A consolidated perspective on multi-microphone speech enhancement and source separation

Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, Alexey Ozerov

A Consolidated Perspective on Multi-Microphone Speech Enhancement and Source Separation

Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov

***Abstract*—Speech enhancement and separation are core problems in audio signal processing, with commercial applications in devices as diverse as mobile phones, conference call systems, hands-free systems, or hearing aids. In addition, they are crucial pre-processing steps for noise-robust automatic speech and speaker recognition. Many devices now have two to eight microphones. The enhancement and separation capabilities offered by these multichannel interfaces are usually greater than those of single-channel interfaces. Research in speech enhancement and separation has followed two convergent paths, starting with microphone array processing and blind source separation, respectively. These communities are now strongly interrelated and routinely borrow ideas from each other. Yet, a comprehensive overview of the common foundations and the differences between these approaches is lacking at present. In this article, we propose to fill this gap by analyzing a large number of established and recent techniques according to four transverse axes: a) the acoustic impulse response model, b) the spatial filter design criterion, c) the parameter estimation algorithm, and d) optional postfiltering. We conclude this overview paper by providing a list of software and data resources and by discussing perspectives and future trends in the field.**

***Index Terms*—Multichannel, array processing, beamforming, Wiener filter, independent component analysis, sparse component analysis, expectation-maximization, postfiltering.**

I. Introduction

SPEECH enhancement and separation are core problems in audio signal processing. Real-world speech signals often involve one or more of the following distortions: reverberation, interfering speakers, and/or noise. In this context, source separation refers to the problem of extracting one or more target speakers and cancelling interfering speakers and/or noise. Speech enhancement is more general, in that it refers to the problem of extracting one or more target speakers and cancelling one or more of these three types of distortion. If one focuses on removing interfering speakers and noise, as opposed to reverberation, the terms of "signal enhancement" and "source separation" become essentially interchangeable. These problems arise in various real scenarios. For instance, spoken communication over mobile phones or hands-free systems requires the enhancement or separation of the near-end speaker's voice with respect to interfering speakers and environmental noises before it is transmitted to the far-end listener. Conference call systems or hearing aids face the same problem, except that several speakers may be considered as targets. Speech enhancement and separation are also crucial pre-processing steps for robust automatic speech recognition and understanding, as available in today's personal assistants, GPS, televisions, video game consoles, and medical dictation devices. More generally, they are believed to be necessary to provide humanoid robots, assistive devices, and surveillance systems with machine audition capabilities. While the above applications require real-time processing, off-line separation of singing voice, drums, and other musical instruments has been successfully used for music information retrieval, upmixing of mono or stereo movie soundtracks to 3D sound formats, and remixing of music recordings. Other applications, e.g. meeting transcription, can be also processed off-line.

With few exceptions such as speech codecs and old sound archives, the input signals are *multichannel*. The number of microphones per device has steadily increased in the last few years. Most smartphones, tablets and in-car hands-free systems are now equipped with two or three microphones. Hearing aids typically feature two microphones per ear and a wireless link [1] to enable communication between the left and right hearing aids, and conference call systems with eight microphones are commercially available. Research prototypes with forty to hundreds of microphones have been demonstrated in lecture halls, office and domestic environments [2]–[6]. The enhancement capabilities offered by these multichannel interfaces are usually greater than those of single-channel interfaces. They make it possible to design *multichannel spatial filters* that selectively enhance or suppress sounds in certain directions (or volumes) by exploiting the spatial diversity, e.g. phase and level differences, or more generally, the different acoustic properties between channels. *Single-channel spectral filters*, in contrast, require much more detailed knowledge about the target and the noise and they usually result in smaller quality improvement. As a matter of fact, it can be shown that the maximum quality improvement theoretically achievable with only two microphones is already much greater than with a single microphone and that it keeps increasing with more microphones [7].

Hundreds of multichannel audio signal enhancement techniques have been proposed in the literature over the last forty years along two historical research paths. *Microphone array processing* emerged from the theory of sensor array processing for telecommunications and it focused mostly on the localization and enhancement of speech in noisy or reverberant environments [8]–[12], while *blind source separation* (BSS) was later popularized by the machine learning community and it addressed "cocktail party" scenarios involving several sound sources mixed together [13]–[18]. These two research

S. Gannot and S. Markovich-Golan are with Bar-Ilan University, Ramat-Gan 5290002, Israel (email: gannot@eng.biu.ac.il, shmuel.markovich@biu.ac.il). E. Vincent is with Inria, 54600 Villers-lès-Nancy, France (e-mail: emmanuel.vincent@inria.fr). A. Ozerov is with Technicolor R&D, 35576 Cesson Sévigné, France (email: alexey.ozerov@technicolor.com).

tracks have converged in the last decade and they are hardly distinguishable today. As will be shown in this overview paper, source separation techniques are not necessarily blind anymore and most of them exploit the same theoretical tools, impulse response models and spatial filtering principles as speech enhancement techniques.

Despite this convergence, most books and reviews have focused on either of these tracks. This article intends to fill this gap by providing a comprehensive overview of their common foundations and their differences. The vastness of the topic requires us to limit the scope of this overview to the following:

- we focus on multichannel recordings made by multiple microphones, as opposed to multichannel signals created by mixing software which do not match the acoustics of real environments;
- we mostly study the enhancement and separation of speech with respect to interfering speech sources and environmental noise in reverberant environments, as opposed to cancelling echoes and reverberation of the target speech;
- we concentrate on truly multichannel techniques based on acoustic impulse response models and multichannel filtering: as such, we only briefly introduce speech and noise models, computational auditory scene analysis (CASA) models, and time-frequency masking techniques used to assist multichannel processing, but do not describe their use for single-channel or channel-wise filtering in depth;
- we do not describe possible use of the enhanced signals for subsequent tasks;
- time difference of arrival (TDOA) estimation and speaker localization of (multiple) sound sources are beyond the scope of this paper.

Readers interested in multichannel signals created by professional mixing software and in the use of source separation as a prior step to audio upmixing and remixing may refer to, e.g., [19]–[21]. Echo cancellation, dereverberation, and CASA are major topics described in the books [22]–[25]. For more information about advanced spectral models and their use for single-channel and channel-wise spectral filtering, see, e.g., [18], [26], [27]. For the use of speech enhancement and musical instrument separation as pre-processing steps for speech recognition and music information retrieval, see, e.g., [28]–[31]. For a survey of TDOA and location estimation techniques, interested readers may refer to [32]–[34].

In spite of its limited scope, this overview still covers a wide field of research. In order to classify existing techniques irrespectively of their origin in microphone array processing or BSS, we adopt four transverse axes: a) the acoustic impulse response model, b) the spatial filter design criterion, c) the parameter estimation algorithm, and d) optional postfiltering. These four modeling and processing steps are common to all techniques, as illustrated in Fig. 1. The structure of the article is as follows. We recall useful elements of acoustics and introduce general notations in Section II. After describing various acoustic impulse response models in Section III, we define the fundamental concepts of spatial filtering in Section IV and review existing design criteria, estimation algorithms, and postfiltering techniques in Sections V, VI, and VII, respectively. We provide a list of resources in Section VIII and conclude in Section IX by summarizing the similarities and the differences between approaches originating from microphone array processing and BSS and discussing perspectives in the field.

II. Elements of acoustics — Notations

From now on, we assume that two or more sound *sources* are simultaneously recorded by two or more microphones. The microphones are assumed to be *omnidirectional*, unless explicitly stated otherwise. The set of microphones is called a microphone *array*. Each recorded signal is called a *channel* and the set of recorded signals is the array *input* signal or the *mixture* signal.

A. Physics

Sound is a variation of air pressure on the order of 10^{-2} Pa for a speech source at a distance of 1 m, on top of the average atmospheric pressure of 10^5 Pa. For such pressure values, the wave equation that governs the propagation of sound in air is linear [35]. This has two implications:

1) the pressure field at any time is the sum of the pressure fields resulting from each source at that time;
2) the pressure field emitted at a given source propagates over space and time according to a linear operation.

Unless clipping occurs, microphones operate linearly to record the pressure value at given point in space. If one considers the pressure field emitted by each source as the target[1], the overall phenomenon is therefore linear.

In the free field, the solution to the wave equation is given by the *spherical wave* model. The waveform $x_i(\tilde{t})$ recorded at point i when emitting a waveform $s_j(\tilde{t})$ at point j is equal to

$$x_i(\tilde{t}) = \frac{1}{\sqrt{4\pi} q_{ij}} s_j\left(\tilde{t} - \frac{q_{ij}}{c}\right) \tag{1}$$

with $\tilde{t}$ denoting continuous time, q_{ij} the distance between points i and j, and c the speed of sound, that is 343 m/s at 20°C. This speed is very small compared to the speed of light, so that propagation delays are not negligible. The recorded waveform differs from the emitted waveform by a delay q_{ij}/c and an attenuation factor of $1/\sqrt{4\pi} q_{ij}$.

In the presence of obstacles, the sound wave is affected in different ways depending on its frequency ν. The *wavelength* $\lambda = c/\nu$ of audio varies from 17 mm at $\nu = 20$ kHz to 17 m at $\nu = 20$ Hz.

When the sound wave hits an object of dimension smaller than λ, it is not affected. When it hits an obstacle of comparable dimension to λ, it is subject to *diffraction*. The wavefront is bended in a way that depends on the shape of the obstacle, its material and the angle of incidence. Roughly speaking, it will take more time for the wave to pass the obstacle and it will be more attenuated than in air. This phenomenon occurs

[1]Loudspeakers and musical instruments such as the trumpet do not operate linearly. These nonlinearities occur within solid parts of the loudspeaker or the instrument, however, before vibration is transmitted to air.

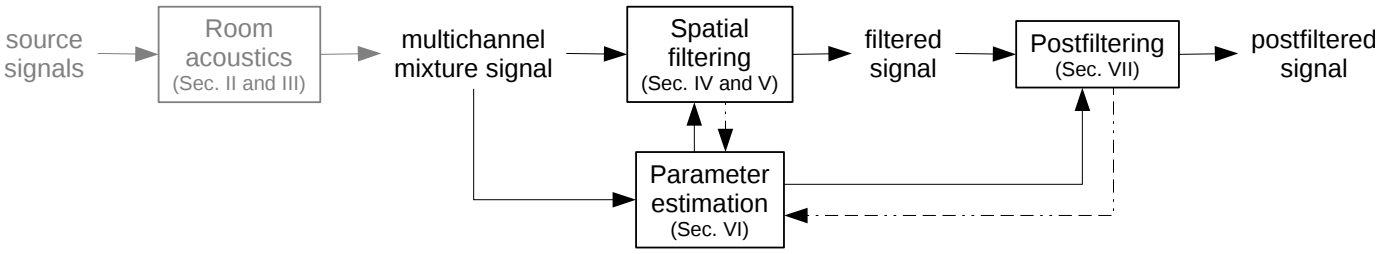


Figure 1. General schema showing acoustical propagation (gray) and the processing steps behind speech enhancement and source separation (black). Plain arrows indicate the processing order common to all algorithms and dashed arrows the feedback loops for certain algorithms.

most notably for hearing aid users, whose torso, head, and pinna, act as obstacles [36]. It also explains *source directivity*, i.e. the fact that the sound emitted by a source depends on direction.

When the wave hits a large rigid surface of dimension larger than λ, it is subject to *reflection*. The direction of the reflected wave is symmetrical to the direction of the incident wave with respect to the surface normal. Only part of the wave power is reflected: the rest is absorbed by the surface. The absorption ratio depends on the material and the angle of incidence [37]. It is on the order of 1% for a tiled floor, 7% for a concrete wall, and 15% for a carpeted floor.

Due to these small values, many successive wave reflections typically occur before the power becomes negligible. This induces multiple propagation paths between each source and each microphone, each with a different delay and attenuation factor. The waves corresponding to different paths are coherent and may result in constructive or destructive interference.

B. Deterministic perspective

Let us now move from the physical domain to discrete time signal processing. We assume that the recorded sound scene consists of J sources and that the number of microphones is equal to I. We adopt the following general notations: scalars are represented by plain letters, vectors by bold lowercase letters, and matrices by bold uppercase letters. The source index, the microphone index, and the time index are denoted by i, j, and t, respectively. The operator ^{T} refers to matrix transposition, and ^{H} to Hermitian transposition.

According to the first linearity assumption in Section II-A, the multichannel mixture signal $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$ can be expressed as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{2}$$

where $\mathbf{c}_j(t) = [c_{1j}(t), \ldots, c_{Ij}(t)]^T$ is the *spatial image* [38] of source j, that is the contribution of that source to the sound recorded at the microphones. This formulation is very general: it applies both to targets and noise, and multiple noise sounds can be modeled either as multiple sources or as a single source [39]. In particular, it is valid for spatially *diffuse* sources such as wind, trucks, or large musical instruments, which emit sound in a large region of space.

In the case of a *point source*, the second linearity assumption makes it possible to express $\mathbf{c}_j(t)$ by linear convolution of a single-channel *source signal* $s_j(t)$ and the vector $\mathbf{a}_j(t,\tau) = [a_{1j}(t,\tau), \ldots, a_{Ij}(t,\tau)]^T$ of *acoustic impulse responses* (AIRs) from the source to the microphones:

$$\mathbf{c}_j(t) = \sum_{\tau=0}^{\infty} \mathbf{a}_j(t,\tau) s_j(t-\tau) \tag{3}$$

This expression only holds for sources such as human speakers which emit sound in a tight region of space. The AIRs result from the summation of the multiple propagation paths and they vary over time due to movements of the source, of the microphones, or of other objects in the environment. When such movements are small, they can be approximated as time-invariant and denoted as $\mathbf{a}_j(\tau)$.

A schematic illustration of the shape of an AIR is provided in Fig. 2. It consists of three successive parts. The first peak is the *direct path* from the source to the microphone, as modeled in (1). It is followed by *early echoes* corresponding to the first few reflections on the room boundaries and the furniture. Subsequent reflections cannot be distinguished from each other anymore and they form an exponentially decreasing tail called *reverberation*. This overall shape is often described by two quantities: the *reverberation time* (RT), that is the time it takes for the reverberant tail to decay by 60 decibels (dB), and the *direct-to-reverberant ratio* (DRR), that is ratio of the power of direct sound (i.e., direct path) to that of the rest of the AIR. The RT depends solely on the room, while the DRR also depends on the source-to-microphone distance. The RT is virtually equal to 0 in outdoor conditions due to the absence of reflection and it is on the order of 50 ms in a car [40], 0.2 to 0.8 s in office or domestic conditions, 0.4 s to 1 s in a classroom, and 1 s or more in an auditorium [41].

Fig. 3 depicts a real AIR measured in a meeting room. It has both positive and negative values and it exhibits a strong first reflection on a table just after the direct path, but its magnitude follows the overall shape in Fig. 2.

C. Statistical perspective

Besides the above deterministic characterization of AIRs, it is useful to adopt a statistical point of view [35], [42]. To do so, we decompose AIRs as

$$a_{ij}(\tau) = e_{ij}(\tau) + r_{ij}(\tau) \tag{4}$$

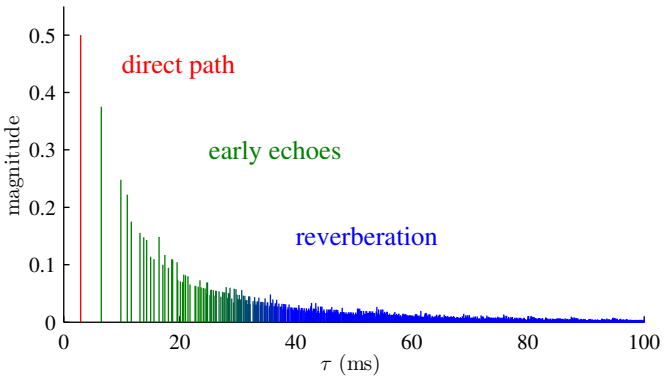


Figure 2. Schematic illustration of the shape of an AIR for a reverberation time of 0.25 s (from [18]).

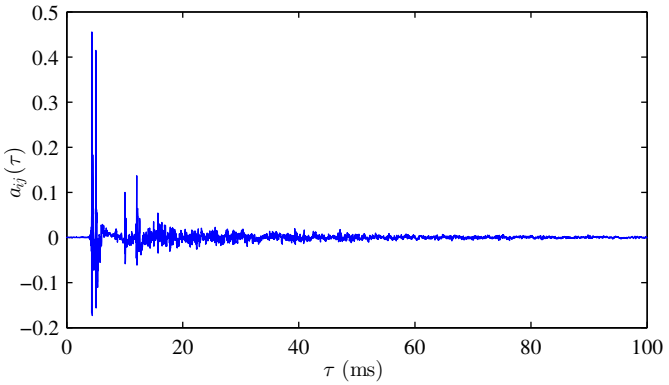


Figure 3. First 0.1 s of a real AIR from the Aachen Impulse Response Database [41] recorded in a meeting room with a reverberation time of 0.23 s with a source-to-microphone distance of 1.45 m.

where $e_{ij}(\tau)$ models the direct path and early echoes and $r_{ij}(\tau)$ models reverberation.

The fact that reverberation results from the superposition of thousands to millions of acoustic paths makes it follow the law of large numbers. This implies three useful properties. Firstly, $r_{ij}(\tau)$ can be modeled as a zero-mean Gaussian noise signal whose amplitude decays exponentially over time according to the room's RT [43]. Secondly, the covariance $\mathbb{E}(r_{ij}(\nu)r_{ij}^*(\nu'))$ between its Fourier transform $r_{ij}(\nu)$ at two different frequencies ν and ν' decays quickly with the difference between ν and ν' [44], [45]. Thirdly, if the room's RT is large enough, the reverberant sound field is diffuse, homogenous and isotropic, which means that it has equal power in all directions of space. This last property makes it possible to compute the normalized correlation between two different channels i and i' in closed-form as [35], [45], [46]

$$\Omega_{ii'}(\nu) = \frac{\mathbb{E}^{\text{spat}}(r_{ij}(\nu)r_{i'j}^*(\nu))}{\sqrt{\mathbb{E}^{\text{spat}}(|r_{ij}(\nu)|^2)}\sqrt{\mathbb{E}^{\text{spat}}(|r_{i'j}(\nu)|^2)}} = \frac{\sin(2\pi\nu\ell_{ii'}/c)}{2\pi\nu\ell_{ii'}/c} \quad (5)$$

where $\mathbb{E}^{\text{spat}}$ denotes *spatial expectation* over all possible absolute positions of the sources and the microphone array in the room, and $\ell_{ii'}$ the distance between the microphones. Note that the result does not depend on j anymore. This quantity known as the *interchannel coherence* is shown in Fig. 4. It is large for small arrays and low frequencies and it increases with microphone distance and frequency. We can further define the $I \times I$ coherence matrix of the diffuse sound field by concatenating all elements from (5) as $(\mathbf{\Omega}(\nu))_{ii'} = \Omega_{ii'}(\nu)$. It is interesting to note that both deterministic and statistical perspectives are valid. The appropriate choice depends on the observation length, and both perspectives can be useful in accomplishing different tasks [47]. We will elaborate on this issue in the subsequent section.

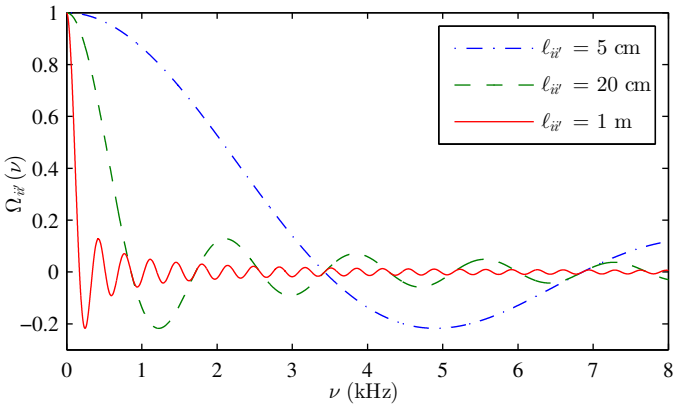


Figure 4. Interchannel coherence $\Omega_{ii'}(\nu)$ of the reverberant part of an AIR as a function of microphone distance $\ell_{ii'}$ and frequency ν.

III. Acoustic Impulse Response Models

The above properties of AIRs can be modeled and exploited to design enhancement techniques. Five categories of models have been proposed in the literature. A model is defined by a parameterization of the AIRs and possible prior knowledge about the parameter values. This prior knowledge can take the form of deterministic constraints, penalty terms which we shall denote by $\mathcal{P}(.)$, or probabilistic priors which we shall denote by $p(.)$.

A. Time-domain models

The simplest approach is to consider the AIRs as finite impulse response (FIR) filters modeled by their time-domain coefficients $\mathbf{a}_j(t,\tau)$ or $\mathbf{a}_j(\tau)$, $\tau \in \{0,\ldots,L-1\}$. The assumed length L is generally on the order of several hundred to a few thousand taps. This model was very popular in the early stages of research [48]–[55]. Recently, interest has revived with *sparse* penalties which account for prior knowledge about the physical properties of AIRs, namely the facts that power concentrates in the direct path and the first early echoes [56]–[60] and that the time envelope decays exponentially [61], but these penalties have not yet been used in a BSS context.

