array}\right]^T;\ p = 1, \ldots, P \tag{4d}$$

$$\underline{\boldsymbol{H}}(k) \triangleq \left[\begin{array}{ccc} \underline{\boldsymbol{h}}_1(k) & \cdots & \underline{\boldsymbol{h}}_P(k) \end{array}\right] \tag{4e}$$

and $(\bullet)^T$ denotes the transpose operator. The vector $\underline{\boldsymbol{h}}_p(k)$ is denoted as the vector containing the RTFs for the p-th source. Define the time-invariant noise correlation matrix and the time-varying microphone signals correlation matrix at the k-th frequency bin as:

$$\underline{\boldsymbol{\Phi}}_{vv}(k) \triangleq \mathrm{E}\left[\underline{\boldsymbol{v}}(\ell, k)\,\underline{\boldsymbol{v}}^H(\ell, k)\right] \tag{5}$$

$$\begin{aligned}\underline{\boldsymbol{\Phi}}_{xx}(\ell, k) &\triangleq \mathrm{E}\left[\underline{\boldsymbol{x}}(\ell, k)\,\underline{\boldsymbol{x}}^H(\ell, k)\right] \\ &= \underline{\boldsymbol{H}}(k)\underline{\boldsymbol{\Lambda}}_s(\ell, k)\,\underline{\boldsymbol{H}}^H(k) + \underline{\boldsymbol{\Phi}}_{vv}(k)\end{aligned} \tag{6}$$

respectively, where

$$\underline{\boldsymbol{\Lambda}}_s(\ell, k) \triangleq \mathrm{diag}\left\{\underline{\lambda}_{s1}(\ell, k), \ldots, \underline{\lambda}_{sP}(\ell, k)\right\} \tag{7}$$

denotes the covariance matrix of the sources' components at the reference microphone,

$$\underline{\lambda}_{sp}(\ell, k) \triangleq \mathrm{E}\left[|\underline{s}_p(\ell, k)|^2\right] \tag{8}$$

denotes the time-varying power of the component of the p-th source at the reference microphone at the (ℓ, k) time-frequency bin, for $p = 1, \ldots, P$ and $\mathrm{diag}\{\bullet\}$ denotes a diagonal matrix with its vector argument on its diagonal.

Our goal is to separate the mixtures of the speech components at the reference microphone and reduce the noise at the outputs $y_p(n)$, for $p = 1, \ldots, P$. In other words, we want to generate estimates of

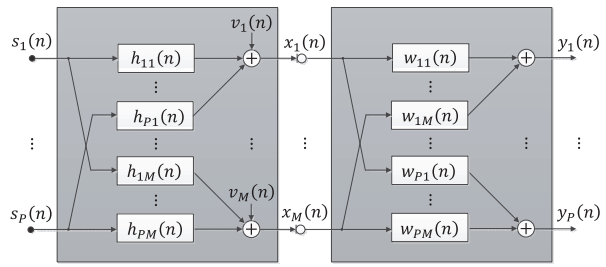


Fig. 1: Source separation scenario.

P desired signals, where in each estimated signal only one speaker is enhanced and all other speakers and noise are attenuated. We assume that the speakers' components at the reference microphone are of sufficient quality, and therefore, do not consider the dereverberation problem. We confine ourselves to solutions consisting of filtering the microphone channels followed by a summation. Hence the output signals can be formulated in the time domain and in the STFT domain, respectively, as:

$$y_p(n) \triangleq \sum_{m=1}^{P} w_{pm}(n) * x_m(n), \tag{9a}$$

$$\underline{y}_p(\ell,k) \triangleq \underline{\boldsymbol{w}}_p^T(k)\underline{\boldsymbol{x}}(\ell,k), \tag{9b}$$

where $y_p(n)$ is an estimate of $s_p(n)$ for $p = 1,\ldots,P$, $w_{pm}(n)$ are two-sided, i.e, possibly non-causal finite impulse response (FIR) filters of length $L = L_c + L_{nc}$, $w_{pm}(k)$ denotes their transformation to the DFT domain and $\underline{\boldsymbol{w}}_p(k)$ is defined by concatenating the transformed p-th signal filters of all microphones $\underline{\boldsymbol{w}}_p^T(k) \triangleq \begin{bmatrix} w_{p1}(k) & \cdots & w_{pM}(k) \end{bmatrix}$. Thereby L_c and L_{nc} are the lengths of the causal and anticausal components, respectively. The STFT frame length is selected such that $L < K$. We assume for simplicity that the speech sources do not move, hence the separating filters can also be time-invariant. This assumption is not mandatory, and separating moving sources with time-varying ATFs can be obtained by extending the algorithms proposed in this contribution. However, this is beyond the scope of the current contribution. A block-diagram of the source separation problem is depicted in Fig. 1.

Furthermore, we assume that for the separation task oracle information regarding the DOAs of the various speakers, denoted θ_p for $p = 1,\ldots,P$, is available. For the suppression of noise and interference, information on noise-only time segments is required. The concatenation of all DOAs into a single vector is defined as $\boldsymbol{\theta} \triangleq [\theta_1,\ldots,\theta_P]^T$. Estimating the sources' DOAs and detecting noise-only time segments are out of the scope of this contribution, please refer to [41], [42] for some detection methods. The required accuracy of the estimates is discussed in Sec. IV-B.

III. LCMV BEAMFORMER

A. Background

Given the RTFs of all sources, and the noise covariance matrix, a LCMV-BF which extracts the p-th source while suppressing all other sources and noise can be defined per frequency bin as in [15] (note that we omit the frequency-bin index for brevity):

$$\underline{\boldsymbol{w}}_p^{\mathrm{LI}} \triangleq \left(\underline{\boldsymbol{\Phi}}_{vv}^{-1}\underline{\boldsymbol{H}}\left(\underline{\boldsymbol{H}}^H\underline{\boldsymbol{\Phi}}_{vv}^{-1}\underline{\boldsymbol{H}}\right)^{-1}\boldsymbol{e}_p\right)^*. \tag{10}$$

The latter BF is defined for $p = 1,\ldots,P$, where $\boldsymbol{e}_p$ is a selection vector, used for extracting the p-th element of a $P\times 1$ vector, defined as:

$$\boldsymbol{e}_p \triangleq \begin{bmatrix} \mathbf{0}_{1\times(p-1)} & 1 & \mathbf{0}_{1\times(P-p)} \end{bmatrix}^T. \tag{11}$$

The superscript $(\bullet)^{\mathrm{LI}}$ stands for LCMV in an ideal scenario where all RTFs are available a priori. Usually, source RTFs as well as noise covariance matrices are not available a priori and need to be estimated from the received signals. After construction, the filters can be transformed back into the time domain and applied as in Fig. 1.

B. Noise covariance whitening (CW)-based RTF estimation

CW is a common method for estimating the RTFs in a supervised scenario [14]–[16], when information about the activity of the sources is available. Given a noisy time segment in which all sources are inactive, the noise covariance matrix is estimated using a sample covariance matrix with recursive averaging, denoted by $\hat{\underline{\boldsymbol{\Phi}}}_{vv}$:

$$\hat{\underline{\boldsymbol{\Phi}}}_{vv}(\ell,k) = \alpha\hat{\underline{\boldsymbol{\Phi}}}_{vv}(\ell-1,k) + (1-\alpha)\underline{\boldsymbol{x}}(\ell,k)\underline{\boldsymbol{x}}^H(\ell,k) \tag{12}$$

where $\underline{\boldsymbol{x}}(\ell,k)$ is a frame containing noise only.