Time-domain modeling of AIRs exhibits several limitations. Firstly, prior knowledge about the spatial position of the sources does not easily translate into constraints on the AIR coefficients [62]. Secondly, the source signals are typically

modeled in the time-frequency domain instead, which forces estimation algorithms to alternate between one domain and the other [63]. Finally, the large number of parameters involved translates into large computational cost [64].

B. Narrowband approximation

To address these limitations, the convolution in the time domain can be approximated by a multiplication in the short-time Fourier transform (STFT) domain [65], provided that the frame length is sufficiently large. Let us denote by $\mathbf{c}_j(n,f)$ and $s_j(n,f)$ the STFT of $\mathbf{c}_j(t)$ and $s_j(t)$, respectively, with $n \in \{1,\dots,N\}$ the frame index, $f \in \{0,\dots,F-1\}$ the discrete frequency bin, N the number of time frames and F the discrete Fourier transform (DFT) length. The so-called *narrowband approximation* [66]–[72] is given by

$$\mathbf{c}_j(n,f) = \mathbf{a}_j(n,f)s_j(n,f) \tag{6}$$

where $\mathbf{a}_j(n,f) = [a_{1j}(n,f),\dots,a_{Ij}(n,f)]^T$ is the concatenation of the acoustic transfer functions (ATFs) from source j to the microphones. The appropriate frame length with respect to the length of the AIR is determined in [7], [73]. The ATFs can be either time-varying or time-invariant. In the former case, they can be represented via a dynamical model [74]. In the latter case, they simplify to $\mathbf{a}_j(n,f) = \mathbf{a}_j(f)$.

The narrowband approximation significantly simplifies estimation algorithms, since the decoupling between frequencies reduces the dimension of the problem. However, it may raise other problems, most notably *gain ambiguity* and *permutation ambiguity*. These ambiguities can be mitigated by smoothing between nearby frequencies [75], [76] or by introducing geometrical (soft or hard) constraints [70], [77], [78]. Interestingly, the latter references demonstrate the common foundations of microphone array and BSS methods for separating speech sources in reverberant environments.

These constraints are based on the fact that, in the absence of echoes and reverberation, the vector of ATFs simplifies to the *steering vector*, that is the DFT of (1):

$$\mathbf{d}_j(f) = \left[\frac{1}{\sqrt{4\pi}q_{1j}}e^{-2\jmath\pi q_{1j}\nu_f/c},\dots,\frac{1}{\sqrt{4\pi}q_{Ij}}e^{-2\jmath\pi q_{Ij}\nu_f/c}\right]^T \tag{7}$$

with $\jmath = \sqrt{-1}$, $\nu_f = f \times f_s/F$ the continuous frequency (in Hz) corresponding to frequency bin $f \in \{0,\dots,F/2\}$, and f_s the sampling frequency. A case of practical interest is the so-called *far-field* case, when the source-to-microphone distances q_{ij} are large compared to the inter-microphone distances $\ell_{ii'}$. The attenuation factors $1/\sqrt{4\pi}q_{ij}$ are then considered as equal, and the steering vector further simplifies (up to this factor) to

$$\mathbf{d}_j(f) = \left[e^{-2\jmath\pi q_{1j}\nu_f/c},\dots,e^{-2\jmath\pi q_{Ij}\nu_f/c}\right]^T. \tag{8}$$

An explicit model of early echoes was also recently proposed in [79].

C. Relative transfer function and interchannel models

An alternative approach to handle the gain ambiguity is to consider the *relative transfer function* (RTF) between channels for a given source. Taking the first microphone as a reference, the vector of RTFs $\tilde{\mathbf{a}}_j(f) = [\tilde{a}_{1j}(f),\dots,\tilde{a}_{Ij}(f)]^T$ for source j is defined as [69]

$$\tilde{\mathbf{a}}_j(f) \triangleq \frac{1}{a_{1j}(f)}\mathbf{a}_j(f). \tag{9}$$

A variant of this representation is to normalize the amplitude and the phase of the ATFs as [80], [81]

$$\bar{\mathbf{a}}_j(f) = \frac{e^{-\jmath\angle a_{1j}(f)}}{\|\mathbf{a}_j(f)\|}\mathbf{a}_j(f). \tag{10}$$

The amplitude normalization in (10), which was also considered in [66], is more robust than in (9) since the normalization factor depends on all channels. The phase normalization remains sensitive to the choice of the reference microphone, though. For a soft selection of the reference channel please refer to [82]. For generalizations of the RTF, see [83], [84].

The RTF encodes the interchannel level difference (ILD), also known as the interchannel intensity difference (IID), in decibels and the interchannel time difference (ITD) in seconds at each frequency [85]:

$$\mathrm{ILD}_{ij}(f) = 20\log_{10}|\tilde{a}_{ij}(f)| \tag{11}$$

$$\mathrm{ITD}_{ij}(f) = \frac{\angle\tilde{a}_{ij}(f)}{2\pi\nu_f} \tag{12}$$

where $\angle$ denotes the phase in radians of a complex number. The ITD is unambiguously defined only below the frequency c/ℓ_{i1}, known as the *spatial aliasing frequency*, with ℓ_{i1} the distance between microphones i and 1. With a sampling rate of 16 kHz, this corresponds to a sensor spacing of less than 4.3 cm. Above that frequency, the phase difference becomes larger than 2π and the ITD can be measured only up to an integer multiple of $1/\nu_f$. For that reason, the interchannel phase difference (IPD)

$$\mathrm{IPD}_{ij}(f) = \angle\tilde{a}_{ij}(f) \tag{13}$$

is often considered instead.

This model is popular for channel-wise filtering in the context of CASA, where the ILD and ITD are called interaural level and intensity differences, respectively, and are influenced by the shape of the pinna, the head and the torso [36]. It has however been used for multichannel filtering too [71], [85]–[87]. The use of level and phase differences retains the information about the source positions while discarding absolute levels and phases which are considered as irrelevant. Indeed, in the absence of echoes and reverberation, the ITD at all frequencies becomes equal to the TDOA $(q_{ij}-q_{1j})/c$, the vector of RTFs becomes equal to the relative steering vector

$$\tilde{\mathbf{d}}_j(f) = \left[1, e^{-2\jmath\pi(q_{2j}-q_{1j})\nu_f/c},\dots,e^{-2\jmath\pi(q_{Ij}-q_{1j})\nu_f/c}\right]^T, \tag{14}$$

and the normalized vector of ATFs $\bar{\mathbf{a}}_j(f)$ becomes equal to the steering vector $\bar{\mathbf{d}}_j(f)$ normalized as in (10). This has been exploited to constrain $\tilde{\mathbf{a}}_j(f)$ in anechoic conditions [88], [89] and to derive penalties over $\tilde{\mathbf{a}}_j(f)$ [85], [90] or $\bar{\mathbf{a}}_j(f)$ [81] in

reverberant conditions, such as

$$\mathcal{P}(\mathbf{a}_j|\mathbf{d}_j) = \sum_{f=0}^{F-1} \|\bar{\mathbf{a}}_j(f) - \bar{\mathbf{d}}_j(f)\|. \quad (15)$$

It should be noted however that such penalties do not match the actual distribution of ILD and IPD in the presence of echoes and reverberation [18, Fig. 2]. The preservation of interaural quantities is especially important in hearing aids, in order to increase speech intelligibility [91] and preserve the spatial awareness of the wearer. For penalties specifically designed for this application area, see [92]–[99].

D. Inter-frame and inter-frequency models

As mentioned above, the narrowband approximation holds only when the frame length is sufficiently long. Time-domain filtering can be exactly implemented in the frequency domain using *overlap and save* techniques [69], [100], provided that the analysis frame-length is larger than the filter length. However, this framework necessitates rectangular windows of different length in the analysis and synthesis stages. This might limit the performance of the separation algorithms, especially in dynamic scenarios.

In the conventional STFT domain [65], [101], [102], it can be shown that time-domain convolution by time-invariant AIRs translates into inter-frame and inter-band convolution [103], [104]:

$$\mathbf{c}_j(n,f) = \sum_{f'=0}^{F-1} \sum_{n'} \mathbf{a}_j(n',f',f) s_j(n-n',f'). \quad (16)$$

Since this expression involves multiple filtering operations, it is beneficial to consider the subband filtering approximation:

$$\mathbf{c}_j(n,f) = \sum_{n'} \mathbf{a}_j(n',f) s_j(n-n',f) \quad (17)$$

which was used to derive speech enhancement and separation algorithms in [105], [106]. Suitable DFT zero-padding makes it equivalent to time-domain filtering [107]–[109], however it introduces a coupling between frequencies. These models have been little used in practice, due to the potentially large number of STFT domain filter coefficients to be estimated.

E. Full-rank covariance model

An alternative approach which partly overcomes the limitations of the narrowband approximation is to model the second-order statistics of the ATFs. Let us consider the narrowband approximation (6) and assume that the source STFT coefficients $s_j(n,f)$ have a zero-mean nonstationary Gaussian distribution with variance $\sigma^2_{s_j}(n,f)$, and they are all independent source-, frame- and frequency-wise (i.e., over j, n and f). Under this *local Gaussian model* (LGM) [110]–[112], it can be shown that the source spatial images $\mathbf{c}_j(n,f)$ follow a zero-mean multivariate nonstationary Gaussian distribution

$$p(\mathbf{c}_j(n,f)|\mathbf{\Sigma}_{\mathbf{c}_j}(n,f)) = \frac{e^{-\mathbf{c}_j^H(n,f)\mathbf{\Sigma}_{\mathbf{c}_j}^{-1}(n,f)\mathbf{c}_j(n,f)}}{|\pi\mathbf{\Sigma}_{\mathbf{c}_j}(n,f)|} \quad (18)$$

with covariance matrix

$$\mathbf{\Sigma}_{\mathbf{c}_j}(n,f) = \sigma^2_{s_j}(n,f)\mathbf{R}_j(n,f) \quad (19)$$

where $\mathbf{R}_j(n,f)$ is the so-called *spatial covariance matrix* [113]. Under the narrowband approximation, $\mathbf{R}_j(n,f) = \mathbf{a}_j(n,f)\mathbf{a}_j^H(n,f)$ is constrained to be a rank-1 matrix. This implies that the channels of $\mathbf{c}_j(n,f)$ are *coherent*, i.e. perfectly correlated.

It was proposed in [113] to relax this constraint and to consider an unconstrained, *full-rank* spatial covariance matrix $\mathbf{R}_j(n,f)$ within the LGM instead. This more flexible formulation is applicable to diffuse sources or reverberated sources whose AIRs are longer than the frame length. In such cases, the sound field spans several directions at each frequency, such that the channels of $\mathbf{c}_j(n,f)$ become incoherent. The diagonal entries of $\mathbf{R}_j(n,f)$ encode the ILD and its off-diagonal entries encode the IPD and the interchannel coherence (IC), that is the correlation between channels.

The spatial covariance can be either time-varying or time-invariant. In the former case, it can be represented via a dynamical model [114]. In the latter case, it simplifies to $\mathbf{R}_j(n,f) = \mathbf{R}_j(f)$.

Due to the increased number of parameters, the estimation of this model is more difficult, especially when the number of microphones I is large. To overcome this difficulty, several approaches have proposed to constrain the full-rank model based on physical AIR characteristics, microphone array geometry, and/or presumed source positions. These constraints are incorporated either via deterministic constraints [113], [115] or probabilistic prior distributions [116] on the model parameters. In [115], $\mathbf{R}_j(f)$ is represented as the weighted sum of rank-1 kernels modeling individual uniformly distributed directions. In [116], the following inverse-Wishart prior is set on $\mathbf{R}_j(f)$ instead:

$$p(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f),m) = \frac{|\mathbf{\Psi}_j(f)|^m |\mathbf{R}_j(f)|^{-(m+I)} e^{-\mathrm{tr}\left[\mathbf{\Psi}_j(f)\mathbf{R}_j^{-1}(f)\right]}}{\pi^{I(I-1)/2}\prod_{i=1}^{I}\Gamma(m-i+1)} \quad (20)$$

where $\Gamma(\cdot)$ is the gamma function, m is the number of degrees of freedom, and $\mathbf{\Psi}_j(f) = (m-I)\bar{\mathbf{R}}_j(f)$ is the inverse scale matrix. Under certain assumptions, the mean value $\bar{\mathbf{R}}_j(f)$ of this distribution can be defined as

$$\bar{\mathbf{R}}_j(f) = \mathbf{d}_j(f)\mathbf{d}_j^H(f) + \sigma^2_{\mathrm{rev}}\mathbf{\Omega}(f) \quad (21)$$

where $\mathbf{d}_j(f)$ is the steering vector in (7), $\mathbf{\Omega}(f)$ is the covariance matrix of a diffuse sound field whose entries $\Omega_{ii'}(\nu_f)$ are given in (5), and σ^2_{rev} is the power of early echoes and reverberation [113]. It was shown that, when the RT is moderate or large, the variance of this distribution is small so that $\mathbf{R}_j(f)$ is similar to $\bar{\mathbf{R}}_j(f)$.

F. Discussion

In summary, various AIR models can be derived from the deterministic and statistical perspectives laid in Sections II-B and II-C, respectively, depending on the frame length. Long frame lengths yield the deterministic narrowband or rank-1

model. As we shall see, up to $I-1$ directional noise sources can then be perfectly eliminated in theory by narrowband spatial filtering. However, the large number of frequency bands and the small number of observed time frames make it difficult to estimate the appropriate filter in practice. Shorter frame lengths result in the statistical full-rank spatial covariance model instead. The amount of directional noise cancellation is then limited. However, this allows for low-latency processing and increases the number of frames available for the estimation of the spatial filter (see early discussion on the influence of the frame length on the coherence [117]). These two perspectives hence complement each other. Actually, they were both adopted for deriving a joint noise reduction and dereverberation algorithm in [47].

IV. Spatial Filtering

In this section we explore some fundamental concepts of array processing. Unless otherwise stated, these definitions are applicable to all arrays (not necessarily microphone arrays). For a comprehensive review on arrays (not specifically for speech applications), the reader is referred to [118].

A. Array Preliminaries

1) Beamformer: Assume that the far-field assumption (8) holds. A *linear spatial filter* is defined by a frequency-dependent vector $\mathbf{w}(f) = [w_1(f), \dots, w_I(f)]^T$ comprising one complex-valued weight per microphone, that is applied to the STFT $\mathbf{x}(n,f)$ of the array input signal $\mathbf{x}(t)$. Its output is equal to $\mathbf{w}^H(f)\,\mathbf{x}(n,f)$ and it can be transformed back into the time-domain by inverse STFT.

Such a filter is called a *beamformer*. The term beamformer originally referred to direction of arrival (DOA) based filters and was later generalized to all linear spatial filters. We will see in Sections V-E and VII that there also exist nonlinear spatial filters, which we will simply call "spatial filters".

2) Beampattern: In the rest of this section, we omit indexes j and f for legibility. Define a spherical coordinate system, with θ the elevation angle measured from the positive z-axis, and ϕ is the azimuth angle:

$$\mathbf{k} = [\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta)]^T. \quad (22)$$

The radius is irrelevant for defining the classical far-field beampattern.

In order to understand the impact of a beamformer on sound sources impinging from different directions, let us consider the special case of a uniform linear array (ULA) lying along the z-axis with inter-microphone distance ℓ and a single far-field source ($J=1$) with wavelength $\lambda = c/\nu_f$ impinging the array from the elevation angle θ. In this case, the direct propagation path is entirely determined by ℓ and θ, as illustrated in Fig. 5. The TDOAs are given by $(q_{ij} - q_{1j})/c = (i-1)\ell\cos(\theta)/c$. We can therefore express the steering vector (14) as

$$\mathbf{d}(\theta,\lambda) = \left[1, e^{-2\jmath\pi\frac{\ell}{\lambda}\cos(\theta)}, \dots, e^{-2\jmath\pi(I-1)\frac{\ell}{\lambda}\cos(\theta)}\right]^T. \quad (23)$$

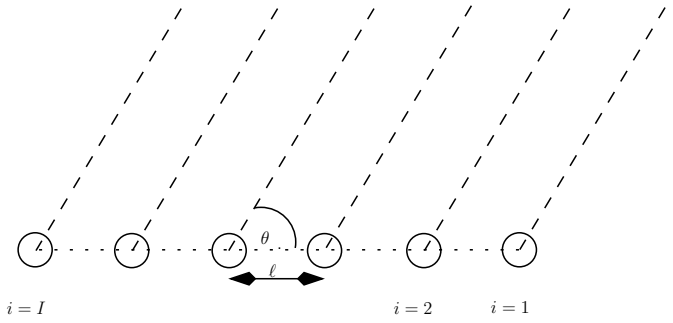


Figure 5. Uniform linear array (along the z-axis) for far-field signals.

The complex-valued response of the array, or *beampattern*, as a function of the angle θ is then given by

$$B\left(\theta; \frac{\ell}{\lambda}\right) = \mathbf{w}^H\mathbf{d}(\theta,\lambda) = \sum_{i=1}^{I} w_i e^{-2\jmath\pi(i-1)\frac{\ell}{\lambda}\cos(\theta)}. \quad (24)$$

Define the delay-and-sum (DS) beamformer, steered towards θ_0, as the beamformer with weights $w_i = \frac{1}{I}e^{-2\jmath\pi(i-1)\frac{\ell}{\lambda}\cos(\theta_0)}$. In this case, the absolute squared beampattern, called *beampower*, is given by

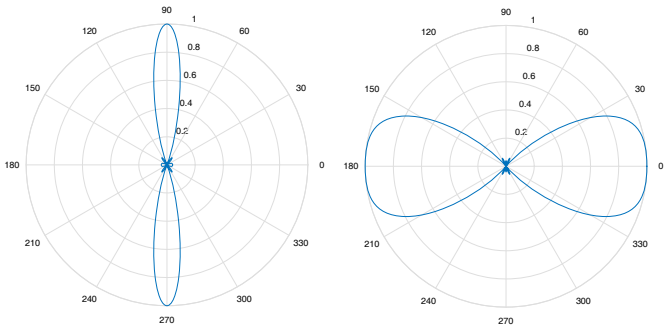
$$\left|B\left(\theta; \frac{\ell}{\lambda}\right)\right|^2 = \left|\frac{\sin\left(I\pi\frac{\ell}{\lambda}(\cos(\theta)-\cos(\theta_0))\right)}{I\sin\left(\pi\frac{\ell}{\lambda}(\cos(\theta)-\cos(\theta_0))\right)}\right|^2. \quad (25)$$

Typical beampatterns as functions of the steering angle and of the ratio $\frac{\ell}{\lambda}$ are depicted in Fig. 6. In Fig. 6(a)-6(b) we set $\frac{\ell}{\lambda} = \frac{1}{2}$. A ULA with $\frac{\ell}{\lambda} = \frac{1}{2}$ is usually referred to as a *standard linear array*. In Fig. 6(a) the steering direction is perpendicular to the array axis and in Fig. 6(b) it is parallel to it. The former look-direction is called *broadside* and the latter *endfire*. Note, that the beampatterns' shape is very distinct. The consequences of setting the inter-microphone distance to a very low value, i.e. $\frac{\ell}{\lambda} \ll 1$, can be deduced from Fig. 6(c), where the beampattern is almost *omnidirectional*, and to a very high value, i.e. $\frac{\ell}{\lambda} \gg 1$, can be deduced from Fig. 6(d), where the beampattern exhibits *grating lobes*, which are the result of spatial aliasing.

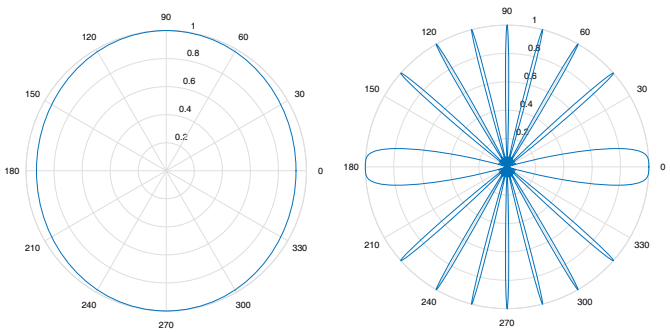
3) Directivity: An important attribute of a beamformer is its *directivity*, defined as the response towards the look direction divided by the integral over all other possible directions. The directivity in dB scale is denoted *directivity index*. In its most general form [119], applicable to any propagation regime (e.g. in a reverberant environment), the directivity at a given frequency can be defined as

$$\mathrm{D}_{\mathrm{gen}}(\mathbf{w},\mathbf{k}) = \frac{|\mathbf{w}^H\mathbf{a}(\mathbf{k}_0)|^2}{\kappa^{-1}\oint_{\mathbf{k}\in\mathcal{K}}|\mathbf{w}^H\mathbf{a}(\mathbf{k})|^2\mathcal{A}(\mathbf{k})d\mathbf{k}} \quad (26)$$

where $\mathbf{a}(\mathbf{k})$ is the vector of ATFs from the three dimensional source position $\mathbf{k}$ to the microphones, $\mathcal{A}(\mathbf{k})$ is the weight for each position $\mathbf{k}$ in the entire space $\mathcal{K}$, and $\mathbf{k}_0$ is the look direction. The normalization factor of the spatial integral is defined as $\kappa \triangleq \oint_{\mathbf{k}\in\mathcal{K}}\mathcal{A}(\mathbf{k})d\mathbf{k}$. In the most familiar definition of directivity, far-field is assumed. The abstract ATF parametrization is replaced by the wave propagation vector



(a) $\theta_0 = 90^o$; $\frac{\ell}{\lambda} = \frac{1}{2}$ (b) $\theta_0 = 0^o$; $\frac{\ell}{\lambda} = \frac{1}{2}$



(c) $\theta_0 = 90^o$; $\frac{\ell}{\lambda} = \frac{1}{32}$ (d) $\theta_0 = 90^o$; $\frac{\ell}{\lambda} = \frac{4}{1}$

Figure 6. Beampower of the DS beamformer for a ULA (along the z-axis).

in (22).

The directivity in spherical coordinates is then given by

$$\mathrm{D}_{\mathrm{sph}}(\mathbf{w}, \phi_0, \theta_0) = \frac{|\mathbf{w}^H \mathbf{d}(\mathbf{k}_0)|^2}{\frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi} \sin(\theta)|B(\phi,\theta)|^2 d\phi d\theta} \tag{27}$$

with $\mathbf{k}_0 = [\sin(\theta_0)\cos(\phi_0), \sin(\theta_0)\sin(\phi_0), \cos(\theta_0)]^T$ the look direction of the array. Assuming that the response in the look direction is equal to 1, this expression simplifies to [118]

$$\mathrm{D}_{\mathrm{sph}}(\mathbf{w}, \phi, \theta) = \left(\mathbf{w}^H \mathbf{\Omega} \mathbf{w}\right)^{-1} \tag{28}$$

where $\mathbf{\Omega}$ is the covariance matrix of a diffuse sound field whose entries $\Omega_{ii'}$ are given in (5).

Maximizing the directivity with respect to the array weights results in[2]

$$\mathrm{D}_{\max}(\phi_0, \theta_0) = \mathbf{d}^H(\phi_0, \theta_0)\mathbf{\Omega}^{-1}\mathbf{d}(\phi_0, \theta_0). \tag{29}$$

As evident from this expression, the directivity may depend on the steering direction. It can be shown that the maximum directivity attained by the standard linear array ($\frac{\ell}{\lambda} = \frac{1}{2}$) is equal to the number of microphones I, which is independent of the steering angle. The array weights in this case are given by $w_i = \frac{1}{I}$, $i = 1, \ldots, I$ assuming broadside look direction. If the directivity of a beamformer significantly exceeds I, it is called super-directive (SD). It was also shown in [120] that for an endfire array with vanishing inter-microphone distance, i.e. $\frac{\ell}{\lambda} \to 0$, the directivity approaches I^2. It was claimed that "it is most unlikely" that any other beamformer can attain higher directivity.

4) Sensitivity: Another attribute of a beamformer is its sensitivity to array imperfections.

Let the source image at the input of the microphone array be $\mathbf{c} = s \cdot \mathbf{a}(\mathbf{k}_0)$, and let $\mathbf{u}$ be the respective noise component. Define the source variance as σ_s^2 and the noise covariance matrix as $\mathbf{\Sigma}_{\mathbf{u}}$. The signal to noise ratio (SNR) at the output of the microphone array is therefore given by:

$$\mathrm{SNR}_{\mathrm{out}} = \frac{\sigma_s^2 |\mathbf{w}^H \mathbf{a}(\mathbf{k}_0)|^2}{\mathbf{w}^H \mathbf{\Sigma}_{\mathbf{u}} \mathbf{w}}. \tag{30}$$

If the noise is spatially-white, i.e. $\mathbf{\Sigma}_{\mathbf{u}} = \sigma_u^2 \mathbf{I}$, then:

$$\mathrm{SNR}_{\mathrm{out}} = \frac{\sigma_s^2}{\sigma_u^2} \frac{|\mathbf{w}^H \mathbf{a}(\mathbf{k}_0)|^2}{\mathbf{w}^H \mathbf{w}} = \mathrm{SNR}_{\mathrm{in}} \frac{|\mathbf{w}^H \mathbf{a}(\mathbf{k}_0)|^2}{\mathbf{w}^H \mathbf{w}} \tag{31}$$

with $\mathrm{SNR}_{\mathrm{in}} = \frac{\sigma_s^2}{\sigma_u^2}$.

Further assuming unit gain in the look direction, the SNR improvement, denoted as white noise gain (WNG), is given by:

$$\mathrm{WNG} = \frac{1}{\mathbf{w}^H \mathbf{w}} = \|\mathbf{w}\|^{-2} \tag{32}$$

where $\|\bullet\|$ stands for the ℓ_2 norm of a vector. It was shown in [121] that the numerical *sensitivity* of an array, i.e. its sensitivity to perturbations of the microphone positions and to the beamformer's weights, is inversely proportional to its WNG:

$$\mathcal{S} = \frac{1}{\mathrm{WNG}} = \|\mathbf{w}\|^2. \tag{33}$$

It was further shown in [121] that there is a tradeoff between the array directivity and its sensitivity and that the SD beamformer suffers from infinite sensitivity to mismatch between the nominal design parameters and the actual parameters. It was therefore proposed to constrain the norm of $\mathbf{w}$ to obtain a more robust design. It should be noted that if the microphone position perturbations are coupled (e.g. if the microphones share the same packaging) a modified norm constraint should be applied to guarantee low numerical sensitivity [122].

B. Array geometries

The ULA is just one possible array geometry among many others. In most algorithms discussed in this survey, no particular array geometry is assumed. Nowadays, microphones can be arbitrarily mounted on a device (e.g., cellphone, tablet, personal computer, hearing aid, smart watch) or several cooperative devices. In many cases, the microphone placement is determined by the product design constraints rather than by acoustic considerations. Ad hoc arrays can also be formed by concatenating several devices, each of which equipped with a small microphone array and limited processing power and communication capabilities. Ad hoc arrays will be briefly discussed in IX-C3.

Despite the fact that arbitrary array constellations are widespread, specific array geometries are still very important and have therefore attracted the attention of both Academia and Industry. We will now briefly describe some of the common microphone array geometries, namely differential and spherical microphone arrays.

[2]The array weights that maximize the directivity are given by the MVDR beamformer (43) with $\mathbf{\Sigma}_{\mathbf{u}} = \mathbf{\Omega}$, which will be discussed in Section V-B.

Differential arrays [123]–[127] are small-sized arrays with microphone spacing significantly smaller than the speech wavelength. They implement the spatial derivative of the sound pressure field and achieve a higher directivity than regular arrays, close to that of the SD beamformer. However, the sensitivity to array imperfections is excessively high. The most commonly used differential arrays implement the first-order derivative, but higher-order geometries exist. A device that can directly measure the sound velocity, i.e. the first-order vector derivative of the sound pressure, is also available [128].

Spherical microphone arrays [129], [130] have also attracted attention, due to their ability to symmetrically analyze tridimensional sound-fields [131]–[133] (see also dual-radius spherical arrays [134]). This analysis is conveniently carried out in the spherical harmonic domain by using the spherical Fourier transform (SFT). The interested reader is referred to a recently published book entirely dedicated to this topic [135].

Finally, crystal-shaped geometries have been used in [136]. They make it possible to diagonalize the (unknown) noise covariance matrix by a fixed, known transform, provided that it meets an isotropy condition.

C. From geometry to linear algebra

The representation of the spatial filtering capabilities of beamformers as a function of the DOA is not very informative for unstructured arrays, whose geometry does not comply with a particular structure, e.g. linear, circular or spherical. Moreover, sound propagation in a reverberant environment is much more intricate than in free field.

The reflections of the sound wave are captured by the AIR. From this perspective, each source can be represented as a vector in a high-dimensional space whose dimension is the number of reflections times the number of microphones. A beamformer can be interpreted as a linear operator in this (abstract) space. The various operations can be interpreted in terms of linear algebra, without resorting to beampatterns as a function of the DOA. One advantage of this perspective lies in the ability to separate desired and interfering sources sharing the same DOA [137], due to the fact that two sources with the same DOA, but with different distances from the microphone array, generally exhibit different reflection patterns. As previously discussed, working in a very high-dimensional space is usually impractical.

It was therefore proposed both in the fields of beamforming and BSS to replace the simple steering vectors (7) and (8) by the ATFs or the respective RTFs. It was shown in [138] that the peak of the RTF in the time domain corresponds to the TDOA between the microphones, provided that the DRR is sufficiently high. Hence the RTF can be viewed as a generalization of the TDOA.

D. Fixed beamforming

The beamformers we have seen thus far are *fixed beamformers* (FBFs), which only rely on the DOA or the RTFs of the target source. FBF designs are suitable when the target direction is known a priori, e.g., in cellphones, cars or hearing aids. In these cases, the beamformer is designed to focus on the target source while minimizing noise and reverberation arriving from other directions. These designs require low computational complexity, but they may be prone to performance degradation when the microphone positions are not accurately known (see Section IV-A). A semi-fixed beamforming approach, suitable for cases when the position of the target source cannot be determined in advance, is to estimate its DOA and to design a FBF steered towards it. Alternatively, the AIRs or the RTFs between the target source position and the microphones can be estimated during a calibration process and used to construct a matched-filter FBF [139].

A common FBF is the DS beamformer [140], which consists of averaging the delay-compensated microphone signals. Although simple, it can be shown to attain the optimal directivity for a spatially-white noise field. The beamwidth and sidelobe levels of the beampattern can be further controlled by spatial windowing of the microphone signals before averaging them. This is simply implemented as a weighted-sum beamformer [141].

Considering a diffuse noise field, or scenarios where the noise field is unknown, a beamformer which steers the beam towards the target while minimizing the interferences arriving from all other directions, can be designed [142], [143]. In the special case of a target located at the endfire of the array with vanishing inter-microphone distance, the directivity of this design approaches I^2 (see discussion in Sec. IV-A). In practice, due to the non-zero inter-microphone distance, the beampattern becomes frequency-dependent. While the DS beamformer has a quasi-omnidirectional beampattern at low frequencies, the beamwidth becomes narrower at higher frequencies. These different beamwidths result in a low-pass effect on the output signal. At very high frequencies the beampattern is also prone to spatial aliasing (see Fig. 6). A first cure to this phenomenon is to split the array into subarrays that cover different frequency bands [2], [123], [144]. In [145] the theory of frequency-invariant beampatterns for far-field beamforming is developed and a practical implementation is presented.

Eigen-filter (non-iterative) design methods for obtaining arbitrary directivity patterns using arbitrary microphones configurations are presented in [146]. Common iterative methods for FBF design, such as least squares (LS), maximum energy and nonlinear optimization, are also explored.

V. Spatial filter design criteria

From now on, we focus on *data-dependent spatial filters*, which depend on the input signal statistics. Compared with FBFs, data-dependent designs attain higher performance due to their ability to adapt to the actual ATFs and the statistics of target and interfering sources. In many cases these spatial filters are also *adaptive*, i.e. time-varying. However, they usually require substantially higher computational complexity. In this section we explore many popular data-dependent spatial filter design criteria. We start in Section V-A with a general framework for the narrowband model and recognize several well-known beamforming criteria as special cases of this

framework. In Section V-B we elaborate on the minimum variance distortionless response (MVDR) and linearly constrained minimum variance (LCMV) beamformers, and in Section V-C on the multichannel Wiener filter (MWF) beamformer and its variant known as the speech distortion weighted multichannel Wiener filter (SDW-MWF). We then proceed with beamforming criteria for inter-frame, inter-frequency, or full-rank covariance models in Section V-D and spatial filter design criteria for sparse speech models in Section V-E. All these criteria rely on a set of parameters such as the RTFs and the second order statistics of the sources, whose estimation will be handled in Section VI.

A. General criterion for the narrowband model

Assume the narrowband approximation in the STFT domain (6) holds. Further assume that the received microphone signals comprise J_p *point* sources of interest and $J - J_p$ noise sources with arbitrary spatial characteristics. Using (2) and (6) and the above assumptions the microphone signals are given by:

$$\mathbf{x}(n,f) = \mathbf{A}(n,f)\mathbf{s}(n,f) + \mathbf{u}(n,f) \tag{34}$$

where $\mathbf{A}(n,f) = \left[\mathbf{a}_1(n,f), \ldots, \mathbf{a}_{J_p}(n,f)\right]$, $\mathbf{s}(n,f) = \left[s_1(n,f), \ldots, s_{J_p}(n,f)\right]^T$, and $\mathbf{u}(n,f) = \sum_{j=J_p+1}^{J} \mathbf{c}_j(n,f)$ is the contribution of all noise sources. The frame index n and the frequency index f are henceforth omitted for brevity, whenever no ambiguity occurs. Denoting by $\mathbf{\Sigma_x} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\}$ the covariance matrix of the received signals, $\mathbf{\Sigma_u} = \mathbb{E}\{\mathbf{u}\mathbf{u}^H\}$ the covariance matrix of the noise signals, and $\mathbf{\Sigma_s} = \mathrm{diag}(\sigma^2_{s_1}, \ldots, \sigma^2_{s_{J_p}})$ the covariance matrix of the signals of interest, assumed to be mutually independent, the following relation holds:

$$\mathbf{\Sigma_x} = \mathbf{A}\mathbf{\Sigma_s}\mathbf{A}^H + \mathbf{\Sigma_u}. \tag{35}$$

In the most general form, define $\mathbf{d} = \mathbf{Q}^H\mathbf{s}$ as the desired output vector, where $\mathbf{Q}$, denoted as the desired response matrix, is a matrix of weights controlling the contributions of the signals of interest at all desired outputs, and $\widehat{\mathbf{d}} = \mathbf{W}^H\mathbf{x}$ the outputs of a filtering matrix $\mathbf{W}$ (note that the desired responses are defined by $\mathbf{Q}^*$). Then, the filtering matrix $\mathbf{W}$ is set to satisfy the following minimum mean square error (MMSE) criterion:

$$\mathbf{W}_{\text{MO-MWF}} = \underset{\mathbf{W}}{\mathrm{argmin}}\, \mathbb{E}\left\{\mathrm{tr}\left((\widehat{\mathbf{d}} - \mathbf{d})(\widehat{\mathbf{d}} - \mathbf{d})^H\right)\right\} = \underset{\mathbf{W}}{\mathrm{argmin}}\left\{(\mathbf{Q}^H - \mathbf{A}\mathbf{W}^H)\mathbf{\Sigma_s}(\mathbf{Q} - \mathbf{W}\mathbf{A}^H) + \mathbf{W}^H\mathbf{\Sigma_u}\mathbf{W}\right\} \tag{36}$$

where the multi-output MWF matrix, $\mathbf{W}_{\text{MO-MWF}}$, is given by:

$$\mathbf{W}_{\text{MO-MWF}} = \mathbf{\Sigma_x}^{-1}\mathbf{A}\mathbf{\Sigma_s}\mathbf{Q} = \left(\mathbf{A}\mathbf{\Sigma_s}\mathbf{A}^H + \mathbf{\Sigma_u}\right)^{-1}\mathbf{A}\mathbf{\Sigma_s}\mathbf{Q}. \tag{37}$$

In the more widely-used scenario, a *single* desired combination of the signals of interest $d = \mathbf{q}^H\mathbf{s}$ is considered, where $\mathbf{q}$, denoted as the desired response vector, is a vector of weights controlling the contribution of the signals at the desired output (note that the desired responses are defined by $\mathbf{q}^*$). Let $\widehat{d} = \mathbf{w}^H\mathbf{x}$ be the output of a beamformer $\mathbf{w}$. The beamformer weights are set to satisfy the following MMSE criterion [147]:

$$\underset{\mathbf{w}}{\mathrm{argmin}}\, \mathbb{E}\{|\widehat{d} - d|^2\} = \underset{\mathbf{w}}{\mathrm{argmin}}\, \mathbb{E}\{|\mathbf{w}^H\mathbf{x} - \mathbf{q}^H\mathbf{s}|^2\}. \tag{38}$$

Several criteria can be derived from (38). Starting from the single desired source case $J_p = 1$, i.e. $d = q^*s_1$, the MWF can be derived by rewriting the MMSE criterion as

$$\mathbf{w}_{\text{MWF}} = \underset{\mathbf{w}}{\mathrm{argmin}}\left\{\left|q - \mathbf{a}_1^H\mathbf{w}\right|^2\sigma^2_{s_1} + \mathbf{w}^H\mathbf{\Sigma_u}\mathbf{w}\right\}. \tag{39}$$

The minimizer of the cost function in (39) is the celebrated MWF:

$$\mathbf{w}_{\text{MWF}} = \left(\sigma^2_{s_1}\mathbf{a}_1\mathbf{a}_1^H + \mathbf{\Sigma_u}\right)^{-1}\sigma^2_{s_1}\mathbf{a}_1 q = \frac{\sigma^2_{s_1}\mathbf{\Sigma_u}^{-1}\mathbf{a}_1}{1 + \sigma^2_{s_1}\mathbf{a}_1^H\mathbf{\Sigma_u}^{-1}\mathbf{a}_1}q. \tag{40}$$

The MWF cost function comprises two terms. The first term is the power of the speech distortion induced by spatial filtering, while the second is the noise power at the output of the beamformer. These two terms are also known as artifacts and interference, respectively, in the source separation literature.

To gain further control on the cost function, a tradeoff parameter may be introduced, resulting in the SDW-MWF cost function [148]:

$$\mathbf{w}_{\text{SDW-MWF}} = \underset{\mathbf{w}}{\mathrm{argmin}}\left\{\left|q - \mathbf{a}_1^H\mathbf{w}\right|^2\sigma^2_{s_1} + \mu\mathbf{w}^H\mathbf{\Sigma_u}\mathbf{w}\right\} \tag{41}$$

where μ is a tradeoff factor between speech distortion and noise reduction. Minimizing the criterion in (41) yields:

$$\mathbf{w}_{\text{SDW-MWF}} = \frac{\sigma^2_{s_1}\mathbf{\Sigma_u}^{-1}\mathbf{a}_1}{\mu + \sigma^2_{s_1}\mathbf{a}_1^H\mathbf{\Sigma_u}^{-1}\mathbf{a}_1}q. \tag{42}$$

By tuning μ in the range $(0, \infty)$, the speech distortion level can be traded for the residual noise level. For $\mu \to \infty$, maximum noise reduction but maximum speech distortion are obtained. Setting $\mu = 1$, the SDW-MWF identifies with the MWF. Finally, for $\mu \to 0$, the SDW-MWF identifies with the MVDR beamformer, with a strict distortionless response $\mathbf{w}^H\mathbf{a}_1 = q$:

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{\Sigma_u}^{-1}\mathbf{a}_1}{\mathbf{a}_1^H\mathbf{\Sigma_u}^{-1}\mathbf{a}_1}q \tag{43}$$

which optimizes the following constrained minimization:

$$\mathbf{w}_{\text{MVDR}} = \underset{\mathbf{w}}{\mathrm{argmin}}\left\{\mathbf{w}^H\mathbf{\Sigma_u}\mathbf{w}\ \text{s.t.}\ \mathbf{a}_1^H\mathbf{w} = q\right\}. \tag{44}$$

More information regarding the SDW-MWF and MVDR beamformers and their relations can be found in [92]. In Section VII we will discuss in details the decomposition of the SDW-MWF into an MVDR beamformer and a subsequent postfiltering stage.

It is also easy to verify [118] that the MVDR and the following minimum power distortionless response (MPDR) criteria are equivalent:

$$\mathbf{w}_{\text{MPDR}} = \underset{\mathbf{w}}{\mathrm{argmin}}\left\{\mathbf{w}^H\mathbf{\Sigma_x}\mathbf{w}\ \text{s.t.}\ \mathbf{a}_1^H\mathbf{w} = q\right\}. \tag{45}$$

The resulting beamformer

$$\mathbf{w}_{\text{MPDR}} = \frac{\mathbf{\Sigma_x}^{-1}\mathbf{a}_1}{\mathbf{a}_1^H\mathbf{\Sigma_x}^{-1}\mathbf{a}_1}q \tag{46}$$

exhibits, however, higher sensitivity to misalignment errors than the MVDR beamformer [149].

Finally, it can be shown [121] that the maximum SNR (MSNR) beamformer that maximizes the output SNR:

$$\mathbf{w}_{\text{MSNR}} = \underset{\mathbf{w}}{\text{argmax}} \left\{ \frac{|\mathbf{a}_1^H \mathbf{w}|^2}{\mathbf{w}^H \boldsymbol{\Sigma}_{\mathbf{u}} \mathbf{w}} \right\}. \tag{47}$$

The MSNR beamformer is given by

$$\mathbf{w}_{\text{MSNR}} = \zeta \boldsymbol{\Sigma}_{\mathbf{u}}^{-1} \mathbf{a}_1 \tag{48}$$

with ζ an arbitrary scalar. The MSNR beamformer identifies with the MVDR and the MPDR beamformers if they satisfy the same constraint $\mathbf{a}_1^H \mathbf{w} = q$. The MSNR beamformer was applied to speech enhancement in [150], [151].