Furthermore, given another noisy time segment in which only the p-th source is active, the covariance matrix of the p-th source plus the noise is estimated in the same way, and denoted $\hat{\underline{\boldsymbol{\Phi}}}_{xx}^{p}(\ell,k)$. We assume that the noisy speech time segment is long and that $\alpha \to 1$. Hence, we assume that $\hat{\underline{\boldsymbol{\Phi}}}_{xx}^{p}(\ell,k)$ converges to an average variance and disregard the dependency on time. Then we apply the generalized eigenvalue decomposition (GEVD) [43] to the matrix $\hat{\underline{\boldsymbol{\Phi}}}_{xx}^{p}$ using $\hat{\underline{\boldsymbol{\Phi}}}_{vv}$ as a rotation matrix and denote the principal eigenvector by $\underline{\boldsymbol{\psi}}_{p1}$. Asymptotically, neglecting estimation errors, assuming that the speakers are static and that the noise is stationary, the principal generalized eigenvalue corresponds to the p-th source component. Its corresponding generalized eigenvector $\underline{\boldsymbol{\psi}}_{p1}$ is parallel to the RTFs vector of the p-th source, i.e., $\underline{\boldsymbol{h}}_p$, rotated by the matrix $\hat{\underline{\boldsymbol{\Phi}}}_{vv}^{-1}$. Finally, an estimate for the RTFs vector of the p-th source is obtained by rotating $\underline{\boldsymbol{\psi}}_{p1}$ back to the microphones' domain using $\hat{\underline{\boldsymbol{\Phi}}}_{vv}$, and normalizing the resulting vector by the element corresponding to the reference microphone, i.e.:

$$\hat{\underline{\boldsymbol{h}}}_p^{\mathrm{CW}} \triangleq \frac{\hat{\underline{\boldsymbol{\Phi}}}_{vv}\underline{\boldsymbol{\psi}}_{p1}}{\boldsymbol{e}_1^T\hat{\underline{\boldsymbol{\Phi}}}_{vv}\underline{\boldsymbol{\psi}}_{p1}}. \tag{13}$$

Concatenating all RTF vectors of the CW method into a single matrix yields:

$$\hat{\underline{\boldsymbol{H}}}^{\mathrm{CW}} \triangleq \begin{bmatrix} \hat{\underline{\boldsymbol{h}}}_1^{\mathrm{CW}} & \cdots & \hat{\underline{\boldsymbol{h}}}_P^{\mathrm{CW}} \end{bmatrix}. \tag{14}$$

Similarly to (10), the estimated RTFs are used for constructing $\underline{\boldsymbol{w}}_p^{\mathrm{LCW}}$, i.e., the LCMV-BF using RTF estimates obtained by the CW method:

$$\underline{\boldsymbol{w}}_p^{\mathrm{LCW}} \triangleq \left(\hat{\underline{\boldsymbol{\Phi}}}_{vv}^{-1}\hat{\underline{\boldsymbol{H}}}^{\mathrm{CW}}\left(\left(\hat{\underline{\boldsymbol{H}}}^{\mathrm{CW}}\right)^H\hat{\underline{\boldsymbol{\Phi}}}_{vv}^{-1}\hat{\underline{\boldsymbol{H}}}^{\mathrm{CW}}\right)^{-1}\boldsymbol{e}_p\right)^* \tag{15}$$

for $p = 1,\ldots,P$.

Note that for time-varying spatial properties of the noise, the accuracy of above mentioned method for estimating the RTF will be hampered. However, the TRINICON based method for RTF estimation (described in the following Sec. IV-B), is robust to non-stationary low-power noises.

IV. TRINICON

A. Background

The TRINICON is a MIMO framework for convolutive mixtures of sources. If applied to BSS, unlike the supervised LCMV-BF, it does not require a priori knowledge of the source RTFs, nor training time-sequences with exclusive activity of single sources for their estimation.

The generic optimization criterion of TRINICON [30] is based on comparing the multivariate probability density function (PDF) of the output signals to a source model PDF. The criterion exploits the

non-stationarity, non-whiteness and non-Gaussianity properties of the sources, and for BSS it is given by:

$$\mathcal{J}(i) = \sum_{i'=0}^{\infty} \beta(i',i) \frac{1}{N} \sum_{n=n_{i'1}}^{n_{i'1}} \log \left\{ \frac{\hat{f}_y(\boldsymbol{y}(n))}{\prod_{p=1}^{P} \hat{f}_{y_p}(\boldsymbol{y}_p(n))} \right\}, \quad (16)$$

where the output signals are split into frames consisting of N data blocks of size D each, i denotes the index of the current frame, corresponding to block index $n \in [n_{i0}, n_{i1}]$, $\beta(i',i)$ denotes the weight of the i'-th frame contribution to the optimization cost function and $\hat{f}_{y_p}$ denotes an estimate of the multivariate PDF of the p-th output (over D consecutive samples. A discussion on choosing D can be found in [30]):

$$\boldsymbol{y}_p(n) \triangleq \left[\begin{array}{ccc} y_p(n) & \cdots & y_p(n-D+1) \end{array} \right]^T, \quad (17)$$

and $\hat{f}_y$ denotes the multivariate PDF of all output signals, i.e.,

$$\boldsymbol{y}(n) \triangleq \left[\begin{array}{ccc} \boldsymbol{y}_1^T(n) & \cdots & \boldsymbol{y}_P^T(n) \end{array} \right]^T. \quad (18)$$

The criterion (16) accounts for the above-mentioned signal properties as follows (please refer to [30] for more details):

- *Non-Gaussianity* is exploited using non-Gaussian PDFs for both sources and output signals.
- *Non-whiteness* is accounted for by considering output cross-relations over D time-lagged output samples for each output signal. Thus, both intra-channel and inter-channel statistical relations of the outputs are modeled.
- *Non-stationarity* of the sources is exploited by averaging over multiple frames (of length N data-blocks), each weighted by the weighting function $\beta(i',i)$ with finite support [30].

The TRINICON criterion is applied in the time domain and the complete set of parameters are optimized jointly. Unlike frequency-domain BSS schemes, where the broadband optimization criterion is split into frequency-bin-wise lower-dimensional optimization problems performed in parallel, the full-length separating filters are optimized simultaneously. An efficient implementation of the time-domain TRINICON criterion which transforms part of the computations into the DFT domain is given in [28]. As with all BSS algorithms, TRINICON is known to work best if P is not greater than two. A GPU-based implementation of two-channel SOS-TRINICON BSS is available in [38], and allows for the parallel execution of dozens to hundreds of BSS units in real-time.

B. TRINICON based RTF estimation

The TRINICON framework can also be used for estimating the RTF of a source in the multiple-speakers scenario, by using only two microphones. In [31], [44], a modified TRINICON criterion which incorporates a geometrical constraint is proposed and denoted GC-TRINICON. Using an approximate DOA of a desired source to construct the constraint helps to "direct" the converged filters to block the desired speaker in one of its output signals. The spatial filters which block the desired source are used for obtaining its RTF.

W.l.o.g., assume that microphones 1 and m are used for estimating the RTF of the p-th source in a scenario comprising P sources. The two outputs of the GC-TRINICON, denoted by $z_p^{1m}(n)$ and $z_{\bar{p}}^{1m}(n)$, are the enhanced p-th source and the blocked p-th source signals, respectively. Note that since only two microphones are used for filtering by the GC-TRINICON, theoretically, only one source can be cancelled at each output. Hence, the filters which construct the output $z_{\bar{p}}^{1m}(n)$, which are designed to block the p-th source, can be used for RTF estimation as follows. Let us denote the GC-TRINICON filters which generate $z_p^{1m}(n)$ and $z_{\bar{p}}^{1m}(n)$ as c_{pa}^{1m}, c_{pb}^{1m} and $c_{\bar{p}a}^{1m}$, $c_{\bar{p}b}^{1m}$, respectively. The corresponding output signals are then given by:

$$z_p^{1m}(n) = c_{pa}^{1m}(n) * x_1(n) + c_{pb}^{1m}(n) * x_m(n) \quad (19)$$

$$z_{\bar{p}}^{1m}(n) = c_{\bar{p}a}^{1m}(n) * x_1(n) + c_{\bar{p}b}^{1m}(n) * x_m(n). \quad (20)$$

Formally, the optimization criterion of the GC-TRINICON is defined similarly to (16) as:

$$\mathcal{J}_{\mathrm{GC}}^{p,1m}(i) = \sum_{i'=0}^{\infty} \beta(i',i) \frac{1}{N} \cdot \sum_{n=n_{i'1}}^{n_{i'1}} \log \left\{ \frac{\hat{f}_z^{p,1m}(\boldsymbol{z}^{p,1m}(n))}{\hat{f}_{z_p}^{1m}(\boldsymbol{z}_p^{1m}(n)) \cdot \hat{f}_{z_{\bar{p}}}^{1m}(\boldsymbol{z}_{\bar{p}}^{1m}(n))} \right\} + \eta \sum_{k=1}^{K} |\underline{c}_{\bar{p}a}^{1m}(k) \underline{\hat{h}}_{pm}^{0}(k) + \underline{c}_{\bar{p}b}^{1m}(k)|^2, \quad (21)$$

where η is a weight controlling the tradeoff between the constraint and the original unconstrained TRINICON criterion. The term $\underline{\hat{h}}_{pm}^{0}$ denotes an initial estimate of the RTF of the p-th source (relating the m-th microphone to the reference microphone), estimated based on the DOA of the source. Eq. (21) defines the p-th source as desired, without loss of generality (w.l.o.g.), $\hat{f}_{z_p}^{1m}$, $\hat{f}_{z_{\bar{p}}}^{1m}$ denote estimates of the multivariate PDFs of the vectors $\boldsymbol{z}_p^{1m}(n)$, $\boldsymbol{z}_{\bar{p}}^{1m}(n)$, respectively, defined over D consecutive samples as:

$$\boldsymbol{z}_p^{1m}(n) \triangleq \left[\begin{array}{ccc} z_p^{1m}(n) & \cdots & z_p^{1m}(n-D+1) \end{array} \right]^T \quad (22)$$

$$\boldsymbol{z}_{\bar{p}}^{1m}(n) \triangleq \left[\begin{array}{ccc} z_{\bar{p}}^{1m}(n) & \cdots & z_{\bar{p}}^{1m}(n-D+1) \end{array} \right]^T \quad (23)$$

and $\hat{f}_z^{p,1m}$ denotes the multivariate PDF of

$$\boldsymbol{z}^{p,1m}(n) \triangleq \left[\begin{array}{cc} \left(\boldsymbol{z}_p^{1m}(n)\right)^T & \left(\boldsymbol{z}_{\bar{p}}^{1m}(n)\right)^T \end{array} \right]^T. \quad (24)$$