Returning to the multi-speaker case, the multiple speech distortion weighted multichannel Wiener filter (MSDW-MWF) criterion can be defined [147]:

$$\mathbf{w}_{\text{MSDW-MWF}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbf{w}^H \boldsymbol{\Sigma}_{\mathbf{u}} \mathbf{w} + \left(\mathbf{q} - \mathbf{A}^H \mathbf{w}\right)^H \boldsymbol{\Lambda} \boldsymbol{\Sigma}_{\mathbf{s}} \left(\mathbf{q} - \mathbf{A}^H \mathbf{w}\right) \right\} \tag{49}$$

where $\boldsymbol{\Lambda} = \text{diag}\left\{\lambda_1, \ldots, \lambda_{J_p}\right\}$ is a diagonal weight matrix with tradeoff factors controlling noise reduction and the deviation from the desired response on its main diagonal. The MSDW-MWF beamformer optimizing the criterion in (49) is given by:

$$\mathbf{w}_{\text{MSDW-MWF}} = \left(\mathbf{A} \boldsymbol{\Lambda} \boldsymbol{\Sigma}_{\mathbf{s}} \mathbf{A}^H + \boldsymbol{\Sigma}_{\mathbf{u}}\right)^{-1} \mathbf{A} \boldsymbol{\Lambda} \boldsymbol{\Sigma}_{\mathbf{s}} \mathbf{q}. \tag{50}$$

Various widely-used beamformers can be derived by setting the values of the weight matrix $\boldsymbol{\Lambda}$ in the generalized MSDW-MWF criterion:

1) By setting $\boldsymbol{\Lambda} = \mathbf{I}$ we get the MWF for estimating a desired combination of all signals of interest $d = \mathbf{q}^H \mathbf{s}$:

$$\mathbf{w}_{\text{M-MWF}} = \left(\mathbf{A} \boldsymbol{\Sigma}_{\mathbf{s}} \mathbf{A}^H + \boldsymbol{\Sigma}_{\mathbf{u}}\right)^{-1} \mathbf{A} \boldsymbol{\Sigma}_{\mathbf{s}} \mathbf{q} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \mathbf{A} \boldsymbol{\Sigma}_{\mathbf{s}} \mathbf{q}. \tag{51}$$

2) Assume now that only one signal of interest exists, i.e. $J_p = 1$, and $\boldsymbol{\Lambda} = \mu^{-1}$. In this case the MSDW-MWF beamformer simplifies to the SDW-MWF beamformer:

$$\begin{aligned} \mathbf{w}_{\text{SDW-MWF}} &= \left(\mathbf{a}_1 \sigma_{s_1}^2 \mathbf{a}_1^H + \boldsymbol{\Sigma}_{\mathbf{u}}\right)^{-1} \mathbf{a}_1 \sigma_{s_1}^2 q \\ &= \frac{\sigma_{s_1}^2 \boldsymbol{\Sigma}_{\mathbf{u}}^{-1} \mathbf{a}_1}{\mu + \sigma_{s_1}^2 \mathbf{a}_1^H \boldsymbol{\Sigma}_{\mathbf{u}}^{-1} \mathbf{a}_1} q. \end{aligned} \tag{52}$$

where the last transition is due to Woodbury identity [152]. The MVDR and MPDR beamformers are obtained from the SDW-MWF as explained above.

3) Selecting $\boldsymbol{\Lambda} = \mu^{-1} \boldsymbol{\Sigma}_{\mathbf{s}}^{-1}$ we obtain at the limit:

$$\begin{aligned} &\lim_{\mu \to 0} \mathbf{w}_{\text{MSDW-MWF}}(\boldsymbol{\Lambda} = \mu^{-1} \boldsymbol{\Sigma}_{\mathbf{s}}^{-1}) = \\ &\boldsymbol{\Sigma}_{\mathbf{u}}^{-1} \mathbf{A} \left(\mathbf{A}^H \boldsymbol{\Sigma}_{\mathbf{u}}^{-1} \mathbf{A}\right)^{-1} \mathbf{q} \end{aligned} \tag{53}$$

which is exactly the LCMV beamformer. It is easily verified that the LCMV beamformer is equivalent to the linearly constrained minimum power (LCMP) beamformer [118]:

$$\mathbf{w}_{\text{LCMP}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \mathbf{A} \left(\mathbf{A}^H \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \mathbf{A}\right)^{-1} \mathbf{q}. \tag{54}$$

The LCMV beamformer optimizes the following criterion:

$$\mathbf{w}_{\text{LCMV}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbf{w}^H \boldsymbol{\Sigma}_{\mathbf{u}} \mathbf{w} \text{ s.t. } \mathbf{A}^H \mathbf{w} = \mathbf{q} \right\}. \tag{55}$$

and the LCMP criterion is obtained by substituting $\boldsymbol{\Sigma}_{\mathbf{u}}$ by $\boldsymbol{\Sigma}_{\mathbf{x}}$ in (55). The LCMP beamformer is known to be much more sensitive to misalignment than the LCMV beamformer [149]. Note, that while an interference source can be perfectly nulled out by adding a proper constraint to the LCMV criterion, its power will only be suppressed by the minimization operation of the MVDR criterion. Interestingly, the MVDR beamformer can also direct an almost perfect null towards an interference source, provided that the respective spatial covariance matrix $\boldsymbol{\Sigma}_{\mathbf{u}}$ is a rank-1 matrix. Similar relations, with the proper modifications, apply to the LCMP and MPDR beamformers.

The reader is referred to [118], [140], [149] for comprehensive surveys of beamforming techniques.

In the next subsections, we will elaborate on specific structures and implementation of widely-used beamformers. In Section V-B two important distortionless beamformers, namely the MVDR and LCMV beamformers, are discussed. We extend the discussion on MMSE beamformers in Section V-C and elaborate on methods to control the level of distortion. Beamforming structures that extend the narrowband approximation are discussed in Section V-D. Spatial filtering criteria that go beyond second-order statistics of the signals are presented in Section V-E.

B. MVDR and LCMV

The desired signal defined in the previous general beamformer formulation consists of a linear combination of the "dry" sources (prior to the filtering process of the AIRs). Hence, the designed beamformer not only aims to reduce the noise, but also aims to de-reverberate the speech signals. Assuming that reverberation alone does not compromise intelligibility, which is the case in many scenarios, the de-reverberation requirement can be relaxed. A modified beamformer can be obtained by redefining the desired signal as a linear combination of the sources as received by some reference microphones. Generally, the reference microphone for each of the sources can be selected differently. Here, for brevity, we assume that the reference microphones are the same for all sources, and arbitrarily select it to be the first microphone. Redefine the modified vector of desired responses as:

$$\tilde{\mathbf{q}} \triangleq \begin{bmatrix} a_{11}^* & \cdots & a_{J_p,1}^* \end{bmatrix}^T. \tag{56}$$

Consider the special case of enhancing a single desired speaker, i.e., $J_p = 1$, contaminated by noise, using the MVDR criterion. Using the definition (56) with $J_p = 1$ in the MVDR criterion (44) results in a modified MVDR beamformer aiming at the enhancement of the desired signal *image* on the

reference microphone $a_{11}s_1$:

$$\tilde{\mathbf{w}}_{\text{MVDR}} \triangleq \frac{\mathbf{\Sigma}_{\mathbf{u}}^{-1}\tilde{\mathbf{a}}_1}{\tilde{\mathbf{a}}_1^H\mathbf{\Sigma}_{\mathbf{u}}^{-1}\tilde{\mathbf{a}}_1} \tag{57}$$

where $\tilde{\mathbf{a}}_1$ denotes the RTF vector of the desired source as defined in (9). In [153] the SNR improvement of an MVDR beamformer is evaluated as a function of the reverberation level at the output. It is concluded that a tradeoff between noise reduction and dereverberation exist, i.e. the highest SNR improvement is obtained when dereverberation is sacrificed.

Consider the multiple speakers scenario, and assume that J_p speakers of interest can be classified into two groups, namely as desired or as interfering speakers. Without loss of generality, we assume that the first J_α sources are desired and denote their respective ATF matrix by $\mathbf{A}_\alpha$. Correspondingly, the last J_β speakers are assumed to be interfering and their respective ATF matrix is denoted as $\mathbf{A}_\beta$. The total number of source of interest therefore satisfies $J_p = J_\alpha + J_\beta$. Similarly to the above, relaxing the dereverberation requirement, the goal of the beamformer is to extract the desired sources as received by the reference microphone while mitigating the interfering speakers and minimizing the background noise. Explicitly, the constraints set can be defined as

$$\tilde{\mathbf{A}}^H\mathbf{w} = \mathbf{q}_{\text{LCMV}} \tag{58}$$

where $\tilde{\mathbf{A}} \triangleq \begin{bmatrix} \tilde{\mathbf{A}}_\alpha & \tilde{\mathbf{A}}_\beta \end{bmatrix}$ comprises the RTFs of the desired and interfering speakers arranged in matrices $\tilde{\mathbf{A}}_\alpha$ and $\tilde{\mathbf{A}}_\beta$, respectively, and $\mathbf{q}_{\text{LCMV}} \triangleq \begin{bmatrix} \mathbf{1}_{1\times J_\alpha} & \mathbf{0}_{1\times J_\beta} \end{bmatrix}^T$. A straightforward computation of the LCMV beamformer requires knowledge of the RTFs of the sources (both desired and interfering). It can be shown (see [137]) that an equivalent constraints set can be formulated as:

$$\begin{bmatrix} \tilde{\mathbf{Q}}_\alpha & \tilde{\mathbf{Q}}_\beta \end{bmatrix}^H\mathbf{w} = \mathbf{q}_{\text{LCMV}} \tag{59}$$

where the matrices $\mathbf{Q}_\alpha$ and $\mathbf{Q}_\beta$ are arbitrary bases spanning the column-subspace of the matrices $\mathbf{A}_\alpha$ and $\mathbf{A}_\beta$, respectively, and $\tilde{\mathbf{Q}}_\alpha$ and $\tilde{\mathbf{Q}}_\beta$ are their normalized counterparts defined as:

$$\tilde{\mathbf{Q}}_\alpha = \text{diag}(Q_{\alpha,11},\ldots,Q_{\alpha,1J_\alpha})^{-1}\mathbf{Q}_\alpha \tag{60a}$$
$$\tilde{\mathbf{Q}}_\beta = \text{diag}(Q_{\beta,11},\ldots,Q_{\beta,1J_\beta})^{-1}\mathbf{Q}_\beta. \tag{60b}$$

The operator $\text{diag}(\cdot)$ denotes a diagonal matrix with the argument on its diagonal and $Q_{\alpha,1j}$ denotes the first element in the j-th column of the matrix $\mathbf{Q}_\alpha$. Constructing the LCMV beamformer with the constraints set in (59) can be shown to be equivalent to the construction with (58). Moreover, using (58) relaxes the requirement for estimating the RTF vectors for each of the sources, and substitutes it with estimating two basis matrices, one for each group of sources (desired and interfering, respectively). A practical method for estimating the basis matrices $\mathbf{Q}_\alpha$ and $\mathbf{Q}_\beta$ is discussed in Section VI-B.

C. MWF, SDW-MWF and parametric MWF

Time-domain implementation of single-source MWF-based speech enhancement is proposed in [154]. The covariance matrix of the received microphone signals comprises speech and noise components. Using generalized singular value decomposition (GSVD), the mixture and noise covariance matrices can be jointly diagonalized [155]. Utilizing the low-rank structure of the speech component, a time-recursive and reduced-complexity implementation is proposed. The complexity can be further reduced by shortening the length of GSVD-based filters.

In later work [156], a similar solution to the SDW-MWF [148] was derived from a different perspective. It is suggested to minimize the noise variance at the output of the beamformer while constraining the maximal distortion incurred to the speech signal, denoted σ_D^2. The beamformer which optimizes the latter criterion is called parametric MWF (PMWF):

$$\mathbf{w}_{\text{PMWF}} = \underset{\mathbf{w}}{\text{argmin}}\left\{\mathbf{w}^H\mathbf{\Sigma}_{\mathbf{u}}\mathbf{w} \text{ s.t. } \mathbb{E}\{|d-\widehat{d}|^2\} \le \sigma_D^2\right\} \tag{61}$$

The expression of the PMWF is identical to that of the SDW-MWF in (42). The relation between the parameters σ_D^2 of the PMWF and μ of the SDW-MWF does not have a closed-form representation in the general case. This relation and the performance of the PMWF are analyzed in [157].

D. Criteria for inter-frame, inter-frequency, or full-rank covariance models

The beamformers we have seen thus far rely on the narrowband approximation. The underlying MMSE criterion can also be used when this approximation does not hold, e.g., with inter-frame, inter-frequency, or full-rank covariance models.

With the full-rank covariance model in Section III-E, for instance, the target signal to be estimated is the vector $\mathbf{c}_j(n,f)$ of STFT coefficients of each spatial source image. Beamforming can then be achieved using a matrix of weights $\mathbf{W}(n,f)$ as $\widehat{\mathbf{c}}_j(n,f) = \mathbf{W}^H(n,f)\mathbf{x}(n,f)$. The MMSE criterion is expressed as

$$\underset{\mathbf{W}}{\text{argmin}}\,\mathbb{E}\{\|\mathbf{W}^H(n,f)\mathbf{x}(n,f) - \mathbf{c}_j(n,f)\|^2\} \tag{62}$$

and the solution is given by the MWF [113]

$$\mathbf{W}_j(n,f) = \mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{\Sigma}_{\mathbf{c}_j}(n,f) \tag{63}$$

with $\mathbf{\Sigma}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J}\mathbf{\Sigma}_{\mathbf{c}_j}(n,f)$. Variants of this criterion involving multiple target speakers and tradeoff between speech distortion and residual noise can be derived similarly to above.

A similar approach can also be used for the inter-frame and inter-frequency models in Section III-D. Beamformers then involve STFT coefficients from multiple frames or frequency bins as inputs and the MWF is obtained using a similar expression to (63) where the covariance matrices represent the covariance between multiple frames or frequency bins [104]–[106], [109]. In [47], the inter-frame model and the full-rank model are combined in a nested MVDR beamforming structure.

E. Sparsity-based criteria

The beamformers we have reviewed thus far are obtained by minimizing power criteria which can be expressed in terms

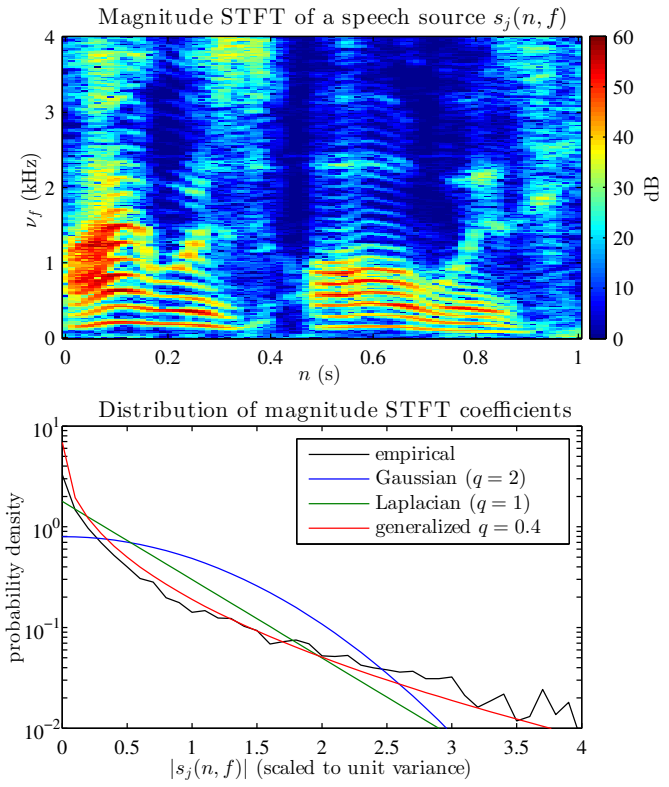


Figure 7. Distribution of the magnitude STFT of a speech source.

of the second-order statistics of the signals. Mathematically speaking, these statistics are sufficient to characterize Gaussian signals. However, audio signals are often nongaussian. Fig. 7 shows that, in the time-frequency domain, the distribution of speech signals is sparse: at each frequency, a few coefficients are large and most are close to zero compared to a Gaussian.

This has inspired researchers to design spatial filters that take this distribution into account. This is typically achieved by optimizing a maximum likelihood (ML) criterion under the narrowband model (6). Three approaches have been proposed.

1) Binary masking and local inversion: A first approach considers that each source is active in a few time-frequency bins so that only few sources are active in each time-frequency bin. The simplest model assumes that a single source $j^\star(n,f)$ is active in each time-frequency bin [85], [158], [159]. If we further assume that $j^\star(n,f)$ is uniformly distributed in $\{1,\dots,J\}$ and that the noise $\mathbf{u}(n,f)$ is Gaussian with covariance $\mathbf{\Sigma}_\mathbf{u}(f)$, the sources $s_j(n,f)$ and the model parameters $\boldsymbol{\theta}=\{\mathbf{a}_j(f),\mathbf{\Sigma}_\mathbf{u}(f)\}$ can be jointly estimated by maximizing the log-likelihood

$$\begin{aligned}\underset{\mathbf{s},\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{nf} &-\log\det(\pi\mathbf{\Sigma}_\mathbf{u}(f)) \\ &-(\mathbf{x}(n,f)-\mathbf{a}_j^\star(f)\,s_j^\star(n,f))^H \\ &\mathbf{\Sigma}_\mathbf{u}^{-1}(f)(\mathbf{x}(n,f)-\mathbf{a}_j^\star(f)\,s_j^\star(n,f)) \quad (64)\end{aligned}$$

where $\mathbf{a}_j^\star(f)$ and $s_j^\star(n,f)$ denote the value of $\mathbf{a}_j(f)$ and $s_j(n,f)$ for $j=j^\star(n,f)$. Given $j^\star(n,f)$ and $\boldsymbol{\theta}$, it turns out that the optimal value of the predominant source is obtained by the MVDR beamformer $s_j^\star(n,f)=\mathbf{w}_{\mathrm{MVDR}}^H(f)\mathbf{x}(n,f)$ where $\mathbf{w}_{\mathrm{MVDR}}(f)$ is given by (43) by identifying $\mathbf{a}_1$ with $\mathbf{a}_j^\star(f)$ and setting $q=1$. The other sources $s_j(n,f)$, $j\neq j^\star(n,f)$, are set to zero. This can be interpreted as a conventional MVDR beamformer followed by a binary postfilter equal to 1 for the predominant source and 0 for the other sources (see Section VII-A).

A variant of this approach assumes that a subset of sources $\mathcal{J}(n,f)\subset\{1,\dots,J\}$ is active in each time-frequency bin where the number of active sources is smaller than the number of microphones I [54], [160]–[163]. The ML criterion can then be written as

$$\begin{aligned}\underset{\mathbf{s},\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{nf} &-\log\det(\pi\mathbf{\Sigma}_\mathbf{u}(f)) \\ &-\left(\mathbf{x}(n,f)-\sum_{j\in\mathcal{J}(n,f)}\mathbf{a}_j(f)\,s_j(n,f)\right)^H \\ &\mathbf{\Sigma}_\mathbf{u}^{-1}(f)\left(\mathbf{x}(n,f)-\sum_{j\in\mathcal{J}(n,f)}\mathbf{a}_j(f)\,s_j(n,f)\right). \quad (65)\end{aligned}$$

Given $\mathcal{J}(n,f)$ and $\boldsymbol{\theta}$, the optimal value of each active source is now obtained by the LCMV beamformer $s_j(n,f)=\mathbf{w}_{\mathrm{LCMV}}^H(f)\mathbf{x}(n,f)$ whose general expression is given later in (73) where $\breve{\mathbf{A}}=[\mathbf{a}_j(f)]_{j\in\mathcal{J}(n,f)}$ and $\mathbf{q}=[0,\dots,1,\dots,0]^T$ with the value 1 in the position corresponding to source j. The activity patterns $j^\star(n,f)$ or $\mathcal{J}(n,f)$ and the model parameters $\boldsymbol{\theta}$ can be estimated using an EM algorithm (see Section VI-C). Alternative solutions include estimating $\boldsymbol{\theta}$ first using, e.g. the techniques in Section VI-B, and subsequently looping over all possible activity patterns and select the one yielding the largest likelihood, or even reestimating $[\mathbf{a}_j(f)]_{j\in\mathcal{J}(n,f)}$ in each time-frequency bin using other criteria than ML [163].

2) ICA and SCA: A second approach assumes that all sources are possibly active but their STFT coefficients $s_j(n,f)$ are independent and identically distributed (i.i.d.) according to a known sparse distribution. The circular generalized Gaussian distribution is a popular choice [164], [165]. It models the phases of the source STFT coefficients as uniformly distributed and their magnitudes as [166], [167]

$$p(|s_j(n,f)|)=q\frac{\beta^{1/q}}{\Gamma(1/q)}e^{-\beta\,|s_j(n,f)|^q} \quad (66)$$

where the parameters $0<q<2$ and $\beta>0$ govern respectively the shape and the variance of the prior and $\Gamma(\cdot)$ is the gamma function. This distribution includes the Laplacian ($q=1$) [72], [168], [169] as a special case and its sparsity increases with decreasing q. It was shown in [164] that $q=0.4$ matches well the distribution of speech, as illustrated in Fig. 7. Generalizations of this distribution [170] and other i.i.d. distributions [67], [76], [168], [171]–[174] have also been used.