An estimate of the RTF of the p-th source relating the m-th microphone (for $m = 2, \ldots, M$) and the reference microphone can be obtained by the GC-TRINICON:

$$\underline{\hat{h}}_{pm}^{\mathrm{GT}}(k) = -\frac{\underline{c}_{\bar{p}a}^{1m}(k)}{\underline{c}_{\bar{p}b}^{1m}(k)} \quad (25)$$

for $p = 1, \ldots, P$ and for all frequency bins $k = 1, \ldots, K$, where the superscript $(\bullet)^{\mathrm{GT}}$ stands for GC-TRINICON.

Finally, we define the initial estimates of the RTFs, i.e., $\underline{\hat{h}}_{pm}^{0}$ for $p = 1, \ldots, P$ and $m = 1, \ldots, M$. The RTFs of the various speakers are initialized as in [44] according to their corresponding relative delays between the microphones and the reference microphone (denoted time difference of arrival (TDOA)) computed according to a priori knowledge or estimates of the speaker DOAs. Given a coarse estimate of the DOA of the p-th speaker, θ_p, the TDOA τ_{pm} of the corresponding direct-arrival components at the m-th microphone and at the reference microphone are explicitly given by:

$$\tau_{pm} = \frac{d_m \sin(\theta_p)}{c} \quad (26)$$

where we assume that a linear array is used to simplify notation, d_m is the distance between the m-th microphone and the reference microphone and c is the sound velocity. Consequently, the p-th speaker RTF of the m-th microphone is initialized as:

$$\underline{\hat{h}}_{pm}^{0}(k) \triangleq \exp\left(-\mathrm{j}2\pi \frac{k}{K} f_s \tau_{pm}\right) \quad (27)$$

for $p = 1, \ldots, P$ and $m = 1, \ldots, M$ where f_s denotes the sampling frequency and we assume far-field wave propagation. Using vector notation, similarly to (4d), the initial estimate of the RTFs vector of the p-th source is defined as:

$$\underline{\hat{\boldsymbol{h}}}_{p}^{0}(k) \triangleq \left[\begin{array}{ccc} \underline{\hat{h}}_{p1}^{0}(k) & \cdots & \underline{\hat{h}}_{pM}^{0}(k) \end{array} \right]^T \quad (28)$$

for $p = 1, \ldots, P$.

It should be noted that the considered criterion in (21) relies on the existence of a dominant direct-path signal, i.e., at least a moderate direct to reverberant ratio (DRR) is assumed. Therefore, unsatisfactory separation results must be expected if reverberation prevails over the direct propagation path. According to [44], the TDOA estimates need to be sufficiently accurate to allow neighbouring sources to be distinguished.

V. Combined LCMV-TRINICON

A simple and straightforward way of combining the LCMV-BF with TRINICON-based RTF estimation is proposed next. First, the RTFs relating the p-th source components at microphones $m = 2, 3, \ldots, M$ to the reference microphone, the first microphone, are estimated by the procedure described in Sec. IV-B. The latter RTFs are denoted $\hat{\underline{h}}_{pm}^{\text{GT}}$, for $m = 2, \ldots, M$. Define $\hat{\underline{h}}_{p1}^{\text{GT}} \triangleq 1$. Similarly to (4d), define the vector of RTFs of the p-th source, estimated using the GC-TRINICON algorithm as:

$$\hat{\underline{\boldsymbol{h}}}_p^{\text{GT}} \triangleq \left[\begin{array}{ccc} \hat{\underline{h}}_{p1}^{\text{GT}} & \cdots & \hat{\underline{h}}_{pM}^{\text{GT}} \end{array} \right]^T. \tag{29}$$

The latter procedure is repeated for all sources, i.e., $p = 1, \ldots, P$. Note, that the GC-TRINICON RTF estimation procedure is applied $P \cdot (M-1)$ times in parallel. Similarly to (4e), define the concatenation of all RTFs into a single matrix by:

$$\hat{\underline{\boldsymbol{H}}}^{\text{GT}} \triangleq \left[\begin{array}{ccc} \hat{\underline{\boldsymbol{h}}}_1^{\text{GT}} & \cdots & \hat{\underline{\boldsymbol{h}}}_P^{\text{GT}} \end{array} \right]. \tag{30}$$

Finally, in a similar manner to (10), define the LCMV-BF which is constructed using the RTFs estimated with GC-TRINICON algorithm as:

$$\underline{\boldsymbol{w}}_p^{\text{LGT}} \triangleq \left(\hat{\underline{\boldsymbol{\Phi}}}_{vv}^{-1} \hat{\underline{\boldsymbol{H}}}^{\text{GT}} \left(\left(\hat{\underline{\boldsymbol{H}}}^{\text{GT}} \right)^H \hat{\underline{\boldsymbol{\Phi}}}_{vv}^{-1} \hat{\underline{\boldsymbol{H}}}^{\text{GT}} \right)^{-1} \boldsymbol{e}_p \right)^* \tag{31}$$

for $p = 1, \ldots, P$, where the superscript $(\bullet)^{\text{LGT}}$ stands for LCMV-BF with RTFs estimates using the GC-TRINICON algorithm, and $\hat{\underline{\boldsymbol{\Phi}}}_{vv}$ is an estimate of the noise covariance matrix, estimated during a noise-only time segment.

VI. Combined LCMV-TRINICON with spatially processed references

In this section we propose an iterative algorithm which combines the TRINICON criterion for estimating the RTFs of the sources and the LCMV beamformer for separating the sources. The algorithm is denoted as the LCMV-SPR-TRINICON. The complexity of the proposed algorithm is analyzed in Sec. VI-D.

A. Spatially processed reference TRINICON-based RTF estimation

Here, we extend the GC-TRINICON by replacing the reference microphone signal with a SPR, which has an improved signal to interference ratio (SIR). The modified RTFs, relating the desired signal components in all microphones with the modified reference, are estimated with higher accuracy as the SIR of the modified reference increases. The RTFs with respect to the microphone signal are obtained by proper normalization of the latter modified RTF estimates. The GC-TRINICON can be obtained as a special case of the SPR-GC-TRINICON with a simple spatial processing which selects the reference microphone.

Consider estimating the RTFs of the p-th speaker, i.e., $\underline{h}_{pm}(k)$, for $m = 1, \ldots, M$. Given a *modified reference* signal:

$$r_p(n) \triangleq \sum_{m=1}^{M} q_{pm}(n) * x_m(n) \tag{32}$$

where at this point $q_{pm}(n)$ is an arbitrary filter. Note that by transforming (32) into the frequency-domain and substituting (2) and (3) we obtain:

$$\underline{r}_p(\ell, k) = \sum_{p=1}^{P} \underline{\boldsymbol{q}}_p^T(k) \underline{\boldsymbol{h}}_p(k) \underline{s}_p(\ell, k) + \underline{\boldsymbol{q}}_p^T(k) \underline{\boldsymbol{v}}(\ell, k). \tag{33}$$

Denote the p-th speaker modified reference RTFs relating the components of the p-th source at $x_m(n)$ and $r_p(n)$ by:

$$\underline{g}_{pm}(k) \triangleq \frac{\underline{h}_{pm}(k)}{\underline{\boldsymbol{q}}_p^T(k) \underline{\boldsymbol{h}}_p(k)}, \tag{34}$$

where we assume that $\underline{\boldsymbol{q}}_p^T(k) \underline{\boldsymbol{h}}_p(k) \neq 0$. Given an initial estimate of $\hat{g}_{pm}^0(n)$, the GC-TRINICON is applied for obtaining a more accurate estimate, denoted $\hat{g}_{pm}(n)$. The modified reference $r_p(n)$ is filtered by the a priori estimate $\hat{g}_{pm}^0(n)$, yielding:

$$r_{pm}(n) \triangleq \hat{g}_{pm}^0(n) * r_p(n) \tag{35}$$

in the time domain and

$$\underline{r}_{pm}(\ell, k) \triangleq \hat{\underline{g}}_{pm}^0(k) \cdot \underline{r}_p(\ell, k) \tag{36}$$

in the STFT domain. By substituting the p-th source component of $\underline{r}_p(\ell, k)$ from (33) in (36), its corresponding component coincides with the p-th source component in $\underline{x}_m(\ell, k)$, up to estimation errors of the initial RTF $\hat{\underline{g}}_{pm}^0(\ell, k)$:

$$\begin{aligned} &\hat{\underline{g}}_{pm}^0(k) \cdot \underline{\boldsymbol{q}}_p^T(k) \underline{\boldsymbol{h}}_p(k) \underline{s}_p(\ell, k) \\ &\quad \approx \frac{\underline{h}_{pm}(k)}{\underline{\boldsymbol{q}}_p^T(k) \underline{\boldsymbol{h}}_p(k)} \cdot \underline{\boldsymbol{q}}_p^T(k) \underline{\boldsymbol{h}}_p(k) \underline{s}_p(\ell, k) = \underline{h}_{pm}(k) \underline{s}_p(\ell, k). \end{aligned}$$

The GC-TRINICON with an initial constraint of a unit impulse, i.e, $\delta(n)$, is then applied to the input signals $r_{pm}(n)$ and $x_m(n)$. The more $\hat{g}_{pm}^0(n)$ is accurate the closer is $r_{pm}(n)$ to the component of the p-th source at the m-th microphone, resulting in an estimated RTF which is closer to a unit impulse. The estimated RTF relating the components of the p-th source at $x_m(n)$ and $r_{pm}(n)$, denoted by $\hat{\delta}_{rpm}(n)$, is used for updating the estimate of the RTF between the components of the p-th source at the m-th microphone and at the reference microphone, i.e.:

$$\hat{g}_{pm}^{\text{ST}}(n) \triangleq \hat{g}_{pm}^0(n) * \hat{\delta}_{rpm}(n), \tag{37}$$

where the superscript $(\bullet)^{\text{ST}}$ stands for SPR-GC-TRINICON. A block-diagram of the SPR-GC-TRINICON is depicted in Fig. 2. The notation $(\bullet)|\theta_p$ is added to the initial RTF in the figure to emphasize that it is computed based on the a priori known DOA. For brevity, we omit this explicit DOA dependency in the text.

We apply the above-mentioned SPR-GC-TRINICON M times, over all microphones with the p-th modified reference. Similarly to (4d), a vector of modified reference RTFs is defined as:

$$\underline{\boldsymbol{g}}_p(k) \triangleq \left[\begin{array}{ccc} \underline{g}_{p1}(k) & \cdots & \underline{g}_{pM}(k) \end{array} \right]^T. \tag{38}$$

Note that the p-th speaker RTFs vector, $\underline{\boldsymbol{h}}_p(k)$, can be expressed in terms of the modified reference RTFs vector, i.e., $\underline{\boldsymbol{g}}_p(k)$, by:

$$\underline{\boldsymbol{h}}_p(k) = \frac{1}{\underline{g}_{p1}(k)} \underline{\boldsymbol{g}}_p(k). \tag{39}$$

Hence, we propose to estimate $\underline{\boldsymbol{h}}_p(k)$ by:

$$\hat{\underline{\boldsymbol{h}}}_p^{\text{ST}}(k) = \frac{1}{\hat{\underline{g}}_{p1}^{\text{ST}}(k)} \hat{\underline{\boldsymbol{g}}}_p^{\text{ST}}(k) \tag{40}$$

where, similarly to (38), $\hat{\underline{\boldsymbol{h}}}_p^{\text{ST}}$ is defined as:

$$\hat{\underline{\boldsymbol{g}}}_p^{\text{ST}} \triangleq \left[\begin{array}{ccc} \hat{\underline{g}}_{p1}^{\text{ST}} & \cdots & \hat{\underline{g}}_{pM}^{\text{ST}} \end{array} \right]^T. \tag{41}$$

A block-diagram for estimating the RTFs of the p-th speaker using the SPR-GC-TRINICON is depicted in Fig. 3.

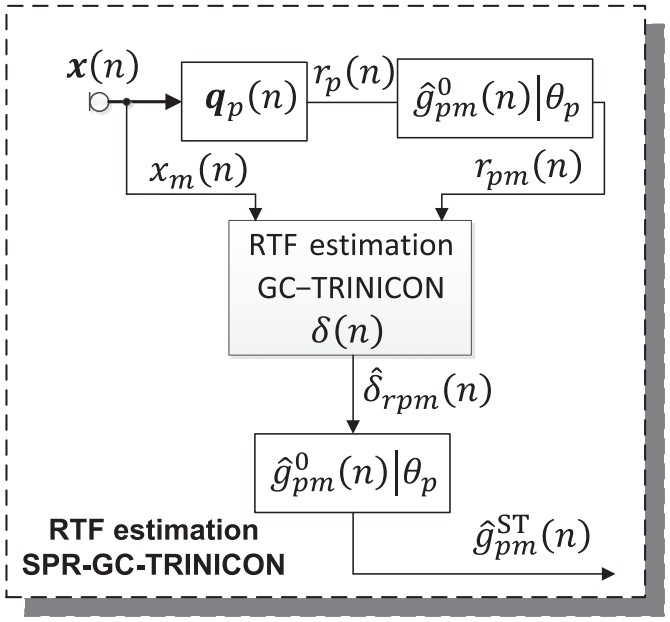


Fig. 2: Block-diagram of SPR-GC-TRINICON for estimating the RTF between the m-th microphone and the modified reference of the p-th source.

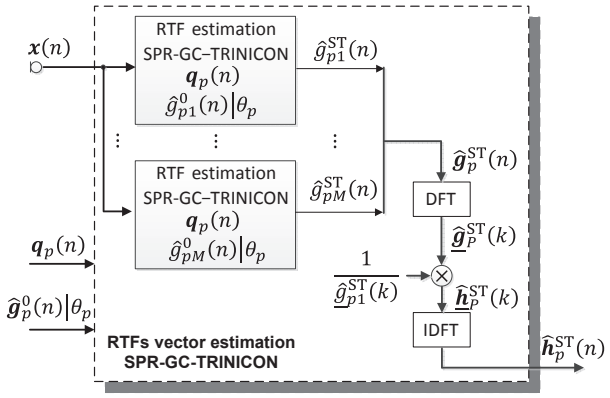


Fig. 3: Block-diagram of SPR-GC-TRINICON for estimating the RTFs of the p-th source.

B. Algorithm

The higher the SIR of the modified reference at the GC-TRINICON is, the more accurate is the estimated RTF. Thus, we propose an iterative two-stage algorithm for speaker separation: 1) estimate vectors of RTFs of all speakers; 2) construct modified reference signals for each of the speakers using LCMV-BFs. The latter two stages can be repeated for improved performance.

Let us consider the i-th iteration of the algorithm. Denote by $\hat{\boldsymbol{h}}_p^{\mathrm{ST},i-1}$ and $\boldsymbol{w}_p^{\mathrm{LST},i-1}$, the estimated RTFs and the LCMV-BFs at iteration $i-1$, for $p=1,\ldots,P$. We apply the SPR-GC-TRINICON, and update the estimated RTF vector of the p-th source, i.e., $\hat{\underline{\boldsymbol{h}}}_p^{\mathrm{ST},i}$, for $p=1,\ldots,P$. We use the p-th output of the previous iteration as the SPR of the current iteration, i.e., the SPR filters used for constructing the SPR are defined as:

$$\underline{\boldsymbol{q}}_p^i(k) \triangleq \underline{\boldsymbol{w}}_p^{\mathrm{LST},i-1}(k) \tag{42}$$

where the superscript *LST* denotes LCMV-SPR-TRINICON and the corresponding SPR is:

$$\underline{r}_p^i(\ell,k) \triangleq \left(\underline{\boldsymbol{q}}_p^i(k)\right)^T \underline{\boldsymbol{x}}(\ell,k) \tag{43}$$

for $p=1,\ldots,P$. Note that the p-th source component at the SPR is designed to coincide with the corresponding component at the reference microphone, up to estimation errors of the RTFs. For this specific selection the SPR of the p-th source is distortionless for $p=1,\ldots,P$. Hence, the estimated RTFs relating the p-th source components at the microphone signals with the corresponding component at the p-th SPR, $\hat{\underline{\boldsymbol{g}}}_p^{\mathrm{ST},i-1}$, can be defined as the RTFs corresponding to the p-th source component at the reference microphone, $\hat{\underline{\boldsymbol{h}}}_p^{\mathrm{ST},i-1}$, i.e.:

$$\hat{\underline{\boldsymbol{g}}}_p^{\mathrm{ST},i-1}(k) \triangleq \hat{\underline{\boldsymbol{h}}}_p^{\mathrm{ST},i-1}(k). \tag{44}$$