In the so-called determined case, when the number of sources J is equal to the number of microphones I, estimating the matrix of ATFs $\mathbf{A}(f)$ is equivalent to estimating the matrix of beamformers $\mathbf{W}(f)=\mathbf{A}^{-1}(f)$, which can be used to jointly recover all sources as $\mathbf{s}(n,f)=\mathbf{W}^H(f)\mathbf{x}(n,f)$. The

optimal beamformers $\mathbf{W}(f)$ can then be estimated in the ML sense as

$$\mathbf{W}_{\mathrm{ICA}}(f) = \underset{\mathbf{W}(f)}{\operatorname{argmax}} \sum_{nf} \log p(\mathbf{x}(n,f)|\mathbf{A}(f)) \tag{67}$$

$$= \underset{\mathbf{W}(f)}{\operatorname{argmax}} \log|\det \mathbf{W}(f)| + \sum_{jnf} \log p(s_j(n,f)) \tag{68}$$

Interestingly, this criterion is equivalent to minimizing the mutual information $I(s_1,\ldots,s_J)$, which is an information-theoretic measure of dependency between random variables [175]. In other words, it results in maximizing the statistical independence of the source signals. For this reason, it was called nongaussianity-based independent component analysis (ICA) [176]–[178]. This is the most common form of ICA, which differs from the nonstationarity-based ICA stemming from the LGM in Section III-E. Minimum mutual information is more general than ML as it can also be applied when the distribution $p(s_j(n,f))$ is unknown. In practice though, most ICA methods rely on ML which is easier to optimize and can also be applied to enhance a single source [179]. The beamformer resulting from nongaussianity-based ICA significantly differs from the ones we have seen so far in that it can never be expressed in terms of the second-order statistics of the signals. Actually, it cannot even be computed in closed-form: parameter estimation and beamforming are tightly coupled as illustrated by the dashed arrow in Fig. 1. Iterative estimation algorithms will be reviewed in Section VI-D.

One limitation of the ICA criterion is that it is invariant with respect to permutation of the sources. Yet, the order of the sources must be aligned across the frequency bins. Linear constraints [70], [77] such as the one used for MVDR and penalty terms constraining $\mathbf{a}_j(f)$ to vary smoothly over frequency [75], [76] or to be close to the anechoic steering vector $\mathbf{d}_j(f)$ [78] have been used to constrain the optimization (68). Post-processing permutation alignment techniques which exploit the additional fact that the source short-term spectra are correlated across frequency bands have also been proposed [81], [180], [181].

In the so-called underdetermined case, when the number of sources J is larger than the number of microphones I, ICA cannot recover all sources anymore and joint ML estimation of $\mathbf{A}(f)$ and $s_j(n,f)$ is difficult. An approximate solution is to obtain $\mathbf{A}(f)$ first using, e.g. the techniques in Section VI-B3, and to subsequently estimate $s_j(n,f)$ in the ML sense:

$$\mathbf{s}_{\mathrm{SCA}}(n,f) = \underset{\mathbf{s}(n,f)}{\operatorname{argmax}} \sum_{jnf} \log p(s_j(n,f)) \tag{69}$$

under the constraint that $\mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{s}(n,f)$. Due to the sparse distribution used, this objective has been denoted sparse component analysis (SCA). In the case when the generalized Gaussian distribution (66) is used, this amounts to minimizing the sum over all sources of the q-th power of the ℓ_q norm of each source

$$\left(\sum_{nf} |s_j(n,f)|^q\right)^{1/q}. \tag{70}$$

The solution cannot be found in closed-form and requires an iterative algorithm in the general case [164], [182]. However, if the shape parameter q is small enough, the corresponding distribution is so sparse that it forces $J-I$ sources to zero in each time-frequency bin and only the remaining I sources indexed by $j \in \mathcal{J}(n,f)$ are nonzero [164]. The nonzero source STFT coefficients are found by *local inversion* of the mixing process, i.e., $[s_j(n,f)]_{j\in\mathcal{J}(n,f)} = [\mathbf{a}_j(n,f)]^{-1}_{j\in\mathcal{J}(n,f)}\mathbf{x}(n,f)$ [72]. This can be interpreted as LCMV beamforming similarly to above. The value of the noise covariance matrix $\mathbf{\Sigma}_{\mathbf{u}}(f)$ does not matter here since the matrix $\mathbf{A}(f) = [\mathbf{a}_j(n,f)]_{j\in\mathcal{J}(n,f)}$ is invertible. An alternative approach that forces certain source STFT coefficients to zero based on the theoretical framework of co-sparsity was proposed in [183].

3) Non i.i.d. models: The assumptions of independence and identical distribution behind ICA and SCA are major limitations: contrary to traditional beamforming approaches based on second-order statistics, they ignore the fact that audio sources exhibit patterns over time and frequency. A few approaches have attempted to relax these two assumptions. The TRINICON framework [184] and the earlier framework in [185] relax the second assumption: the source signals are assumed to be independently distributed according to a sparse distribution but the parameters of this distribution vary over time. Independent vector analysis (IVA) [186]–[188] relaxes the first assumption instead: it models the correlation between the source STFT coefficients across frequency using a multivariate sparse distribution, which results in the minimization of the sum over all sources of the q-th power of the mixed $\ell_{p,q}$ norm of each source

$$\left(\sum_{n}\left(\sum_{f}|s_j(n,f)|^p\right)^{q/p}\right)^{1/q}. \tag{71}$$

This model provides a principled approach to solving the permutation problem of ICA. Mixed norms have also been used for underdetermined separation in [63]. However, these approaches have little been pursued due to the limited range of spectro-temporal characteristics they can model and the increased optimization difficulty.

F. Summary

In Table I a summary of important single output spatial filters, discussed above, can be found.

VI. Parameter Estimation Algorithms and Implementation

In this section we will explore some widely-used structures and estimation procedures for implementing the beamformers and the spatial filters discussed in Section V. We discuss the generalized sidelobe canceller (GSC) structure, often used for implementing MVDR and LCMV beamformers in Sec. VI-A. The estimation of the speech presence probability (SPP), the (spatial) second-order statistics of the various signals, and the RTFs of the signals of interest are discussed in Sec. VI-B. Although, traditionally, the extraction of geometry information

Beamformer	Criterion	Solution	# Hard Constraints	Variants
MWF	(39)	(40)	-	**SDW-MWF** (42), **MO-MWF** (37)(63)
MVDR	(44)	(43)	1	**MPDR** (46), **MSNR** (48)
LCMV	(55)	(53)	J_p	**LCMP** (54)
ICA	(68)	no closed form	-	**TRINICON**, **IVA**

Table I
SUMMARY OF BEAMFORMERS FOR SPEECH ENHANCEMENT.

and signals' activity patterns were only used by microphone array processing methods, in recent years they were also adopted by the BSS community. We will elaborate on the differences and similarities of these paradigms in Sec. IX-A. Numerous statistical estimation criteria for estimating the various components of the spatial filters, such as maximum likelihood (ML), maximum a posteriori (MAP), and variational Bayes (VB), are discussed in Sec. VI-C.

A. The generalized sidelobe canceller

In its most general form the LCMV beamformer optimizes the following criterion (see also (55)):

$$\mathbf{w}_{\text{LCMV}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \mathbf{w}^H \mathbf{\Sigma}_{\mathbf{u}} \mathbf{w} \text{ s.t. } \check{\mathbf{A}}^H \mathbf{w} = \mathbf{q} \right\} \tag{72}$$

where $\check{\mathbf{A}}$ is a general constraint matrix (not necessarily equal to the source ATFs) and $\mathbf{q}$ is the desired response. The criterion in (72) minimizes the noise at the beamformer output subject to a set of linear constraints. The multiple constraint set generalizes the simpler MVDR criterion to allow for further control on the beampattern, beyond the response towards the array look-direction. Several alternatives for constraint selection are listed in [140], including beam derivative constraint [189], eigenvector constraint [190] and volume constraint [191]. Since adaptive constrained minimization can be a cumbersome task (see e.g. [192]) it was proposed in [193] to decompose the MVDR beamformer into separate (and orthogonal) beamformers responsible for satisfying the constraint and for noise power minimization. The resulting structure is called GSC. While the existence of such a decomposition was only proven for the MVDR beamformer in [193], it was later extended to the more general LCMV beamformer in several publications. A short and elegant proof that all LCMV beamformers can be decomposed into a GSC structure is given in [194].

The LCMV beamformer for an arbitrary constraint matrix and a desired response vector $\mathbf{q}$ is given by:

$$\mathbf{w}_{\text{LCMV}} = \mathbf{\Sigma}_{\mathbf{u}}^{-1} \check{\mathbf{A}} \left(\check{\mathbf{A}}^H \mathbf{\Sigma}_{\mathbf{u}}^{-1} \check{\mathbf{A}} \right)^{-1} \mathbf{q}. \tag{73}$$

Now, the beamformer can be recast as a sum of two orthogonal beamformers:

$$\mathbf{w}_{\text{LCMV}} = \mathbf{w}_0 - \mathbf{w}_n \tag{74}$$

where $\mathbf{w}_0 \in \text{Span}\{\check{\mathbf{A}}\}$, $\mathbf{w}_n \in \text{Null}\{\check{\mathbf{A}}\}$, and $\text{Span}\{\check{\mathbf{A}}\}$ and $\text{Null}\{\check{\mathbf{A}}\}$ are respectively the column space and the null space of the constraint matrix $\check{\mathbf{A}}$. Such an orthogonal decomposition always exists [195]. The rank of the $\text{Span}\{\check{\mathbf{A}}\}$ is J_p and the rank of $\text{Null}\{\check{\mathbf{A}}\}$ is $I - J_p$. Any vector in $\text{Null}\{\check{\mathbf{A}}\}$ can be further decomposed as $\mathbf{w}_n = \mathbf{B}\mathbf{g}$. The columns of the $I \times (I - J_p)$ matrix $\mathbf{B}$ span $\text{Null}\{\check{\mathbf{A}}\}$ and $\mathbf{g}$ is a $(I - J_p) \times 1$ weight vector. The matrix $\mathbf{B}$ is usually referred to as the blocking matrix (BM), as it blocks all constrained signals.

Using this decomposition, the output of the beamformer is given by:

$$\widehat{d} = \mathbf{w}^H \mathbf{x} = \quad \mathbf{w}_0^H \mathbf{x} - \mathbf{g}^H \underbrace{\mathbf{B}^H \mathbf{x}}_{\mathbf{e}}. \tag{75}$$

The signals $\mathbf{e} = \mathbf{B}^H \mathbf{x}$, usually referred to as *noise reference* signals, lie in $\text{Null}\{\check{\mathbf{A}}\}$, i.e. they comprise noise-only components.

The GSC implementation hence consists of two branches, as depicted in Fig. 8. The upper branch is responsible for satisfying the constraint set, and is usually denoted FBF. It should, however, be stressed that in some scenarios the constraint matrix is time-varying, e.g. when the sources are free to move. Even in such scenarios, the term FBF, although inaccurate, will still be used. A widely-used FBF is the perpendicular to the constraint set:

$$\mathbf{w}_0 = \check{\mathbf{A}} \left(\check{\mathbf{A}}^H \check{\mathbf{A}} \right)^{-1} \mathbf{q}. \tag{76}$$

Other alternatives will be discussed later.

A straightforward implementation of the BM is given by selecting the first $I - J_p$ columns of the projection matrix to the null subspace of $\check{\mathbf{A}}$, given by:

$$\mathbf{B} = \left(\mathbf{I}_{I \times I} - \check{\mathbf{A}} \left(\check{\mathbf{A}}^H \check{\mathbf{A}} \right)^{-1} \check{\mathbf{A}}^H \right) \begin{bmatrix} \mathbf{I}_{(I-J_p) \times (I-J_p)} \\ \mathbf{0}_{J_p \times (I-J_p)} \end{bmatrix}. \tag{77}$$

It is easy to verify that $\mathbf{B}^H \check{\mathbf{A}} = \mathbf{0}$.

The role of the filters $\mathbf{g}$ is to minimize the noise power at the output of the beamformer. Note that, in the ideal case, the constrained signals do not leak to the output of the BM, hence the noise power can be reduced by unconstrained minimization. The decoupling between the application of the constraint and the minimization of the noise power is the most important attribute of the GSC structure, emphasizing the importance of avoiding the leakage of the constrained signals at the output of the BM. Such a leakage contributes to the *self-cancellation* phenomenon, often resulting in desired signals distortion. The noise canceller (NC) filters $\mathbf{g}$ can be calculated using the MWF with noise reference signals $\mathbf{e}$ as inputs and the output of the FBF as the desired signal:

$$\mathbf{g} = \left(\mathbf{B}^H \mathbf{\Sigma}_{\mathbf{u}} \mathbf{B} \right)^{-1} \mathbf{B}^H \mathbf{\Sigma}_{\mathbf{u}} \mathbf{w}_0. \tag{78}$$

However, the NC is usually implemented using adaptive

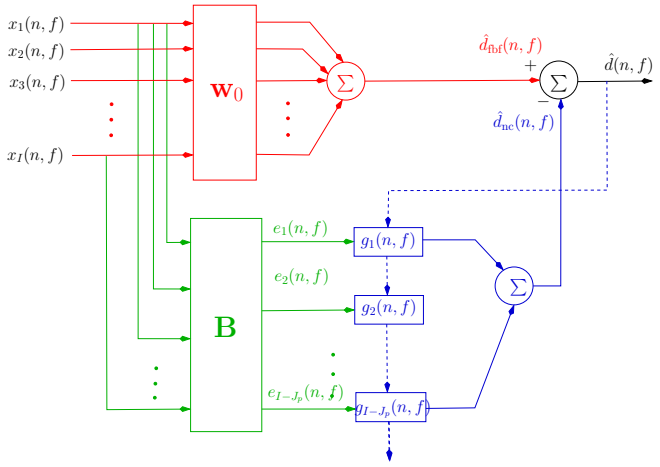


Figure 8. GSC structure for implementing the LCMV beamformer.

filters, most commonly the least mean squares (LMS) algorithm [196].

The GSC structure is widely used in the microphone array literature. In [197] a time-domain GSC is applied to enhance a desired signal impinging the array from a fixed look-direction in a car environment. In [66], a STFT-domain subspace tracking procedure is utilized for estimating the ATFs of the desired source (confined to a small predefined area), which are necessary for implementing the FBF and BM blocks. The subspace tracking procedure was later utilized to address the multiple moving sources scenario in [198]. In [199] a robust GSC is implemented in the time domain, employing adaptive BM, constrained to a predefined DOA, and norm-constrained NC. A frequency-domain implementation of an equivalent structure is given in [200]. A GSC beamformer implemented in the STFT domain is proposed in [69] with the ATFs substituted by the RTFs that can be estimated from the received signals utilizing speech nonstationarity. This structure was extended to the two speaker case (i.e. two sources of interest: one desired and one interfering) with RTFs estimated using speech nonstationarity as well [201]. Later, the multiple speaker case was addressed in [137] with the RTFs estimated using eigenvalue decomposition (EVD). An efficient implementation of the BM with the smallest possible number of filters can be found in [202]. The case of multiple speakers is also addressed in [203] by assuming disjoint activity of the various sources in the STFT domain and applying clustering procedure to localize the sources. In [204] a generalization of both the SDW-MWF and the GSC, called spatially preprocessed SDW-MWF (SP-SDW-MWF) is proposed. Applying this generalized form offers improved robustness to errors in the estimated RTFs.

B. SPP-based second order statistics and RTF estimation

The beamformers and the other spatial filters defined in the previous sections assume that certain parameters are available for their computation, namely the RTFs of the speakers, the covariance matrices of the background noise and the speakers, and/or the cross-covariance between the mixture signals and the desired signal. Numerous methods exist for estimating these parameters. Many of them rely on estimating the SPP for determining noise and speech time-frequency bins combined with speaker classification (for the multiple speakers case) in a first stage and independently estimating the various model parameters in a second stage. In the following, we review SPP estimation, and proceed with the estimation of covariances matrices and RTFs.

1) Estimating the speech presence probability: Many speech enhancement algorithms, implemented in the STFT domain, require information regarding the temporal-spectral activity of the speech signals. Contrary to the voice activity detection (VAD) problem where low resolution is sufficient, high-resolution activity estimation in both time and frequency is required here for proper enhancement.

We first consider the estimation of the SPP in a single-speaker scenario using a single microphone and then discuss the multi-microphone scenario. In this scenario, the STFT of one of the microphone signals is given by:

$$x = c + u \tag{79}$$

where c is the spatial image of the speech source and u is the sum of all noise components. The microphone and sources indices are omitted for brevity and the time and frequency indices are (n, f) unless otherwise stated. Denote the speech activity and absence hypotheses in time-frequency bin (n, f) as $\mathcal{H}_s$ and $\mathcal{H}_u$, respectively. The problem at hand can be viewed as a classical hypothesis testing problem.

Denote the *a posteriori* SNR as:

$$\gamma(n,f) \triangleq \frac{1}{|\mathcal{N}_n| \cdot |\mathcal{F}_f|} \sum_{n' \in \mathcal{N}_n, f' \in \mathcal{F}_f} \frac{|x(n',f')|^2}{\sigma_u^2(n',f')}. \tag{80}$$

where σ_u^2 denotes the variance of the noise component and $\mathcal{N}_n$, $\mathcal{F}_f$ are sets of time and frequency indices, respectively, defining a *neighborhood* of time-frequency points around (n, f). By averaging over a neighborhood of time-frequency points, the smoothness property of speech activity is utilized to reduce fluctuations in γ, which will reduce fluctuations in the SPP, and avoid distortion artifacts at the output of the enhancement algorithms. Assuming that the STFT coefficients of speech and noise are Gaussian distributed and independent over time and frequency [205], γ approximately follows a chi-squared distribution with $r = 2|\mathcal{N}_n| \cdot |\mathcal{F}_f| c_{\text{dof}}$ degrees of freedom, where c_{dof} is a correction factor resulting from the correlation between time and frequency bins due to the STFT overlap factor and the analysis window.

Numerous methods exist for estimating the noise power spectral density σ_u^2, e.g. by conventional spectrum estimation methods during speech-free time segments, by minimum statistics [206], by the improved minima controlled recursive averaging [207] or by an improved MMSE criterion [208].

Let $p \in [0, 1]$ denote the probability that hypothesis $\mathcal{H}_s$ is true, i.e. speech is present. This probability can be calculated as

$$p \triangleq \mathbb{P}\{\mathcal{H}_s|\gamma\} = \frac{\Lambda}{1+\Lambda} \tag{81}$$

where Λ is the generalized likelihood ratio, defined as

$$\Lambda \triangleq \frac{q \cdot \mathbb{P}\left\{\gamma|\mathcal{H}_s\right\}}{(1-q) \cdot \mathbb{P}\left\{\gamma|\mathcal{H}_s\right\}} \tag{82}$$

with $q = \mathbb{P}\left\{\mathcal{H}_s\right\}$ the *a priori* probability of speech presence.

For computing Λ, we further assume that the speech power is homogeneously distributed over the neighborhood of time-frequency bin (n, f). Define the *a priori* SNR as

$$\xi \triangleq \frac{1}{|\mathcal{N}_n| \cdot |\mathcal{F}_f|} \sum_{n \in \mathcal{N}_n, f \in \mathcal{F}_f} \frac{\mathbb{E}\left\{|c(n,f)|^2\right\}}{\sigma_u^2(n,f)}. \tag{83}$$

Using these definitions and the assumption that speech and noise STFT coefficients are Gaussian distributed, it can be shown (see [209]) that Λ is given by

$$\Lambda = \frac{q}{1-q}\left(\frac{1}{1+\xi}\right)^{\frac{r}{2}} \exp\left(\frac{\xi}{1+\xi} \cdot \frac{r}{2}\gamma\right). \tag{84}$$

Substituting (84) in (81) yields

$$p = \left\{1 + \frac{1-q}{q}(1+\xi)^{\frac{r}{2}} \exp\left(-\frac{\zeta}{1+\xi}\frac{r}{2}\right)\right\}^{-1} \tag{85}$$

where

$$\zeta \triangleq \gamma\xi. \tag{86}$$

The *a priori* SNR and *a priori* SNR can be signal-dependent [210]–[212] or fixed to typical values designed to meet certain false-alarm and miss-detection rates [209].

This approach can be extended to multichannel SPP estimation under the narrowband approximation [213], where multivariate Gaussian distributions are assumed for the speech and noise components. The resulting SPP is calculated using (85) with ζ and ξ redefined as

$$\xi \triangleq \mathrm{tr}\left\{\mathbf{\Sigma}_\mathbf{u}^{-1}\mathbf{\Sigma}_\mathbf{c}\right\} \tag{87}$$

$$\zeta \triangleq \mathbf{x}^H\mathbf{\Sigma}_\mathbf{u}^{-1}\mathbf{\Sigma}_\mathbf{c}\mathbf{\Sigma}_\mathbf{u}^{-1}\mathbf{x} \tag{88}$$

where

$$\mathbf{\Sigma}_\mathbf{c} \triangleq \mathbf{A}\mathbf{\Sigma}_\mathbf{s}\mathbf{A}^H \tag{89}$$

is the covariance matrix of the images of the sources of interest.

Note that estimating SPP relies on estimates of SNR and consequently the estimated power spectral densities (PSDs) of speech and noise. Straightforward incorporation of SPP in the latter power spectral density (PSD) estimates, results in a feedback which might increase false estimation and, in severe cases, might cause the estimated SPP to converge to one of its limits (either 0 or 1) without the ability to recover. Among possible solutions to this problem (see [206], [207] and [208]) are: 1) estimating the noise PSD independently of the estimated SPP; 2) using a fixed a priori SNR and fixed a priori SPP, independent of previous data-dependent SPP estimates; 3) constraining the minimal and maximal values of the SPP, and thus effectively limiting the period contaminated by these errors; 4) incorporating spatial information on the speaker in estimating the SPP (such as coherence [214] and position [215], [216]). The latter spatial information can also be utilized to classify the active speakers in a multiple speakers scenario.