Given the updated estimates of RTFs vectors, $\hat{\underline{\boldsymbol{h}}}_p^{\mathrm{ST},i}(k)$ for $p=1,\ldots,P$, we update the LCMV-BFs following (10):

$$\underline{\boldsymbol{w}}_p^{\mathrm{LST},i} \triangleq \left(\hat{\underline{\boldsymbol{\Phi}}}_{vv}^{-1}\left(\hat{\underline{\boldsymbol{H}}}^{\mathrm{ST},i}\right)^H\left(\left(\hat{\underline{\boldsymbol{H}}}^{\mathrm{ST},i}\right)^H\hat{\underline{\boldsymbol{\Phi}}}_{vv}^{-1}\hat{\underline{\boldsymbol{H}}}^{\mathrm{ST},i}\right)^{-1}\boldsymbol{e}_p\right)^* \tag{45}$$

where, similarly to (4e), the matrix $\hat{\underline{\boldsymbol{H}}}^{\mathrm{ST},i}$ is defined as:

$$\hat{\underline{\boldsymbol{H}}}^{\mathrm{ST},i} \triangleq \left[\begin{array}{ccc}\hat{\underline{\boldsymbol{h}}}_1^{\mathrm{ST},i} & \cdots & \hat{\underline{\boldsymbol{h}}}_P^{\mathrm{ST},i}\end{array}\right]. \tag{46}$$

The $M\times P$ dimensional matrix $\underline{\boldsymbol{W}}^{\mathrm{LST},i}(k)$ is defined as the concatenation of the separating filters $\underline{\boldsymbol{w}}_p^{\mathrm{LST},i}(k)$ for all sources ($p=1,\ldots,P$) at the i-th iteration:

$$\underline{\boldsymbol{W}}^{\mathrm{LST},i} \triangleq \left[\begin{array}{ccc}\underline{\boldsymbol{w}}_1^{\mathrm{LST},i} & \cdots & \underline{\boldsymbol{w}}_P^{\mathrm{LST},i}\end{array}\right]. \tag{47}$$

This iterative procedure is repeated I times.

C. Initialization

Finally, we define the initialization stage of the algorithm. The initial RTF estimates are defined according to the given or estimated DOAs of the various sources, as in (28), for $p=1,\ldots,P$. The initial separation filters, i.e., $\underline{\boldsymbol{w}}_p^{\mathrm{LST},0}$ for $p=1,\ldots,P$, are defined similarly to (45) using (28), where:

$$\hat{\underline{h}}_{pm}^{\mathrm{ST},0} \triangleq \hat{\underline{h}}_{pm}^{0} \tag{48}$$

for $p=1,\ldots,P$ and $m=1,\ldots,M$.

A block-diagram of the iterative LCMV-SPR-TRINICON method for RTFs estimation is depicted in Fig. 4. The notation $\hat{\underline{\boldsymbol{H}}}^{\mathrm{ST},0}(k)|\boldsymbol{\theta}$ in the figure is meant to emphasize that the initial RTF estimates are constructed from the a priori DOA information. For brevity, we omit this explicit DOA dependency in the text.

D. Complexity analysis of the combined LCMV-SPR-TRINICON method

In this section we analyze the complexity of the proposed combined LCMV-SPR-TRINICON method. The number of computations depends on the length of the time-segment which is used for constructing the beamformer, denoted T, and the number of iterations I.

The computation is comprised of several steps: 1) transforming the microphone signals to the STFT domain; 2) constructing the noise covariance matrix; 3) estimating the RTFs using the proposed SPR-TRINICON method; 4) constructing the LCMV beamformers for separating the sources; 5) generating the separated sources. Note that the computation of steps $3-5$ is repeated for each iteration, i.e. I times. Furthermore, the computation of all steps, except for the construction of the LCMV beamformers, depend linearly on the length of the time-segment, T.

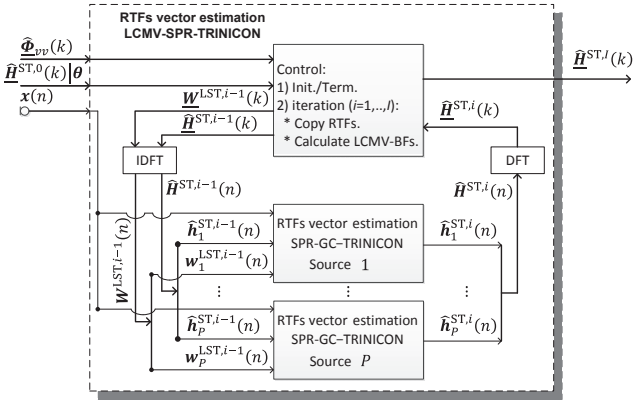


Fig. 4: Block-diagram of the iterative LCMV-SPR-TRINICON method for RTFs estimation.

For a detailed computational analysis of the TRINICON algorithm please refer to [45]. Note that SPR-GC-TRINICON comprises of $M \cdot P$ applications of the GC-TRINICON. Denote the complexity per sample of the TRINICON algorithm by C_{T}. According to [45], for the parameters that we use here, $C_{\mathrm{T}} \approx 2 \times 10^4$. The complexity of constructing a single output LCMV beamformer (see [46]) is approximately $\frac{2}{3}(M^3 + P^3) + 2M^2P$.

The complexity of the various steps is summarized in Table. I. Clearly from this table, the complexity of the proposed LCMV-SPR-TRINICON method is dominated by the RTFs' estimation process, which can be approximated as $I \cdot M \cdot P \cdot C_{\mathrm{T}} \cdot T$. However, an efficient GPU real-time implementation of the TRINICON algorithm is described in [38].

Step	Number of computations	
	Analytic expression	Empirical value
1) STFT of the inputs.	$\frac{1}{1-\kappa} M \log_2 K \cdot T$	1.3×10^5
2) Construct noise covariance matrices.	$\frac{1}{2(1-\kappa)} M^2 T$	5.8×10^5
A single iteration:		
3) SPR-GC-TRINICON	$M \cdot P \cdot C_{\mathrm{T}} \cdot T$	2.9×10^9
4) Construct LCMV	$\left(\frac{2}{3}\left(M^3 + P^3\right) + 2M^2P\right) \frac{K \cdot P}{2}$	4.5×10^6
5) Generate separated sources.	$\frac{1}{1-\kappa} P \cdot M \cdot T$	5.8×10^5

TABLE I: Complexity of the combined LCMV-SPR-TRINICON algorithm detailed for each of its components. Empirical values refer to the parameters used in the nominal case of the experimental study for a 1s time-segment (i.e., for: $T = 8000$; $I = 3$; $P = 3$; $M = 6$; $K = 2048$; $\kappa = 75\%$; $L = 2048$; $N = 4096$; 75% overlap between TRINICON blocks).

VII. Experimental Study

We evaluate the performance of the proposed algorithms in an experimental study. In Sec. VII-A we describe the compared algorithms and the performance criteria. Two types of scenarios are considered for the testing: 1) simulated scenarios using speech sources and acoustic impulse responses (IRs) taken from published databases (see Sec. VII-B); 2) real-life recordings of real persons recorded in an acoustic lab (see Sec. VII-C). Examples of input and output signals of the various algorithms can be found in [47].

A. Compared algorithms and evaluation criteria

We compare the following algorithms:

- Ideal LCMV-BF given the true RTFs and noise covariance matrix, denoted by *LI* in short. The results of this algorithm serve as an upper bound for the achievable performance.
- Practical LCMV-BF with CW estimates of RTFs (see Sec. III-B), denoted by *LCW* in short. Perfectly detected single-talk time segments for each of the speakers are used for estimating the RTFs.
- An LCMV-BF computed from the steering vectors constructed from the a priori knowledge of the sources DOA, denoted by *LD* in short. These weights are denoted as $\underline{\boldsymbol{W}}^{\mathrm{LD}}$ and are equal to the initialization weights of the proposed LCMV-SPR-TRINICON algorithm, i.e. $\underline{\boldsymbol{W}}^{\mathrm{LD}} \triangleq \underline{\boldsymbol{W}}^{\mathrm{LST},0}$ (see Sec. VI-C).
- MCNMF in convolutive mixtures [39], representing state-of-the-art algorithms of the BSS family and denoted by *NMF* in short. [1]
- Proposed algorithm which combines LCMV-TRINICON (see Sec. V), denoted by *LGT* in short.
- Proposed combined LCMV-SPR-TRINICON algorithm (see Sec. VI), denoted by *LST* in short.

The window length of the STFT is set to $K = 2048$ and the overlap is set to $\kappa = 75\%$. In all algorithms, except for the ideal LCMV-BF, the noise covariance matrix defined by (5) is estimated by time-recursive sample covariance matrix during a noise-only time segment at the beginning of the recording.

An efficient SOS realization of the TRINICON update rule [30] is adopted here by applying D-variate Gaussian PDF models in (21). The filter length is set to $L = 2048$, the block size for correlation matrix estimation is $N = 4096$ samples, and the desired speakers' DOAs are available to the algorithm (through *oracle* knowledge). The overlap between TRINICON blocks is 75%.

The MCNMF is trained using the separated speakers components, which are artificially contaminated by an additive white Guassian noise at a SNR of 10dB.

The performance criteria measures are (see [49]) weighted signal-to-distortion ratio (WSDR), weighted signal-to-interference ratio (WSIR), weighted signal-to-noise ratio (WSNR) and normalized RTF error (NRE). For this we denote the long-term averaged weighted power of noise and the various speaker components at the reference microphone and at the output by $\lambda_v^{\xi,\mathrm{A}}$ and $\left\{\lambda_{s,p}^{\xi,\mathrm{A}}\right\}_{p=1}^{P}$, respectively, where $\xi \in \{i, o\}$ stands for input reference microphone and output, respectively, and the superscript $(\cdot)^{\mathrm{A}}$ denotes the A-weighting [50]. The long-term weighted distortion between the p-th source component at the p-th output signal and its corresponding component at the reference microphone is denoted by $\lambda_{s,p}^{d,\mathrm{A}}$. Formally, the WSIR and WSNR criteria are defined for the p-th source as:

$$\mathrm{WSIR}_p^{\xi} \triangleq \frac{\lambda_{s,p}^{\xi,\mathrm{A}}}{\sum_{p' \neq p} \lambda_{s,p'}^{\xi,\mathrm{A}}} \tag{49a}$$

$$\mathrm{WSNR}_p^{\xi} \triangleq \frac{\lambda_{s,p}^{\xi,\mathrm{A}}}{\lambda_v^{\xi,\mathrm{A}}} \tag{49b}$$

where $p = 1, \ldots, P$ and $\xi \in \{i, o\}$ stands for input and output, respectively. The WSDR of the p-th output signal is defined as:

$$\mathrm{WSDR}_p \triangleq \frac{\lambda_{s,p}^{i,\mathrm{A}}}{\lambda_{s,p}^{d,\mathrm{A}}}. \tag{50}$$

The NRE criterion for assessing the performance of an RTF estimate $\hat{\boldsymbol{h}}_p(k)$ at the k-th frequency-bin is defined as:

$$\mathrm{NRE}_p(k) \triangleq \frac{\|\hat{\boldsymbol{h}}_p(k) - \boldsymbol{h}_p(k)\|^2}{\|\boldsymbol{h}_p(k)\|^2}. \tag{51}$$

[1] We would like to express our gratitude to D. Kounades-Bastian for sharing his code, extending the original MCNMF implementation by A. Ozerov and C. Févotte [48] from stereophonic input to any number of microphones.

Then, by averaging over all frequencies, a broad-band NRE is defined as:

$$\text{NRE}_p \triangleq \frac{1}{K}\sum_k \text{NRE}_p(k). \quad (52)$$

B. Simulated scenarios using speakers and IRs database

Various test-scenarios are simulated by convolving recorded speakers from the WSJCAM0 database [51] with RIRs drawn from a database [52]. We use a uniform linear array of $M \in \{4, 6, 8\}$ microphones with 8cm inter-microphone spacing.

The nominal testing scenario takes place in a reverberant room of size 6m $\times$ 6m $\times$ 2.4m with a reverberation time (RT) of 360ms and white Gaussian diffuse noise is generated using the simulator in [53]. Unless noted otherwise, speech signals of $P = 3$ sources with equal average power, arranged on a 1m radius circle, are received by a linear microphone array comprising $M = 6$ microphones, located at the center of the circle, with a WSNR of 14dB. The average DRR is 10dB. The signal duration is 1min, with all sources being simultaneously active. Only for the LCW algorithm, 1min time segments of each of the sources contaminated with noise are used for RTFs estimation. For each scenario, the performance measures are averaged over 10 Monte-Carlo experiments, in which the DOAs of the sources are randomly selected (with a minimum angle of 45^o between adjacent sources). The default number of iterations of the LST is 3. Unless stated otherwise, these parameters are used in all experiments.

We examine and compare the various algorithms versus the number of speakers and number of microphones in Sec. VII-B1. In Sec. VII-B2 we evaluate the performance of the various algorithms versus RT. For BSS algorithms, LGT and LST, we test the performance with different DOA initialization errors in Sec. VII-B3. Finally, we explore the convergence of the proposed algorithm LST over iterations in Sec. VII-B4.

1) Performance depending on number of speakers and microphones: We consider three scenarios in order to evaluate the performance of the algorithms versus number of speakers and microphones:

- $P = 2$ speakers and $M = 4$ microphones.
- $P = 3$ speakers and $M = 6$ microphones.
- $P = 4$ speakers and $M = 8$ microphones.

In practice, the more microphones are used, a more robust beamformer is obtained. For diffuse noises, this also improves the output SNR significantly.

The results are depicted in Table II. All algorithms obtain similar WSNR improvement. Regarding WSIR improvement, evidently the ideal LI outperforms all algorithms, and the supervised LCW performs almost as well. This can be attributed to the fact that the NRE of supervised algorithm LCW is fixed regardless of the number of sources, as it uses single-talk time segments for RTFs estimation. However, NRE of the proposed algorithms, LGT and LST, degrade as the number of sources increases. Yet, the proposed LST is consistently better than LGT, and obtains $13 - 15.7$dB improvement in WSIR.

2) Performance depending on RT: Considering the nominal scenario, we evaluate the algorithms' performance with different RT of 160ms, 360ms and 610ms and corresponding DRRs of 20dB, 10dB and 7dB, respectively. Results are depicted in Table III. Performance of all algorithms degrade as RT increases. This results from longer RIRs which require more parameters to estimate and due to violation of the multiplicative transfer function (MTF) assumption in the frequency domain. Supervised algorithms outperform unsupervised algorithms, however, the difference diminishes as RT increases. The proposed LST outperforms LGT with respect to WSIR improvement by $1.6 - 6.9$dB.

3) Sensitivity to DOA errors: Next, we evaluate the sensitivity of BSS algorithms to DOA errors of $0^o, 5^o, 10^o$ during initialization. Results are depicted in Table IV. Performance of both LST and LGT is robust to DOA errors of 5^o. For DOA errors of 10^o, WSIR improvement of LGT is degraded by 3.0dB, whereas for LST the degradation is only by 1.3dB.

4) Performance depending on number of iterations: Finally, we check the performance of the proposed LST algorithm with different numbers of iterations $I = 1, 2, \ldots, 5$, see results in Table V. Clearly, performance improves as number of iterations increase, however after 3 iterations the performance doest not increase much.

In all of the scenarios abovementioned, the performance of the proposed LST algorithm is comparable to the performance of the state-of-the-art MCNMF algorithm. The WSIR improvement is roughly

P	Alg.	WSNR [dB]		WSIR [dB]		WSDR	NRE
		Val.	Imp.	Val.	Imp.	[dB]	[dB]
2	In.	13.6	–	0.0	–	–	–
	LI	20.7	7.1	18.3	18.3	14.0	–
	LCW	20.7	7.1	18.5	18.5	13.9	−27.4
	LD	20.6	7.0	7.9	7.9	7.3	−7.3
	NMF	20.7	6.9	14.3	4.3	9.0	–
	LGT	20.5	6.9	14.5	14.5	11.2	−16.5
	LST	20.7	7.1	15.7	15.7	12.2	−18.7
3	In.	14.4	–	−3.6	–	–	–
	LI	22.6	8.2	14.0	17.7	13.1	–
	LCW	22.5	8.1	14.1	17.7	12.9	−26.4
	LD	22.6	8.2	4.1	7.1	6.0	−6.3
	NMF	21.1	6.6	11.5	15.1	7.9	–
	LGT	22.5	8.1	7.9	11.4	8.1	−10.1
	LST	22.8	8.4	10.8	14.4	10.8	−14.3
4	In.	13.8	–	−5.5	–	–	–
	LI	21.9	8.1	13.3	18.9	13.1	–
	LCW	21.9	8.1	13.3	18.8	13.1	−25.7
	LD	21.7	7.9	1.1	6.7	7.1	−5.6
	NMF	21.2	7.1	10.0	15.4	7.9	–
	LGT	21.6	7.8	4.5	10.6	6.2	−7.8
	LST	22.2	8.4	7.4	13.0	8.5	−11.0

TABLE II: Performance depending on the number of speakers where In., Val. and Imp. stand for Input, Value and Improvement, respectively.