2) Estimating second order statistics: The noise covariance matrix can be estimated by recursively averaging instantaneous covariance matrices weighted according to the SPP:

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_\mathbf{u}(n,f) =& \lambda_u'(n,f)\widehat{\mathbf{\Sigma}}_\mathbf{u}(n-1,f) \\ &+ (1-\lambda_u'(n,f))\,\mathbf{x}(n,f)\mathbf{x}^H(n,f).\end{aligned} \tag{90}$$

where

$$\lambda_u'(n,f) \triangleq (1-p(n,f))\,\lambda_u + p(n,f) \tag{91}$$

is a time-varying recursive averaging factor and λ_u is selected such that its corresponding estimation period ($\frac{1}{1-\lambda_u}$ frames) is shorter than the stationarity time of the noise. Alternatively, a hard binary weighting, obtained by applying a threshold to the SPP, can be used instead of the soft weighting.

Define the hypothesis that speaker j is present as $\mathcal{H}_{s_j}(n,f)$, and its corresponding a posteriori probability as $p_j(n,f) \triangleq \mathbb{P}\left\{\mathcal{H}_{s_j}(n,f)|\mathbf{x}(n,f)\right\}$. Similarly to (90), the covariance matrix of the spatial image of source j, denoted $\mathbf{\Sigma}_{\mathbf{c}_j}(n,f) \triangleq \sigma_{s_j}^2(n,f)\mathbf{a}_j(f)\mathbf{a}_j^H(f)$, can be estimated by

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{\mathbf{c}_j}(n,f) =& \lambda_{s_j}'(n,f)\widehat{\mathbf{\Sigma}}_{\mathbf{c}_j}(n-1,f) \\ &+ (1-\lambda_{s_j}'(n,f))(\mathbf{x}(n,f)\mathbf{x}^H(n,f) - \widehat{\mathbf{\Sigma}}_\mathbf{u}(n-1,f))\end{aligned} \tag{92}$$

where

$$\lambda_{s_j}'(n,f) \triangleq (1-p_j(n,f))\,\lambda_s + p_j(n,f) \tag{93}$$

is a time-varying recursive-averaging factor, and λ_s is selected such that its corresponding estimation period ($\frac{1}{1-\lambda_s}$ frames) is shorter than the *coherence time* of the AIRs of speaker j, i.e. the time period over which the AIRs are assumed to be time-invariant. Note that: 1) usually the estimation period is longer than the speech nonstationarity time, therefore, although the spatial structure of $\mathbf{\Sigma}_{\mathbf{c}_j}(n,f)$ is maintained, the estimated variance is an average of the speech variances over multiple time periods; 2) the estimate $\widehat{\mathbf{\Sigma}}_{\mathbf{c}_j}(n,f)$ keeps its past value when speaker j is absent.

Individual SPPs for each of the speakers can be approximated from their estimated positions [215]–[217]. Given that any of the speakers is active (i.e. hypothesis $\mathcal{H}_s$ is true and hence $p(n,f)$ is high), and assuming that each time-frequency bin is dominated by at most one speaker signal (i.e. the time-frequency sparsity assumption), the a posteriori probabilities are obtained by

$$p_j(n,f) = p(n,f) \cdot \mathbb{P}\left\{\mathcal{H}_{s_j}|\mathbf{x}(n,f), \mathcal{H}_s\right\}. \tag{94}$$

Next, assuming that source position is the *sufficient statistics* embedded in $\mathbf{x}(n,f)$ for classifying the active source we obtain:

$$p_j(n,f) \approx p(n,f) \cdot \mathbb{P}\left\{\mathcal{H}_{s_j}|\widehat{\mathbf{r}}(n,f), \mathcal{H}_s\right\} \tag{95}$$

where $\widehat{\mathbf{r}}(n,f)$ denotes position estimate of source active in time-frequency bin (n,f). A plethora of methods exist for estimating source positions, however, this topic is beyond the scope of this overview paper. By adopting a Gaussian model

for the error of the estimated position, and by applying Bayes rule, the SPP in (95) can be reformulated as

$$p_j(n,f) = p(n,f) \cdot \frac{\pi_{s,j}\mathcal{N}\left(\widehat{\mathbf{r}}|\boldsymbol{\mu}_j^r, \boldsymbol{\Sigma}_j^r\right)}{\sum_{j'}\pi_{s,j'}\mathcal{N}\left(\widehat{\mathbf{r}}|\boldsymbol{\mu}_{j'}^r, \boldsymbol{\Sigma}_{j'}^r\right)} \tag{96}$$

where $\mathcal{N}$ denotes the Gaussian distribution and $\pi_{s,j}$, $\boldsymbol{\mu}_j^r$, $\boldsymbol{\Sigma}_j^r$ are the prior probability, the mean and the covariance matrix of the position of speaker j, respectively, for $j = 1, \ldots, J_p$. The parameters J_p, $\pi_{s,j}$, $\boldsymbol{\mu}_j^r$, $\boldsymbol{\Sigma}_j^r$ for all j are estimated by an expectation-maximization (EM) algorithm. The individual SPP estimates can also utilize DRR estimates [218] affected by the proximity of the sources to the microphone array.

3) Estimating the relative transfer function: Two common approaches for RTF estimation are the covariance subtraction (CS) [154], [219] and the covariance whitening (CW) [137], [220] methods. Here, for brevity we assume a single speaker scenario. Both of these approaches rely on estimated noisy speech and noise-only covariance matrices, i.e. $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{u}}$. Given the estimated covariance matrices, CS estimates the speaker RTF by

$$\tilde{\mathbf{a}}_{\text{CS}} \triangleq \frac{1}{\mathbf{i}_1^H(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{u}})\mathbf{i}_1}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{u}})\mathbf{i}_1 \tag{97}$$

where $\mathbf{i}_1 = \begin{bmatrix} 1 & \mathbf{0}_{1\times I-1} \end{bmatrix}^T$ is an $I \times 1$ selection vector for extracting the component of the reference microphone, here assumed to be the first micrphone. The CW approach estimates the RTF by: 1) applying the generalized eigenvalue decomposition (GEVD) to $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$ with $\widehat{\boldsymbol{\Sigma}}_{\mathbf{u}}$ as the whitening matrix; 2) de-whitening the eigenvector corresponding to the strongest eigenvalue, denoted $\tilde{\mathbf{a}}_{\mathbf{u}}$, namely $\widehat{\boldsymbol{\Sigma}}_{\mathbf{u}}^{-1}\tilde{\mathbf{a}}_{\mathbf{u}}$; 3) normalizing the de-whitened eigenvector by the reference microphone component. Explicitly:

$$\tilde{\mathbf{a}}_{\text{CW}} \triangleq (\mathbf{i}_1^H\widehat{\boldsymbol{\Sigma}}_{\mathbf{u}}^{-1}\tilde{\mathbf{a}}_{\mathbf{u}})^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{u}}^{-1}\tilde{\mathbf{a}}_{\mathbf{u}}. \tag{98}$$

A preliminary analysis and comparison of the CS and CW methods can be found in [221].

Alternative methods utilize the speech nonstationarity property, assuming that the noise has slow time-varying statistics. In [69], the problem of estimating the RTF of microphone i is formulated as a LS problem where the l-th equation utilizes $\widehat{\sigma}_{x_i x_1}^l$, the estimated cross-PSD of microphone i and the reference microphone in the l-th time segment. This cross-PSD satisfies:

$$\widehat{\sigma}_{x_i x_1}^l = \tilde{a}_i(\widehat{\sigma}_{x_1}^l)^2 + \widehat{\sigma}_{u_i x_1}^l + \epsilon_i^l \tag{99}$$

where we use the relation $\mathbf{x} = \tilde{\mathbf{a}}x_1 + \mathbf{u}$. The unknowns are $\tilde{a}_i$, i.e. the required RTF, and $\widehat{\sigma}_{u_i x_1}^l$, which is a nuisance parameter. ϵ_i^l denotes the error term of the l-th equation. Multiple LS problems, one for each microphone, are solved for estimating the vector RTF. Note that, the latter method, also known as the *nonstationarity*-based RTF estimation, does not require a prior estimate of the noise covariance, since it simultaneously solves for RTF and the noise statistics. Similarly, a weighted least squares (WLS) problem with exponential weighting can be defined and implemented using a recursive least squares (RLS) algorithm [138]. Considering speech sparsity in the STFT domain, in [219] the SPPs were incorporated into the weights of the WLS problem, resulting in a more accurate solution. Sparsity of speech signals in the frequency domain increases the convergence time of RTF estimation methods, until sufficient signal energy is collected in the entire band. In [60] time-domain sparsity of the RTF is utilized for reducing convergence time, by interpolating missing spectral components.

In [85], [87], the RTFs of multiple speakers are obtained by clustering ILD and ITD information across all time-frequency bins. This approach is refined in [71], [72] by detecting single-source time intervals using a rank criterion, estimating the RTFs or ATFs in each time interval, and clustering these estimates to obtain a single estimate per source. A variant of the latter technique is applied in [54] for anechoic mixtures. A similar approach is applied in [86], [169] where mixtures of STFT coefficients, normalized as in (10), are clustered. The largest clusters are then used for obtaining the RTF estimates. Beyond clustering, a second step of resolving permutation across frequencies is required. In [222] the clustering and permutation alignment are performed in a single step. In [223], rather than clustering and classification, RTFs are estimated based on instantaneous observation vectors projected to the signals subspace constructed by smoothing past observation vectors. An analysis and evaluation of ICA methods for RTF estimation is available in [224].

C. EM, VB, and MM

In contrast with SPP-based approaches which independently estimate the model parameters, some approaches jointly estimate all parameters according to some criterion. Among them, early approaches to ICA were based on time-delayed decorrelation [68], [225] or quadratic spatial contrasts [70], [226], which inspired ML-based approaches [53], [227]. Many approaches based on ML and alternative statistical estimation criteria such as maximum a posteriori (MAP) and variational Bayes (VB) have then been proposed [53], [110]–[113], [115], [116], [159], [227]–[237]. In many cases, the resulting optimization problems cannot be solved in closed-form. General nonlinear optimization techniques such as gradient ascent or the Newton method are impractical due to the large number of parameters. The EM algorithm [238] is a popular optimization method which iteratively breaks down the problem into several smaller optimization problems involving subsets of parameters, which are solved separately using closed-form updates. Many variants of EM have been proposed which, e.g., break down the problem in a different way or use nonlinear optimization techniques to solve each subproblem. We do not detail all of them here but rather describe the main criteria and illustrate them in the case of the LGM in Section III-E and the binary activation model in Section V-E.

1) ML and MAP criteria: Let us denote by $\boldsymbol{\theta}$ the set of model parameters. When no prior information about the model parameters is given, $\boldsymbol{\theta}$ is often estimated in the ML sense as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, p(\mathcal{X}|\boldsymbol{\theta}). \tag{100}$$

The likelihood $p(\mathcal{X}|\boldsymbol{\theta})$ can be expressed as

$$p(\mathcal{X}|\boldsymbol{\theta}) = \int p(\mathcal{X}|\mathcal{C},\boldsymbol{\theta})p(\mathcal{C}|\boldsymbol{\theta})d\mathcal{C} \tag{101}$$

where $\mathcal{X} = \{\mathbf{x}(n,f)\}_{nf}$ and $\mathcal{C} = \{\mathbf{c}_j(n,f)\}_{jnf}$ denote the set of STFT coefficients of the mixture and the spatial images of all sources, respectively. The set of possible parameter values $\boldsymbol{\Theta}$ may be the full parameter space or incorporate prior knowledge by means of deterministic constraints as in [113], [115].

For example, in the case of the LGM, $\boldsymbol{\theta}$ may consist of the spatial covariances and the source variances,

$$\boldsymbol{\theta} = \left\{\{\mathbf{R}_j(f)\}_{jf}, \{\sigma^2_{s_j}(n,f)\}_{jnf}\right\}. \tag{102}$$

The prior distribution of the source spatial images is given by $p(\mathcal{C}|\boldsymbol{\theta}) = \prod_{jnf} p(\mathbf{c}_j(n,f)|\boldsymbol{\Sigma}_{\mathbf{c}_j}(n,f))$ in (18), and $p(\mathcal{X}|\mathcal{C},\boldsymbol{\theta})$ is the Dirac distribution corresponding to the mixing equation (2). The likelihood is then expressed as

$$p(\mathcal{X}|\boldsymbol{\theta}) = \prod_{nf} p(\mathbf{x}(n,f)|\boldsymbol{\Sigma}_{\mathbf{x}}(n,f)) \tag{103}$$

with $\boldsymbol{\Sigma}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} \boldsymbol{\Sigma}_{\mathbf{c}_j}(n,f)$.

When some prior knowledge about the model parameters is provided via a prior distribution, e.g., $p(\boldsymbol{\theta}) = \prod_{jf} p(\mathbf{R}_j(f))$ with $p(\mathbf{R}_j(f))$ given by (20), the MAP criterion may be used instead:

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{X}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}).\end{aligned} \tag{104}$$

Note that in practice $p^{\gamma}(\boldsymbol{\theta})$ is usually considered in the above equation rather than $p(\boldsymbol{\theta})$ where the parameter γ controls the strength of the prior [116]. Note also that the MAP criterion generalizes the ML criterion, since (104) reduces to (100) when a non-informative uniform prior $p(\boldsymbol{\theta}) \propto 1$ is considered. In the following, we therefore formulate the EM algorithm in its most general form for the MAP criterion.

2) EM and GEM algorithms: Let

$$\mathcal{L}(\boldsymbol{\theta}) \triangleq \log[p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})] \tag{105}$$

be the logarithm of the MAP objective. In most cases maximizing $\mathcal{L}(\boldsymbol{\theta})$ has no closed-form solution. An intuition behind the EM algorithm is as follows. Given that $\mathcal{L}(\boldsymbol{\theta})$ is difficult to optimize, it is possible, for a range of models, to consider a set of unknown data $\mathcal{Z}$ called *latent data* such that replacing the observed data likelihood $p(\mathcal{X}|\boldsymbol{\theta})$ in (105) by the *complete data* likelihood $p(\mathcal{X},\mathcal{Z}|\boldsymbol{\theta})$ makes the optimization much easier. Since the value of the latent data is unknown, the complete data log-likelihood $\log p(\mathcal{X},\mathcal{Z}|\boldsymbol{\theta})$ is replaced by its expectation over $\mathcal{Z}$ given the current model parameters and the measurements $\mathcal{X}$.

More precisely the EM algorithm consists in iterating several times the following two steps [238]:

- **E-step**: Compute an auxiliary function as

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(l)}) = \mathbb{E}_{\mathcal{Z}|\mathcal{X},\boldsymbol{\theta}^{(l)}} \log p(\mathcal{X},\mathcal{Z}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}). \tag{106}$$

- **M-step**: Update the model parameters as the maximum of the auxiliary function:

$$\boldsymbol{\theta}^{(l+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(l)}) \tag{107}$$

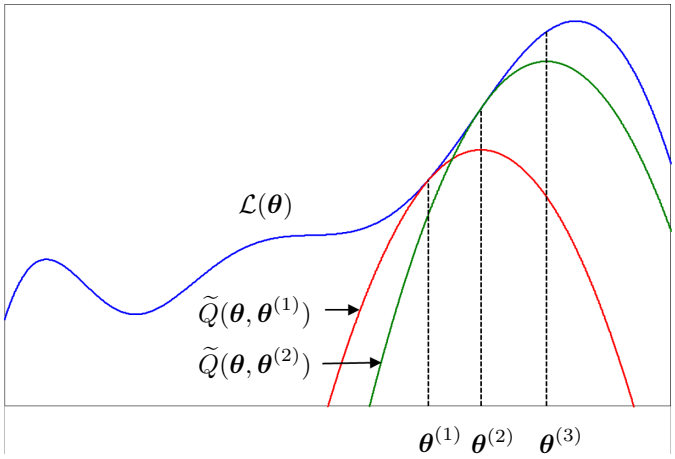


Figure 9. Graphical illustration of the EM algorithm.

where $\boldsymbol{\theta}^{(l)}$ denotes the estimated model parameters at the l-th iteration of the algorithm. Let us add to the auxiliary function $Q(\boldsymbol{\theta},\boldsymbol{\theta}')$ an additional term that is independent on $\boldsymbol{\theta}$ and thus does not change its optimization in (107):

$$\widetilde{Q}(\boldsymbol{\theta},\boldsymbol{\theta}') = Q(\boldsymbol{\theta},\boldsymbol{\theta}') - \mathbb{E}_{\mathcal{Z}|\mathcal{X},\boldsymbol{\theta}'} \log p(\mathcal{Z}|\mathcal{X},\boldsymbol{\theta}'). \tag{108}$$

The auxiliary function $\widetilde{Q}(\boldsymbol{\theta},\boldsymbol{\theta}')$ is proven [238] to satisfy

$$\widetilde{Q}(\boldsymbol{\theta},\boldsymbol{\theta}') \leq \mathcal{L}(\boldsymbol{\theta}) \quad \text{and} \quad \widetilde{Q}(\boldsymbol{\theta}',\boldsymbol{\theta}') = \mathcal{L}(\boldsymbol{\theta}'), \tag{109}$$

i.e., it is a lower bound of $\mathcal{L}(\boldsymbol{\theta})$ that is tight at the current estimate $\boldsymbol{\theta}'$. These properties are enough to prove that the cost function $\mathcal{L}(\boldsymbol{\theta})$ is non-decreasing under the update (107), i.e., $\mathcal{L}(\boldsymbol{\theta}^{(l+1)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(l)})$. This can be intuitively understood from the illustration in Fig. 9. As any other nonlinear optimization strategy, the EM algorithm does not guarantee convergence to a global maximum. Providing an appropriate initialization of the parameters $\boldsymbol{\theta}$ is therefore very important.

In the case when the M-step is not tractable in closed-form, one can replace (107) by any update such that $Q(\boldsymbol{\theta}^{(l+1)},\boldsymbol{\theta}^{(l)}) \geq Q(\boldsymbol{\theta}^{(l)},\boldsymbol{\theta}^{(l)})$. This algorithm still guarantees that $\mathcal{L}(\boldsymbol{\theta})$ is non-decreasing over the iterations and is referred to as generalized expectation-maximization (GEM) algorithm [238]. These updates may result from gradient ascent, the Newton method, or explicit optimization over a discretized set, for instance [159], [239].

In the particular case when the complete data distribution belongs to the so-called *exponential family* of distributions [238], [240], the EM or GEM algorithm can be reformulated as computing the conditional expectation of the *sufficient statistics* representing the distribution (E-step) and maximizing the complete data posterior as a function of these statistics (M-step). With Gaussian or discrete models, the sufficient statistics are typically zeroth-, first-, and second-order moments. For more details, see [238], [240]. Although

this reformulation does not change the final algorithm, it can simplify its derivation. Most EM algorithms considered in the literature and all the EM algorithms considered here fall into this particular case.

In summary, a particular EM algorithm and its output depend on many factors such as the estimation criterion (ML or MAP), the choice of latent data, the auxiliary function update strategy in the case of the GEM algorithm, the parameter initialization, and the number of iterations. To illustrate the variety of possible EM/GEM implementations, we detail below three algorithms for the LGM in Section III-E and the binary activation model in Section V-E which differ by the choice of the latent data. We use the ML criterion (100) and assume for simplicity that both the spatial covariance matrices $\mathbf{R}_j(f)$ and the source variances $\sigma_{s_j}^2(n,f)$ are unconstrained. Adding deterministic or probabilistic constraints on $\mathbf{R}_j(f)$ and/or $\sigma_{s_j}^2(n,f)$ would only affect the M-step. For examples of modified M-step updates resulting from such constraints, see [115], [116] and [228], [232], [234], respectively. The modifications to the M-step resulting from constraints on the source variances are also briefly addressed in Section VII-C.

3) Source spatial image EM: Let us consider the LGM in (18)–(19). A first approach [113] which is applicable when $\mathbf{R}_j(f)$ is full-rank is to consider the source spatial images $\mathbf{c}_j(n,f)$ as latent data. One iteration of the resulting GEM algorithm[3] can be written as follows:

- **E-step**: Compute the expected sufficient statistics

$$\widehat{\mathbf{c}}_j(n,f) = \mathbf{W}_j^H(n,f)\mathbf{x}(n,f) \tag{110}$$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{c}_j}(n,f) = \widehat{\mathbf{c}}_j(n,f)\widehat{\mathbf{c}}_j^H(n,f) + (\mathbf{I} - \mathbf{W}_j^H(n,f))\mathbf{\Sigma}_{\mathbf{c}_j}(n,f) \tag{111}$$

  where $\mathbf{I}$ is the $I \times I$ identity matrix and $\mathbf{W}_j(n,f)$ is the MWF defined in (63).
- **M-step**: Update the model parameters

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{\sigma_{s_j}^2(n,f)} \widehat{\mathbf{\Sigma}}_{\mathbf{c}_j}(n,f) \tag{112}$$

$$\sigma_{s_j}^2(n,f) = \frac{1}{I}\operatorname{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_{\mathbf{c}_j}(n,f)) \tag{113}$$

Note that spatial filtering is performed in (110) as part of the E-step. This can be thought of as a feedback loop from spatial filtering to parameter estimation, as illustrated in Fig. 1.

4) Subsource EM: A second approach [228] is based on the observation that spatial covariance matrices $\mathbf{R}_j(f)$ of arbitrary rank R can be non-uniquely represented as

$$\mathbf{R}_j(f) = \mathbf{A}_j(f)\mathbf{A}_j^H(f) \tag{114}$$

where $\mathbf{A}_j(f)$ is an $I \times R$ complex-valued matrix. The source spatial image $\mathbf{c}_j(n,f)$ can then be expressed as

$$\mathbf{c}_j(n,f) = \mathbf{A}_j(f)\mathbf{z}_j(f) \tag{115}$$

where $\mathbf{z}_j(n,f) = [z_{j1}(n,f),\ldots,z_{jR}(n,f)]^T$ is an $R \times 1$ vector of *subsource* STFT coefficients. Assuming that $z_{jr}(n,f)$ is Gaussian distributed with zero mean and variance $\sigma_{s_j}^2(n,f)$ for all r, it can be verified that the covariance of $\mathbf{c}_j(n,f)$ equals the expression in (19). By concatenating these quantities into an $I \times RJ$ matrix $\mathbf{A}(f) = [\mathbf{A}_1(f),\ldots,\mathbf{A}_J(f)]$ and an $RJ \times 1$ vector $\mathbf{z}(n,f) = [\mathbf{z}_1(n,f)^T,\ldots,\mathbf{z}_J(n,f)^T]^T$ and including an additive noise term $\mathbf{u}(n,f)$[4], one can rewrite the mixing equation as

$$\mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{z}(n,f) + \mathbf{u}(n,f). \tag{116}$$

This reformulation is therefore equivalent to the original LGM formulation.

Considering the subsources $\mathbf{z}(n,f)$ as latent data, an EM algorithm was derived in [228]. One iteration of this algorithm can be written as:

- **E-step**: Compute the expected sufficient statistics

$$\widehat{\mathbf{z}}(n,f) = \mathbf{W}^H(n,f)\mathbf{x}(n,f) \tag{117}$$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{xz}}(n,f) = \mathbf{x}(n,f)\widehat{\mathbf{z}}^H(n,f) \tag{118}$$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{zz}}(n,f) = \widehat{\mathbf{z}}(n,f)\widehat{\mathbf{z}}^H(n,f) + (\mathbf{I} - \mathbf{W}^H(n,f)\mathbf{A}(f))\mathbf{\Sigma}_{\mathbf{z}}(n,f) \tag{119}$$

  where $\mathbf{W}(n,f)$ is the multi-output MWF defined in (37)

$$\mathbf{W}(n,f) = \mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{A}(f)\mathbf{\Sigma}_{\mathbf{z}}(n,f) \tag{120}$$

  with

$$\mathbf{\Sigma}_{\mathbf{x}}(n,f) = \mathbf{A}(f)\mathbf{\Sigma}_{\mathbf{z}}(n,f)\mathbf{A}^H(f) + \mathbf{\Sigma}_{\mathbf{u}}(f) \tag{121}$$

$$\mathbf{\Sigma}_{\mathbf{z}}(n,f) = \operatorname{diag}(\underbrace{[\sigma_{s_j}^2(n,f),\ldots,\sigma_{s_j}^2(n,f)]}_{R \text{ times}}{}_{j=1}^{J}) \tag{122}$$

- **M-step**: Update the model parameters

$$\mathbf{A}(f) = \left(\sum_{n=1}^{N} \widehat{\mathbf{\Sigma}}_{\mathbf{xz}}(n,f)\right)\left(\sum_{n=1}^{N} \widehat{\mathbf{\Sigma}}_{\mathbf{zz}}(n,f)\right)^{-1} \tag{123}$$

$$\sigma_{s_j}^2(n,f) = \frac{1}{R}\sum_{k=(j-1)R+1}^{jR} \left(\widehat{\mathbf{\Sigma}}_{\mathbf{zz}}(n,f)\right)_{kk} \tag{124}$$

5) Binary activation EM: Let us now consider the binary activation model in (64). As explained in Section V-E, given the index $j^\star(n,f)$ of the active source and the model parameters $\boldsymbol{\theta}$, the optimal value of the predominant source $s_j^\star(n,f)$ is obtained by the MVDR beamformer. The log-likelihood then simplifies to

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{nf} -\log\det(\pi\mathbf{\Sigma}_{\mathbf{u}}(f)) - \mathbf{x}^H(n,f)\mathbf{\Sigma}_{\mathbf{u}}^{-1}(f)\mathbf{x}(n,f) + \frac{|\mathbf{a}_j^{\star H}(f)\mathbf{\Sigma}_{\mathbf{u}}^{-1}(f)\mathbf{x}(n,f)|^2}{\mathbf{a}_j^{\star H}(f)\mathbf{\Sigma}_{\mathbf{u}}^{-1}(f)\mathbf{a}_j^\star(f)} \tag{125}$$

Considering the indexes $j^\star(n,f)$ of the active sources as latent data, the following EM algorithm can be derived:

- **E-step**: Compute the posterior probability of $j^\star(n,f)$

[3] This is indeed a GEM algorithm since a single iteration of the updates (112) and (113) does not lead to the maximum of (107).

[4] This additive noise term is necessary here, otherwise the complete data likelihood becomes singular and $\mathbf{A}(f)$ remains stuck to its initial value [112]. In practice, the covariance of $\mathbf{u}(n,f)$ is assumed to be diagonal $\mathbf{\Sigma}_{\mathbf{u}}(f) = \sigma_u^2(f)\mathbf{I}$ and it is decreased over the iterations in an annealing fashion.

$$\gamma_j(m,k) \propto \exp\left(\frac{|\mathbf{a}_j^H(f)\mathbf{\Sigma}_\mathbf{u}^{-1}(f)\mathbf{x}(n,f)|^2}{\mathbf{a}_j^H(f)\mathbf{\Sigma}_\mathbf{u}^{-1}(f)\mathbf{a}_j(f)}\right) \quad (126)$$

- **M-step**: Update the model parameters

$$\mathbf{a}_j(f) = \frac{\sum_{n=1}^N \gamma_j(n,f) s_j^*(n,f)\mathbf{x}(n,f)}{\sum_{n=1}^N \gamma_j(n,f)|s_j(n,f)|^2} \quad (127)$$

$$\mathbf{\Sigma}_\mathbf{u}(f) = \frac{1}{N}\sum_{j=1}^J \sum_{n=1}^N \gamma_j(n,f)(\mathbf{x}(n,f) - \mathbf{a}_j(f)\, s_j(n,f)) (\mathbf{x}(n,f) - \mathbf{a}_j(f)\, s_j(n,f))^H \quad (128)$$

with $s_j(n,f)$ updated by MVDR beamforming (43) given $\mathbf{a}_j(f)$ and $\mathbf{\Sigma}_\mathbf{u}(f)$.

All the above EM algorithms are usually referred to as *batch* algorithms, since they are exploiting all the signal samples at once. In contrast, algorithms exploiting only the current and previous audio samples are referred as *online*, and they become crucial for many practical applications such as, e.g. real-time signal separation on a portable device. Online variants of these algorithms were considered in [241], [242] based on the theory in [243]–[245]. These approaches rely either on computing expectations of sufficient statistics by averaging them over time with some forgetting factor [241] and/or by recomputing them from the most recent block of time frames [242]. Online EM algorithms for related problems were also introduced in [246], [247].

6) VB criterion and algorithm: In contrast to ML/MAP, the VB criterion [248] does not rely on finding a point estimate of the model parameters $\boldsymbol{\theta}$, but consists in computing directly the posterior distribution of the source STFT coefficients while marginalizing over all possible model parameters:

$$p(\mathcal{C}|\mathcal{X}) = \int p(\mathcal{C},\boldsymbol{\theta}|\mathcal{X}) d\boldsymbol{\theta}. \quad (129)$$

This leads to more accurate estimation, since the point estimate $\widehat{\boldsymbol{\theta}}$, as in the ML and MAP criteria (100), (104), is replaced by its posterior distribution $p(\boldsymbol{\theta}|\mathcal{X})$.

The computation of the integral in (129) is intractable. To overcome this difficulty, variational approximations [248] are usually applied. They consist in replacing the true posterior distribution $p(\mathcal{C},\boldsymbol{\theta}|\mathcal{X})$ by a factored approximation $q(\mathcal{C},\boldsymbol{\theta}) = q(\mathcal{C})q(\boldsymbol{\theta})$ [237]. The integral is then simply computed as $p(\mathcal{C}|\mathcal{X}) \approx q(\mathcal{C})$. The optimal factored approximation is obtained by minimizing the Kullback-Leibler (KL) divergence with the true distribution:

$$\operatorname*{argmin}_q \mathrm{KL}(q(\mathcal{C},\boldsymbol{\theta})\|p(\mathcal{C},\boldsymbol{\theta}|\mathcal{X})) = \operatorname*{argmax}_q \int q(\mathcal{C},\boldsymbol{\theta}) \log \frac{p(\mathcal{C},\boldsymbol{\theta}|\mathcal{X})}{q(\mathcal{C},\boldsymbol{\theta})} d\mathcal{C} d\boldsymbol{\theta} \quad (130)$$

A VB algorithm that iteratively optimizes this objective was proposed in [231], [237] for the LGM. This algorithm is similar to EM except for the following difference: while EM alternatively estimates the posterior distribution of the latent data and the parameter values $\widehat{\boldsymbol{\theta}}$, VB alternatively estimates the posterior distribution of the latent data and the posterior distribution $q(\boldsymbol{\theta})$ of the parameters. The latter approach is less sensitive to local maxima and overfitting in theory. Finally, note that variational approximations are not only used to optimize the VB criterion, but also can be employed to reduce the computational complexity of a classical EM algorithm [105].

7) MM algorithms: EM is a special case of a more general optimization strategy named minorization-maximization (MM) [249]. The MM principle consists in iteratively constructing and maximizing an auxiliary function $Q(\boldsymbol{\theta},\boldsymbol{\theta}')$ that is required to satisfy (109). However, in contrast to the EM algorithm, the auxiliary function does not need to be constructed as in (106). As such, the MM algorithm leads to a broader family of updates than EM/GEM and it can even be applied to other estimation criteria than ML, MAP or VB. In particular, it was applied to rank-1 and full-rank LGM in [115], [168], [188], [232].

D. Other nonconvex optimization algorithms

The algorithms we have seen so far are applicable to energy-based estimation criteria or Gaussian models, which translate into estimating the second-order statistics of the signals. Nongaussianity-based ICA and its extensions such as TRINICON and IVA (see Sections V-E2 and V-E3) stand apart: due to the assumed continuous sparse distribution, the resulting beamformers cannot be expressed in terms of the second-order statistics of the signals and they cannot even be computed in closed-form. General nonlinear optimization algorithms such as gradient ascent must then be employed. In practice, the data is first whitened and one then searches for a unitary demixing matrix using so-called natural gradient ascent [165], [171], [177], [178], [184]. The same problem arises with SCA for under-determined mixtures, which requires the minimization of the ℓ_p norm or the mixed $\ell_{p,q}$ norm of complex-valued data. For general p and q, gradient descent and pseudo-Netwon techniques were used in [164], [182]. For $p,q \in \{1,2\}$, sparse decomposition algorithms based on proximal gradient [61], [63], reweighted ℓ_1 [183], or greedy methods such as basis pursuit [169] are generally preferred.

VII. Postfiltering, Masking and Joint Spatial-Spectral Estimation

The performance of certain beamformers is limited when the undesired signals are not point sources or when there are too many interfering sources. Moreover, some beamformers suffer from the existence of nonstationary interference, due to the larger observation time required to estimate signal statistics. Single-channel enhancement methods can achieve nonlinear spatial and/or spectral filtering and usually adapt much faster to changes in the interference characteristics. In this section, we explore the use of such algorithms as postfilters applied at the output of the beamformers [27]. We then proceed by presenting single microphone separation algorithms utilizing spatial information and conclude in presenting joint spatial-spectral estimators. Beamformers with a subsequent postfiltering stage, utilizing both spatial and spectral information, adopt some of the single-channel speech separation methodologies, and therefore usually lead

to improved performance as compared with both multichannel and single-channel algorithms. Note, that some of the modern multichannel techniques, reviewed in this paper, are utilizing the entire reflection pattern of the speech propagation rather than resorting to DOA-based steering vector, and are therefore capable of separation sources with identical DOA.

A. Postfiltering

The SDW-MWF was defined in (42) and its definition is repeated here for readability:

$$\mathbf{w}_{\text{SDW-MWF}} = \frac{\sigma_{s_1}^2 \boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\mathbf{a}_1}{\mu + \sigma_{s_1}^2 \mathbf{a}_1^H \boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\mathbf{a}_1} q.$$

By selecting the desired response as $q = a_{11}^*$ the desired component at the beamformer output becomes $a_{11}s_1$.

Using the Woodbury identity [152] we can decompose the SDW-MWF, with the predefined q, into:

$$\mathbf{w}_{\text{SDW-MWF}} = \underbrace{\frac{\boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\tilde{\mathbf{a}}_1}{\tilde{\mathbf{a}}_1^H \boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\tilde{\mathbf{a}}_1}}_{\mathbf{w}_{\text{MVDR-RTF}}} \times \underbrace{\frac{\sigma_{d_s}^2}{\sigma_{d_s}^2 + \mu\sigma_{d_u}^2}}_{w_{\text{SDW-SWF}}} \tag{131}$$

where $\mathbf{w}_{\text{MVDR-RTF}}$ is the MVDR beamformer using the RTF vector and $w_{\text{SDW-SWF}}$ is a single-channel postfilter. Further define, $\sigma_{d_s}^2 = \sigma_{s_1}^2|a_{11}|^2$ as the power of desired speech component at the output of the MVDR-RTF beamformer and $\sigma_{d_u}^2 = \mathbf{w}_{\text{MVDR-RTF}}^H \boldsymbol{\Sigma}_{\mathbf{u}} \mathbf{w}_{\text{MVDR-RTF}}$ as the respective noise power. This decomposition [92], [143] constitutes the motivation for applying a linear postfilter (speech distortion weighted single-channel Wiener filter (SDW-SWF) in this case) at the output of an MVDR beamformer.

A plethora of postfilters can be found in the literature, differing in the procedures for estimating the speech and noise statistics.

Zelinski [250] proposed the following procedure, assuming that the MVDR is distortionless[5], namely $\sigma_{d_s}^2 = \sigma_{s_1}^2$. He further assumed that the noise is spatially-white[6], namely $\boldsymbol{\Sigma}_{\mathbf{u}} = \sigma_u^2\mathbf{I}$. Under these assumptions the cross-PSD between microphones $i \neq i'$ is given by $\sigma_{x_i x_{i'}} = \sigma_{s_1}^2$ and the PSD of the microphone signals is given by $\sigma_{x_i}^2 = \sigma_{s_1}^2 + \sigma_u^2$. Both PSDs can be recursively estimated from the microphone signals. The Zelinski postfilter is finally given by

$$w_{\text{Zel}} = \frac{\frac{2}{I(I-1)}\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I}\Re(\widehat{\sigma}_{x_i x_{i'}})}{\frac{1}{I}\sum_{i=1}^{I}\widehat{\sigma}_{x_i}^2}. \tag{132}$$

with $\Re(\cdot)$ the real part of a complex number, applied here to ensure real-values speech-PSD estimation. It was proposed in [251] to substitute the Wiener filter proposed by Zelinski, by a combined spectral subtraction and Wiener postfilter. These structures were further analyzed and improved in [252].

McCowan and Bourlard [142] substituted the spatially-white noise field assumption by a diffuse noise field assumption instead. Hence, $\sigma_{u_i u_{i'}} = \sigma_u^2 \Omega_{ii'}$ where $\Omega_{ii'}$ is given in (5). The auto- and cross-PSDs of the microphone signals are now given by $\sigma_{x_i x_{i'}} = \sigma_{s_1}^2 + \sigma_u^2\Omega_{ii'}$, $i \neq i'$, and $\sigma_{x_i}^2 = \sigma_{s_1}^2 + \sigma_u^2$, respectively. With these definitions, the McCowan and Bourlard postfilter is given by

$$w_{\text{MB}} = \frac{\frac{2}{I(I-1)}\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I}\widehat{\sigma}_{s_i s_{i'}}}{\frac{1}{I}\sum_{i=1}^{I}\widehat{\sigma}_{x_i}^2} \tag{133}$$

where

$$\widehat{\sigma}_{s_i s_{i'}} = \frac{\Re(\widehat{\sigma}_{x_i x_{i'}}) - \frac{1}{2}\Re(\Omega_{ii'})(\widehat{\sigma}_{x_i}^2 + \widehat{\sigma}_{x_{i'}}^2)}{1 - \Re(\Omega_{ii'})}.$$

Both postfilters [142], [250] use overestimated noise PSD, since they use the input signals rather than the beamformer output for the estimation. An improved postfilter is proposed by Leukimmiatis et al. [253]:

$$w_{\text{Leuk}} = \frac{\widehat{\sigma}_{s_1}^2}{\widehat{\sigma}_{s_1}^2 + \widehat{\sigma}_u^2 \mathbf{w}_{\text{MVDR}}^H \boldsymbol{\Omega} \mathbf{w}_{\text{MVDR}}} \tag{134}$$

with

$$\widehat{\sigma}_{u_i u_{i'}} = \frac{\frac{1}{2}(\widehat{\sigma}_{x_i}^2 + \widehat{\sigma}_{x_{i'}}^2) - \Re(\widehat{\sigma}_{x_i x_{i'}})}{1 - \Re(\Omega_{ii'})}$$

$$\widehat{\sigma}_u^2 = \frac{2}{I(I-1)}\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I}\widehat{\sigma}_{u_i u_{i'}}.$$

A generalized formulation of these postfilters and an EM-based ML estimation procedure a proposed in [235].

A mathematical justification for applying nonlinear postfiltering (provided that it can be stated as an MMSE estimator of a *nonlinear* function of the desired signal) is given in [254] (see also related discussion in [118]). Assuming that the desired source and the noise signals are jointly complex-Gaussian, the conditional probability of $\mathbf{x}$ given s_1 may be expressed as:

$$p(\mathbf{x}|s_1; \sigma_{s_1}^2, \boldsymbol{\Sigma}_{\mathbf{u}}, \mathbf{a}_1) = \frac{1}{\det(\pi\boldsymbol{\Sigma}_{\mathbf{u}})} \exp\left\{-(\mathbf{x} - \mathbf{a}_1 s_1)^H \boldsymbol{\Sigma}_{\mathbf{u}}^{-1}(\mathbf{x} - \mathbf{a}_1 s_1)\right\} \tag{135}$$

it can be shown that a sufficient statistics (in the Bayesian sense) for estimating s_1 in MMSE sense is the output of the MVDR beamformer:

$$p(\rho(s_1)|\mathbf{x}; \sigma_{s_1}^2, \boldsymbol{\Sigma}_{\mathbf{u}}, \mathbf{a}_1) = p(\rho(s_1)|\mathbf{w}_{\text{MVDR}}^H\mathbf{x}; \sigma_{s_1}^2, \boldsymbol{\Sigma}_{\mathbf{u}}, \mathbf{a}_1) \tag{136}$$

where $\rho(\cdot)$ is some nonlinear function. This relation states that the MMSE estimator of $\rho(s_1)$ given the microphone signals can be evaluated by applying the MVDR beamformer to the microphone signals and subsequently applying a single-channel postfilter to its output. By setting $\rho(\cdot)$ to the unity function we simply get the result in (131) that the MWF can be decomposed as an MVDR followed by a single channel Wiener filter. By setting $\rho(\cdot)$ to the absolute value function we obtain the Ephraim and Malah short-time spectral amplitude estimator [210], and by setting $\rho(\cdot)$ to the logarithm of the absolute value the Ephraim and Malah log-spectral amplitude estimator [255] is obtained. This property of the estimator constitutes the justification of applying any proper postfilter to the output of the MVDR beamformer.

[5]Zelinski assumed a simple free-field propagation and hence the RTFs degenerate to delay-only filters.

[6]In this case the MVDR actually degenerates to the DS beamformer.

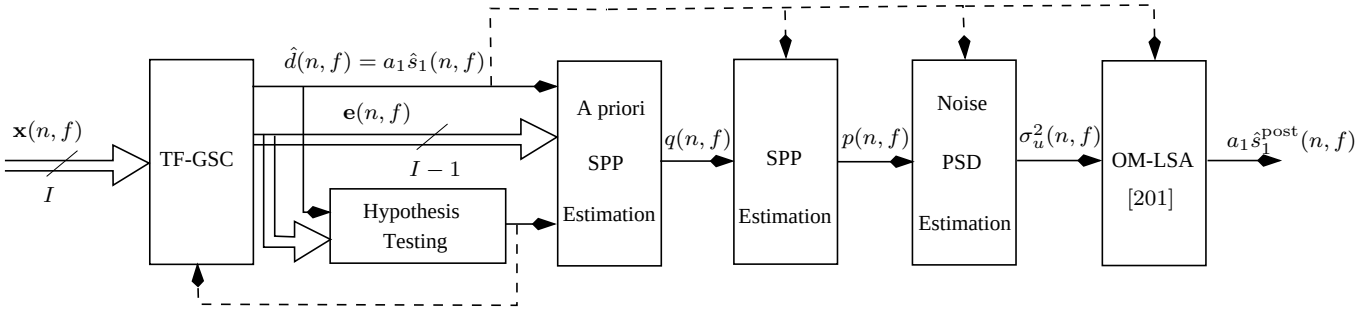


Figure 10. Postfilter incorporating spatial information [256], [257].