RT	Alg.	WSNR [dB]		WSIR [dB]		WSDR	NRE
[ms]		Val.	Imp.	Val.	Imp.	[dB]	[dB]
160	In.	15.8	–	−3.7	–	–	–
	LI	22.3	9.0	27.5	31.1	25.9	–
	LCW	22.3	9.0	26.9	30.6	25.2	−37.3
	LD	22.4	9.1	7.8	11.5	6.8	−5.2
	NMF	20.1	6.8	17.9	21.5	13.7	–
	LGT	22.3	9.0	9.2	12.9	9.6	−9.2
	LST	22.6	9.3	16.2	19.8	16.9	−16.1
360	In.	14.2	–	−3.7	–	–	–
	LI	22.5	8.3	13.9	17.4	12.7	–
	LCW	22.5	8.2	14.0	17.5	12.7	−26.0
	LD	22.5	8.3	3.8	7.3	4.0	−4.2
	NMF	21.0	6.6	11.3	14.8	7.9	–
	LGT	22.5	8.3	7.7	11.3	8.0	−10.0
	LST	22.7	8.5	10.6	14.1	10.4	−13.8
610	In.	13.3	–	−3.4	–	–	–
	LI	23.3	10.0	8.0	11.4	7.7	–
	LCW	23.2	9.9	8.5	11.9	8.0	−17.2
	LD	23.3	10.0	2.3	5.7	1.7	−3.0
	NMF	22.1	6.3	8.0	11.5	5.6	–
	LGT	23.0	9.7	5.2	8.6	5.8	−9.9
	LST	23.4	10.1	6.8	10.2	7.1	−11.1

TABLE III: Performance depending on the reverberation time where In., Val. and Imp. stand for Input, Value and Improvement, respectively.

DOA err.	Alg.	WSNR [dB] Val.	WSNR [dB] Imp.	WSIR [dB] Val.	WSIR [dB] Imp.	WSDR [dB]	NRE [dB]
	In.	14.5	–	−3.6	–	–	–
0^o	LD	22.4	7.8	4.4	8.0	4.8	–
	LGT	22.2	7.6	7.5	11.1	7.8	−10.1
	LST	22.6	8.1	10.4	14.1	10.4	−14.1
5^o	LD	22.5	8.1	4.0	7.6	2.9	–
	LGT	22.7	8.3	7.0	10.6	6.9	−7.9
	LST	23.1	8.7	10.5	14.1	10.2	−12.7
10^o	LD	22.2	7.7	2.3	5.8	1.9	–
	LGT	21.9	7.4	4.5	8.1	5.0	−5.0
	LST	22.9	8.4	9.2	12.8	8.9	−9.8

TABLE IV: Performance depending on the DOA error where In., Val. and Imp. stand for Input, Value and Improvement, respectively.

Alg.	Iter.	WSNR [dB] Val.	WSNR [dB] Imp.	WSIR [dB] Val.	WSIR [dB] Imp.	WSDR [dB]	NRE [dB]
In.		14.4	–	−3.7	–	–	–
LST	1	19.4	5.1	9.2	12.9	9.0	−8.1
	2	19.6	5.2	10.8	14.5	10.8	−11.7
	3	20.1	5.7	11.5	15.2	11.5	−14.2
	4	19.4	5.0	11.8	15.5	11.7	−15.6
	5	19.8	5.4	11.6	15.2	11.1	−16.4

TABLE V: Performance depending on the number of iterations where In., Val. and Imp. stand for Input, Value and Improvement, respectively.

equal $\pm$ 1dB, the WSNR improvement is higher in the LST by 1.5dB and the WSDR is higher in the LST by 2dB.

C. Real-life recorded scenario

The algorithms are also tested in a real-life recording with live persons. The setup is comprised of a 16cm diameter circular array comprising $M = 8$ microphones positioned on top of a table at the center of a 6m $\times$ 6m $\times$ 2.4m room with a reverberation time of 0.3s. 3 persons, two male and one female, are sitting at a distance of 1m from the array at angles of $[0^o, 60^o, 120^o]$. 8 loud-speakers emitting statistically independent stationary noises with "speech like" spectrum are arranged at the 4 corners of the room and against its 4 walls. The loud-speakers are facing the walls to make the noise field more diffuse. The persons and noises are recorded separately and mixed offline according to the desired levels. Here, speech levels are equal and the noise level is 14dB weaker. The performance of the various algorithms is depicted in Table VI. It is evident from these results that the proposed algorithm performs well in a real-life scenario. The LST outperforms the MCNMF: by 2dB in WSNR; by 5dB in WSIR; and by 4dB in WSDR. The spectrograms of the first speaker at the input and at the output of the LST algorithm are depicted in Fig. 5. Clearly, the desired signal is enhanced significantly.

Alg.	WSNR [dB] Val.	WSNR [dB] Imp.	WSIR [dB] Val.	WSIR [dB] Imp.	WSDR [dB]
In.	14.7	–	−2.7	–	–
LI	25.0	10.3	21.0	23.7	19.7
LCW	22.7	11.1	20.0	22.7	18.7
LD	22.9	8.2	9.3	12.0	9.7
NMF	21.0	6.3	13.3	16.0	13.3
LGT	22.9	8.2	10.7	13.4	10.8
LST	23.0	8.3	18.3	21.0	17.3

TABLE VI: Performance comparison in a real-life scenario where In., Val. and Imp. stand for Input, Value and Improvement, respectively.

VIII. Conclusions

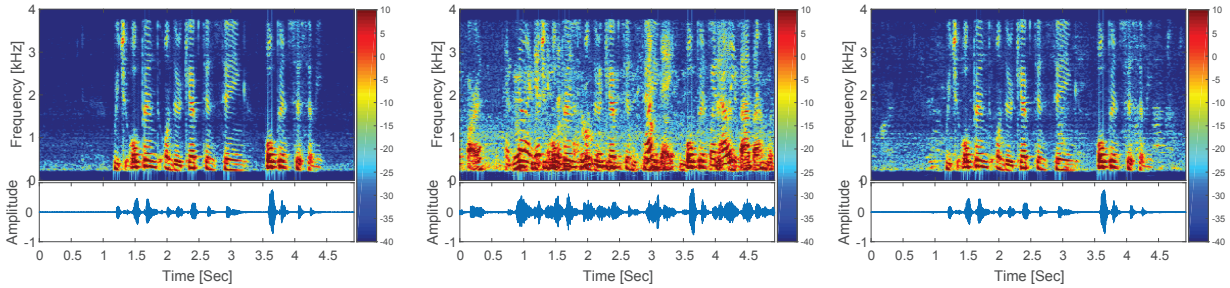
The problem of source separation using LCMV beamforming with supervised and unsupervised filter optimization is considered. Supervised algorithms which require single-talk time segments for estimating RTFs, of course, exhibit better performance than unsupervised. However, their application is restricted to scenarios where such time segments exist and reliable information on source activity patterns is available or can be estimated.

In this paper, we propose two methods in which the TRINICON framework is used for circumventing both requirements, and separate the sources given only a simple noise-only time segment detector and approximate information of the DOAs of the sources. The first proposed algorithm, denoted LCMV-TRINICON, estimates the RTFs of the sources using multiple two-channel GC-TRINICON units where one of the channels is a reference microphone. The estimated RTFs are used in constructing multiple LCMV-BFs for extracting each of the sources separately. In the second algorithm, similarly to the first algorithm, multiple two-channel GC-TRINICON units are used for estimating RTFs. However, instead of using one of the microphones as reference, we propose to use the outputs of the LCMV-BFs as SPRs. Assuming that SINR is improved at the output of the BFs, the error of new estimates of RTFs will be reduced. This procedure can be repeated iteratively where the separated signals at each iteration serve as the SPRs of the next iteration, thereby, continuously improving RTF estimates and the separation performance. For estimating the necessary covariance matrix of noise and interference for both algorithms, a simple detection of noise-only time segments is required.

An experimental study verifies the efficacy and applicability to real-life scenarios of the proposed algorithms. The performance of the proposed method is comparable to the state-of-the-art MCNMF algorithm. The evaluation relative to various algorithms which work under idealized conditions (such as using perfect activity knowledge and perfect noise knowledge by the algorithms ideal LCMV, DOA based LCMV, and LCMV-CW or as using training data by the MCNMF) showed that the proposed combined LCMV-SPR-TRINICON gets close to the behaviour of these algorithms. The performance of the proposed algorithm is close to the optimum while offering real-time capability and requiring no training, only coarse estimates of the sources' DOAs.

References

[1] S. L. Gay and J. Benesty, Eds., *Acoustic signal processing for telecommunication*. Kluwer, 2000.
[2] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.
[3] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.
[4] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech processing in modern communication: Challenges and perspectives*. Springer, 2010.
[5] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech Enhancement*, ser. Signals and Communication Technology. Berlin: Springer, 2005, pp. 199–228.
[6] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel Wiener filter for multiple sources scenarios," in *the 27th convention of the Israeli Chapter of IEEE*, Eilat, Israel, Nov. 2012.
[7] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.
[8] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.
[9] B. R. Breed and J. Strauss, "A short proof of the equivalence of LCMV and GSC beamforming," *IEEE Signal Processing Lett.*, vol. 9, no. 6, pp. 168–169, 2002.



(a) 1st source component at the reference microphone.

(b) reference microphone.

(c) 1st source enhanced.

Fig. 5: Spectrograms of the proposed combined LCMV-SPR-TRINICON algorithm in a real-life scenario.

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

[11] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech enhancement*. Springer, 2005, pp. 199–228.

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

[13] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 1072–1077, 2010.

[14] R. Serizel, M. Moonen, B. V. Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE Trans. Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[15] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[16] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 60, pp. 233–246, Jan. 2012.

[17] T. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[18] P. O'Grady, B. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, pp. 18–33, 2005.

[19] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind speech separation*. Springer, 2007.

[20] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "Convolutive blind source separation methods," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 1065–1094.

[21] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

[22] E. Vincent, M. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI Global, 2010, pp. 162–185.

[23] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[24] H. Sawada, R. Mukai, S. F. G. M. de la Kethulle de Ryhove, S. Araki, and S. Makino, "Spectral smoothing for frequency-domain blind source separation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2003, pp. 311–314.

[25] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive BSS of short mixtures by ICA recursively regularized across frequencies," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 624–639, 2011.

[26] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[27] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 715–726, 2007.

[28] 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 and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2005.

[29] ——, "Bind source separation algorithms for convolutive mixtures exploiting nongaussianity, nonwhiteness, and nonstationarity," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2003, pp. 275–278.

[30] ——, "Blind source separation for convolutive mixtures: A unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Kluwer Academic Pulishers, 2004, pp. 255–293.

[31] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Int. Workshop on Comp. Advances in Multi-Sensor Adapt. Proc. (CAMSAP)*, Aruba, Dutch Antilles, Dec. 2009, pp. 253–256.

[32] K. Reindl and W. Kellermann, "Linearly-constrained multichannel interference suppression algorithms derived from a minimum mutual information criterion," in *IEEE China Summit and Intl. Conf. on Signal and Information Proc. (ChinaSIP 2013)*, Jul. 2013.

[33] 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)*, New Paltz, USA, Oct. 2013.

[34] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2009, pp. 233–236.

[35] F. Nesta and M. Omologo, "Generalized state coherence transform for multidimensional tdoa estimation of multiple sources," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 246–260, 2012.

[36] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Processing*, vol. 86, pp. 1260–1277, June 2006.

[37] H. B. R. Aichner and W. Kellermann, "Exploiting narrowband efficiency for broadband convolutive blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–9, September 2006.

[38] C. Anderson, S. Meier, W. Kellermann, P. Teal, and M. Poletti, "A GPU-accelerated real-time implementation of TRINICON-BSS for multiple separation units," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014.

[39] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[40] S. Markovich-Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2010, pp. 201–204.

[41] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.
[42] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.
[43] M. Benard and R. Bronson, "Matrix methods: An introduction," 1970.
[44] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ICA-based blocking matrix for improved noise estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, 2014.
[45] R. Aichner, "Acoustic blind source separation in reverberant and noisy environments," Ph.D. dissertation, Friedrich-Alexander University, Erlangen-Nuremberg, Germany, 2007.
[46] S. Markovich-Golan, S. Gannot, and I. Cohen, "Low-complexity addition or removal of sensors/constraints in LCMV beamformers," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1205–1214, 2012.
[47] S. Gannot. (2016, Aug.) Audio samples for *Iterative combined LCMV-TRINICON* paper. [Online]. Available: http://www.eng.biu.ac.il/gannot/speech-enhancement/
[48] A. Ozerov and C. Févotte. (2010, Mar.) Multichannel nonnegative matrix factorization toolbox (in Matlab). [Online]. Available: http://www.irisa.fr/metiss/ozerov/Software/multi_nmf_toolbox.zip
[49] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Emerald, 2012.
[50] "Electroacoustics-sound level meters-part 1: Specifications (IEC 61672-1: 2002)," 2003.
[51] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 81–84.
[52] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*. IEEE, 2014, pp. 313–317.
[53] E. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

Shmulik Markovich-Golan Received the B.Sc. (Cum Laude) and M.Sc. degrees in electrical engineering from the Technion – Israel Institute of Technology, Haifa, Israel, in 2002 and 2008, respectively. He received the Ph.D. degree from Bar-Ilan University, Ramat Gan, Israel, in 2013.

Dr. Markovich-Golan is involved in a post-doctoral research under the supervision of Prof. Sharon Gannot at Bar-Ilan University, Israel, and under the supervision of Prof. Walter Kellermann at the University of Erlangen-Nuremberg, Germany. He joined Intel corporation in 2015 where he leads an audio algorithms development team. He is a co-recipient of two best paper awards.

His research interests include multi-channel speech processing using centralized or ad hoc distributed microphone arrays for voice communication and speech recognition.

Sharon Gannot (S'92-M'01-SM'06) received his B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, Haifa, Israel in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel in 1995 and 2000 respectively, all in Electrical Engineering. In 2001 he held a post-doctoral position at the department of Electrical Engineering (ESAT-SISTA) at K.U.Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is a Full Professor at the Faculty of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory and the Signal Processing Track.

Prof. Gannot is the recipient of Bar-Ilan University outstanding lecturer award for 2010 and 2014. He is also a co-recipient of seven best paper awards.

Prof. Gannot has served as an Associate Editor of the EURASIP Journal of Advances in Signal Processing in 2003-2012, and as an Editor of several special issues on Multi-microphone Speech Processing of the same journal. He has also served as a guest editor of ELSEVIER Speech Communication and Signal Processing journals. Prof. Gannot has served as an Associate Editor of IEEE Transactions on Speech, Audio and Language Processing in 2009-2013. Currently, he is a Senior Area Chair of the same journal. He also serves as a reviewer of many IEEE journals and conferences.

Prof. Gannot is a member of the Audio and Acoustic Signal Processing (AASP) technical committee of the IEEE since Jan., 2010. Currently, he serves as the committee vice-chair, and will become the TC chair in Jan, 2017. He is also a member of the Technical and Steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the general co-chair of IWAENC held at Tel-Aviv, Israel in August 2010. Prof. Gannot has served as the general co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in October 2013. Prof. Gannot was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013 and EUSIPCO 2013. Prof. Gannot research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation; dereverberation; single microphone speech enhancement and speaker localization and tracking.

Walter Kellermann is a professor for communications at the University of Erlangen-Nuremberg, Germany, since 1999. He received the Dipl.-Ing. (univ.) degree in Electrical Engineering from the University of Erlangen-Nuremberg, in 1983, and the Dr.-Ing. degree from the Technical University Darmstadt, Germany, in 1988. From 1989 to 1990, he was a postdoctoral Member of Technical Staff at AT&T Bell Laboratories, Murray Hill, NJ. In 1990, he joined Philips Kommunikations Industrie, Nuremberg, Germany, to work on hands-free communication in cars. From 1993 to 1999, he was a Professor at the Fachhochschule Regensburg, where he also became Director of the Institute of Applied Research in 1997. In 1999, he cofounded DSP Solutions, a consulting firm in digital signal processing, and he joined the University Erlangen-Nuremberg as a Professor and Head of the Audio Research Laboratory. He authored or coauthored 16 book chapters, 250+ refereed papers in journals and conference proceedings, as well as 50+ patents, and is a co-recipient of nine best paper awards. His current research interests include speech signal processing, array signal processing, adaptive filtering, and its applications to acoustic humanmachine interfaces. Dr. Kellermann served as an Associate Editor and Guest Editor to various journals, including the IEEE Transactions on Speech and Audio Processing from 2000 to 2004, the IEEE Signal Processing Magazine in 2015, and presently serves as Associate Editor to the EURASIP Journal on Applied Signal Processing. He was the General Chair of seven mostly IEEE-sponsored workshops and conferences. He served as a Distinguished Lecturer of the IEEE Signal Processing Society (SPS) from 2007 to 2008. He was the Chair of the IEEE SPS Technical Committee for Audio and Acoustic Signal Processing from 2008 to 2010, a Member of the IEEE James L. Flanagan Award Committee from 2011 to 2014, a Member of the SPS Board of Governors (2013-2015), and is currently Vice President Technical Directions of the IEEE Signal Processing Society (2016-2018). He was awarded the Julius von Haast Fellowship by the Royal Society of New Zealand in 2012 and the Group Technical Achievement Award of the European Association for Signal Processing (EURASIP) in 2015. In 2016, he was a Visiting Fellow at Australian National University, Canberra, Australia. He is an IEEE Fellow.