A multichannel speech enhancement method comprising an MVDR (implemented in a GSC structure [69]) followed by a modified version of the log-spectral amplitude estimator [211] (OMLSA) is presented in [256], [257]. As depicted in Fig. 10, the required SPP is estimated by incorporating spatial information through the GSC structure. In brief, deviation from stationarity is calculated for both the FBF and BM outputs. A larger change in the lower branch power indicates a change in noise statistics, while a larger measure of non-stationarity at the upper branch indicates speech occurrences. This information can be utilized to enhance the performance of the single-channel speech enhancement algorithm that now incorporates the more reliable spatial information. Moreover, the SPP decisions can be fed back to the beamformer to better control its adaptation. Hence, if non-speech segments are detected, the NC can be updated and if speech is detected, the RTF estimator can be updated, allowing for source tracking.

A statistical analysis of two-channel postfilter estimators in isotropic noise field can be found in [258]. Other nonlinear postfilters can be found in [78], [259], [260]. The underdetermined case, with more sources than microphones, is addressed in [261].

B. Separation by Single-Microphone Masking using Spatial Information

In this section, we explore separation methods that do not apply spatial filtering, but rather apply single microphone separation techniques, usually based on binary masking, that utilize multichannel information. The use of binary masking imitates the perceptual effect of masking in the human auditory system [158]. Furthermore, as speech signals tend to be *W-disjoint* orthogonal in the STFT domain [85], [262], it can be assumed that each time-frequency bin is solely dominated by a single speaker. It is therefore possible to separate the sources by clustering the time-frequency bins and applying a binary mask. The importance of the ideal binary mask as a goal for CASA is summarized in [263]. For further discussion on methods that exploit speech sparsity the reader is referred to Section V-E.

Since first proposed in the early 2000's [85], [158], many algorithms, adopting the masking paradigm, have been proposed. These contributions differ in several aspects: 1) the features used; 2) the clustering procedures; and 3) the type of masking applied. In this section, we will briefly describe the various components of separation by masking using spatial information.

1) Feature Vectors: Masking-based speaker separation algorithms are usually implemented using dual microphone structures, imitating the binaural hearing. The first stage of any separation algorithm, based on clustering, is the feature extraction. ITD, ILD, or a combination thereof, are the most widely used features [158], [264]–[266]. Other popular features are the absolute value and phase the ratios of the two microphone signals in the STFT domain [85], [262], [267], [268], TDOA [269], and single-channel cues (pitch) [270].

2) Classification and Clustering Procedures: A key point in the application of the masking is to assign each time-frequency bin of the mixture to the corresponding speaker. Supervised classification is adopted in [158] using hypothesis testing. In [271], time-frequency bins of the interference source are identified using a BM in a GSC structure trained during speech absence periods. In [266], a mapping between source locations and binaural cues is trained and an inverse mapping is inferred by applying the variational expectation-maximization (VEM) approach.

In [85], [262] an unsupervised clustering approach is adopted, where speakers are clustered according to their different propagation filters (attenuation and delay). In [81] these parameters are estimated by minimizing a cost function resulting in time-frequency bin clustering. Finally, popular clustering techniques, e.g. k-means [80] and Gaussian mixture model (GMM)-EM [264], [268] are widely used in the context of time-frequency bin clustering. For further details on EM-based speech separation methods the reader is referred to Section VI-C.

3) Masking: The last stage in the separation is the application of the mask to separate the various sources. The number of sources can be larger than the number of microphones (undetermined case), but the W-disjoint orthogonality assumption is violated if the number of speakers increases (the percentage of time-frequency bins satisfying this assumption is analyzed in [85], where it is shown to drop below 80% for more than 5 speakers). In most algorithms a binary mask is applied. In [265], a soft mask is proposed based on a mapping between the ITD values and the relative contribution of the target source.

C. Joint spatial-spectral estimation

Following early studies in [105], [272], many recent speech enhancement and source separation approaches now rely on joint modeling and exploitation of spatial and spectral information [112], [115], [183], [228], [229], [233], [242], [273]–[275]. Spatial information is usually represented by spatial models such as those detailed throughout this review. Spectral information relating to the source time-frequency characteristics is modeled for example by constraining the source spectrograms to be low-rank, sparse, and/or to have an excitation-filter structure. Joint modeling of these two types of information potentially improves parameter estimation and subsequent enhancement performance, since spectral information can improve the estimation of the spatial parameters and vice-versa. It also helps solving certain limitations of purely spatial models, such as permutation ambiguity.

The spectral models employed for joint spatial-spectral multichannel estimation are often similar to the spectral models used for single-channel source separation. These models include, but are not restricted to, autoregressive (AR) models [276]–[278], pitch models [273], [274], GMMs [105], [233], hidden Markov models (HMMs) [279], and nonnegative matrix factorization (NMF) [112], [115], [229], [242], [275]. Due to the flexibility of statistical approaches, spectral models allowing statistical formulations are usually easier to integrate within a joint spatial-spectral estimation framework. The choice of a particular model depends on the type of sources to be separated [280]. As such, a general statistical source separation framework was proposed in [228], which allows to combine in a principled way appropriate spatial-spectral models for different sources.

NMF is one of the most popular spectral models for audio source separation [26], [281]. As illustrated in Fig. 11, it approximates a nonnegative source power spectrogram as a product of two nonnegative matrices. One can see from the figure that this decomposition allows for a good approximation of a speech signal while using only a few parameters. We provide below an example of a joint spatial-spectral estimation corresponding to the ML criterion (100) with each source j described by a full-rank unconstrained spatial model $\mathbf{R}_j(f)$ and source variances $\sigma^2_{s_j}(n,f)$ modeled by NMF as

$$\sigma^2_{s_j}(n,f) = \sum_{k=1}^{K_j} b_{jk}(f) h_{jk}(n) \tag{137}$$

with $b_{jk}(f)$ the basis spectra and $h_{jk}(n)$ the time activations [112], [275], [281]. The set of parameters to be estimated is $\boldsymbol{\theta} = \{\{\mathbf{R}_j(f)\}_f, \{b_{jk}(f)\}_{f,k}, \{h_{jk}(n)\}_{k,n}\}_j$. This can be achieved by the following GEM algorithm [242] (one iteration is given below):

- **E-step**: Compute expected sufficient statistics $\widehat{\mathbf{c}}_j(n,f)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_j}(n,f)$ as in (110) and (111).
- **M-step**: Update $\mathbf{R}_j(f)$ as in (112), and update $b_{jk}(f)$ and $h_{jk}(n)$ via multiplicative update (MU) rules [281] as:

$$b_{jk}(f) = b_{jk}(f) \frac{\sum_n \sigma^{-4}_{s_j}(n,f) \widehat{\sigma}^2_{s_j}(n,f) h_{jk}(n)}{\sum_n \sigma^{-2}_{s_j}(n,f) h_{jk}(n)} \tag{138}$$

$$h_{jk}(n) = h_{jk}(n) \frac{\sum_f \sigma^{-4}_{s_j}(n,f)) \widehat{\sigma}^2_{s_j}(n,f) b_{jk}(f)}{\sum_f \sigma^{-2}_{s_j}(n,f) b_{jk}(f)} \tag{139}$$

where $\widehat{\sigma}^2_{s_j}(n,f)$ is computed as in (113).

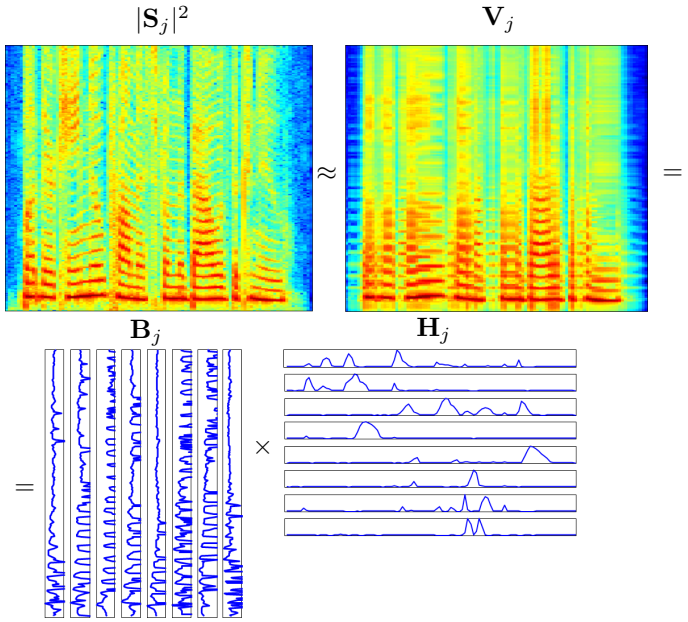


Figure 11. NMF structuring of source variances $\sigma^2_{s_j}(n,f)$ as in (137) can be represented as a factorization of $F \times N$ nonnegative matrix $\mathbf{V}_j \triangleq [\sigma^2_{s_j}(n,f)]_{n,f}$ into a product of $F \times K_j$ and $K_j \times N$ (here $K_j = 8$) nonnegative matrices $\mathbf{B}_j \triangleq [b_{jk}(f)]_{f,k}$ and $\mathbf{H}_j \triangleq [h_{jk}(n)]_{k,n}$, respectively. This decomposition is applied to every source (here, a speech signal), not to the mixture.

Note again that the E-step involves spatial filtering and postfiltering (or, more precisely, joint spatial-spectral filtering (110)) which allows feedback to the parameter estimation. Indeed, the spectral parameters estimated in the previous iteration affect the spatial parameters estimated in the current iteration via $\widehat{\mathbf{c}}_j(n,f)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_j}(n,f)$. Note also that the multiplicative updates (138)–(139) differ from those of single-channel NMF by the fact that they are applied to the estimated power spectrum of each source $\widehat{\sigma}^2_{s_j}(n,f)$ instead of to the observed power spectrum of the mixture. Recently, this algorithm was extended to VB estimation in [231], [237] and in [74], [282] (also considering dynamic scenarios).

VIII. Resources and Results

Over the years, the above speech enhancement and source separation techniques have led to a number of software tools, which are referenced on repositories such as LVA Central[7] and the wiki of ISCA's Special Interest Group on Robust Speech

[7] http://lvacentral.inria.fr/

Table II
SOME AIR DATASETS.

Name	# AIRs	# microphones	# environments	# mic. positions	# speaker positions	interpolation possible	real noise
RWCP [283][9]	364	84	7	1	9	no	no
SiSEC [284][10]	∼50	2	5	1	∼20	no	no
AIR [41][11]	214	2	8	1	13	no	no
Binaural RIR [285][12]	2920	8	5	1	365	yes	yes
CAMIL [266][13]	32400	2	1	16200	1	yes	no
CHiME2 [286][14]	242	2	1	1	121	yes	yes
RIRDB [287][15]	1872	24	3	3	26	no	no

Table III
SOME REAL MULTICHANNEL AUDIO DATASETS, FROM [293].

Name	scenario	duration (h)	# microphones	# environments	speaker dynamics	speaker overlap
Aurora-3[21]	car	∼20	4	1	static	no
AMI [294][22]	meeting	100	16	3	static	yes
DICIT [295][23]	TV order	6	16	1	moving	no
COSINE [296][24]	discuss.	38	20	8	moving	yes
SWC [5][25]	game	7	92	1	moving	yes
CHiME3 [297][26]	tablet	19	6	4	moving	babble

Table IV
EVALUATION SOFTWARE.

Name	Implemented metrics
PESQ [299]	perceptual speech quality (PESQ)
PEMO-Q [300]	perceptual similarity metric (PSM)
STOI [301][27]	short-time objective intelligibility (STOI)
Loizou's [11][28]	segmental signal-to-noise ratio log-likelihood ratio cepstrum distance composite measure
BSS Eval [38][29]	signal-to-distortion ratio (SDR) signal-to-interference ratio (SIR) signal-to-artifacts ratio (SAR)
PEASS [302][30]	overall perceptual score (OPS) target-related perceptual score (TPS) interference-rel. perceptual score (IPS) artifacts-related perceptual score (APS)

Processing[8]. In the following, we provide a non-exhaustive list of popular resources, databases and results, which will be useful for readers to get an idea of the typical performance that may be achieved and to start their own work in the field.

A. Datasets

A first approach to evaluation is to generate the test signals by convolving clean speech signals with AIRs as in (3), summing them together as in (2), and possibly adding real recorded noise. This makes it possible to control the source positions and the room characteristics, which is useful in a development stage. Table II lists a few AIR datasets. Each has its own advantages, depending whether one is interested in a variety of environments, in a large number of microphones, in various speaker-microphone geometries, or in real noise. A variant of this approach which enables even more precise control of the setup is to use artificial AIRs [288], [289] simulated using, e.g., Roomsim [290][16] or Room generator[17] and the respective spherical harmonic domain variant SMIRgen [291][18]. Alternatively, one might record each source separately and then sum all source images together [284], [292]. The series of Signal Separation Evaluation Campaigns (SiSEC) has shown that these three variants lead to comparable separation quality [284]. Although they are often applied to non-moving sources only, the generation of source movements by AIR interpolation has recenty been justified in [286] and implemented in Roomsimove[19] and Signal generator[20].

A second approach is to consider real recorded (and mixed) signals. This is useful in a test stage, but it makes it harder to evaluate the results due to the fact that the true source signals are unknown. Table III lists some datasets for which the target speakers have been recorded by a close-talk microphone so as to provide approximate ground truth. Once again, each dataset has its own advantages, depending whether one is interested in a particular use case, in the amount of data, in the number of microphones, or in the presence of speech overlap. For more details and more datasets, see [293]. The use of mobile robots has recently been proposed as a promising approach towards recording larger real datasets [298].

For either approach, the enhancement and separation qual-

[8] https://wiki.inria.fr/rosp/
[9] http://research.nii.ac.jp/src/en/RWCP-SSD.html
[10] https://sisec.inria.fr/
[11] http://www.ind.rwth-aachen.de/de/forschung/tools-downloads/aachen-impulse-response-database/
[12] http://medi.uni-oldenburg.de/hrir/
[13] https://team.inria.fr/perception/category/data/
[14] http://spandh.dcs.shef.ac.uk/chime_challenge/chime2013/
[15] http://www.eng.biu.ac.il/~gannot/RIR_DATABASE
[16] http://sourceforge.net/projects/roomsim/
[17] http://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator
[18] http://www.audiolabs-erlangen.de/fau/professor/habets/software/smir-generator
[19] http://www.loria.fr/~evincent/Roomsimove.zip
[20] http://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator
[21] http://catalog.elra.info/index.php?cPath=37_40
[22] http://groups.inf.ed.ac.uk/ami/
[23] http://shine.fbk.eu/resources/dicit-acoustic-woz-data
[24] http://melodi.ee.washington.edu/cosine/
[25] http://mini.dcs.shef.ac.uk/data-2/
[26] http://spandh.dcs.shef.ac.uk/chime_challenge/
[27] http://amtoolbox.sourceforge.net/doc/speech/taal2011.php
[28] http://www.crcpress.com/product/isbn/9781466504219
[29] http://bass-db.gforge.inria.fr/bss_eval/
[30] http://bass-db.gforge.inria.fr/peass/

ity can be evaluated by conducting subjective listening tests [302]–[304] or by comparing the estimated source signals with the true source signals using objective performance metrics. Table IV lists a few metrics. According to the study in [302], the frequency-weighted segmental SNR, the composite metric in [11] and the OPS metric of PEASS [302] exhibit the highest correlation with subjective assessment of overall quality.

B. Results

Some attempts were carried out to compare different source separation methods e.g. [284], [305], beamforming methods e.g. [306], and source separation methods vs. beamforming methods [307]. Most of the efforts were carried out in the source separation research community, where six international Signal Separation Evaluation Campaigns (SiSEC) have been run since 2007 [284], [308]–[313]. This allows an objective comparison of different source separation approaches on the same data. Moreover, several source separation methods were used as pre-processing for speech recognition within a series of speech separation and recognition challenges [305]. It was also proposed to objectively evaluate the performance measures of the corresponding source separation methods through the SiSEC campaigns [284].

As for beamforming methods, to the best of our knowledge, there is neither an evaluation campaign nor an evaluation paper comparing the performance of many different beamformers with objective performance metrics. Some studies [306] evaluate different beamformers as pre-processors for a speech recognizer via speech recognition performance metrics. However, since the task used for this evaluation is different from the primary goal of beamforming, it is difficult to conclude from these studies about the performance of the beamformers. In the coming years it is expected that an evaluation campaign will become available in the beamforming community as well. Finally, the study in [307] compared some source separation methods with some beamforming methods. These methods are evaluated for separation of one or two target speech sources from a diffuse background noise.

Hereafter we summarize some results from SiSEC campaigns that are relevant for this review. An overview and comparative analysis of SASSEC [308], SiSEC 2008 [309] and SiSEC 2010 [310] can be found in [284]. Table V summarizes the results of five SiSEC campaigns for the "Underdetermined speech and music mixtures" task over the same dataries that were reused from one campaign to another. The SASSEC campaign [308] is not considered here, since it used a different dataset. We also excluded instantaneous mixtures (only convolutive mixtures are considered) since they are not realistic, as well as music sources (only speech sources are considered) since they are out of the scope of this review. Finally, we excluded the results of partial source separation submissions, i.e. when the corresponding dataset was not processed entirely, since comparing such average results with others is meaningless. As such, the figures in Table V do not coincide in general with the figures from the corresponding SiSEC papers. Details about the considered dataries may be found in the SiSEC papers and the corresponding web pages. We here, very briefly, recall their main characteristics (we use the dataset names from the original campaigns):

- **test**: synthetic or live recorded[31] mixtures; 2 microphones with 1 m or 5 cm spacing; 3 or 4 sources; 0.13 s or 0.25 s RT.
- **test2**: synthetic mixtures; 2 microphones with 20 cm or 4 cm spacing; ; 3 or 4 sources; 0.13 s or 0.38 s RT.
- **test3**: synthetic mixtures; 3 microphones with 50 cm or 5 cm spacing; 4 sources; 0.13 s or 0.38 s RT.

One can draw the following conclusions from the results in Table V. First, separating the live recorded mixtures does not seem to be more difficult than separating the synthetic ones. Second, the performance of the algorithms does not always increase from one campaign to another. This may be explained by the fact that the participants do not usually resubmit the method they have already tried in previous campaigns. Third, none of the methods can overcome the performance ceiling of 5.5 dB SDR and 40 OPS. Finally, one can notice that the two evaluation metrics may behave very differently (see, e.g. the results of two submissions for SiSEC 2011 in Table V).

IX. Summary and perspectives

This section concludes this survey article. First, we discuss the two major algorithmic families introduced in this survey, namely microphone array processing and BSS and their differences and similarities. Then, we provide some guidelines on the selection of the proper algorithm, based on the acoustic scenario and available resources. We conclude this section and the entire article by reviewing some current and future trends in the field.

A. Microphone array processing and BSS: Differences and similarities

Two main paradigms for speech enhancement and source separation were explored in this survey, namely microphone array processing and BSS. We claim here that recent trends are showing that these two paradigms are converging by borrowing ideas from each other.

Concerning the signal models, array processing methods traditionally utilized the spatial resolution of the array as a function of the DOA while BSS methods were originally designed for instantaneous mixtures (no delay, echoes or reverberation). It was then proposed, in the field of array processing, to substitute the simple DOA-based propagation model by ATFs and RTFs [66], [69]. In parallel, BSS methods developed from instantaneous mixtures to convolutive mixtures also modeled by ATFs and RTFs [67], [321]. Under this

[31] For the synthetic mixtures the source images are artificially produced by convolving the sources with the corresponding AIRs. For the live recorded mixtures the source images are physically recorded. In both cases the mixtures are produced synthetically by adding the corresponding source images.

[32] http://sisec2008.wiki.irisa.fr/

[33] http://sisec2010.wiki.irisa.fr/

[34] http://sisec2011.wiki.irisa.fr/

[35] http://sisec2013.wiki.irisa.fr/

[36] http://sisec.inria.fr/

[37] The system is an extended version of the reference.

Table V
SISEC 2008 - 2015 RESULTS FOR CONVOLUTIVE SPEECH MIXTURES OF "UNDER-DETERMINED SPEECH AND MUSIC MIXTURES" TASK. GRAY CELLS MEAN THAT EITHER THE DATASET OR THE EVALUATION METRIC WAS NOT CONSIDERED DURING THE CORRESPONDING EVALUATION CAMPAIGN, "-" SIGN MEANS THAT THE CORRESPONDING DATASET WAS NOT PROCESSED ENTIRELY, AND THE HIGHEST SCORES ARE IN BOLD.

SiSEC	Authors of submission	Method	Average of SDR metric [38]				Average of OPS metric [302]			
			test		test2	test3	test		test2	test3
			Syn.	Live	Syn.	Syn.	Syn.	Live	Syn.	Syn.
2008 [309][32]	Mandel	[314]	1.14	1.91						
	Weiss	[315]	1.46	2.26						
	El Chami	[316]	3.02	2.95						
2010 [310][33]	Ozerov, Nesta and Vincent	[228][37]	1.77	2.66	1.72		**21.2**	37.9	26.8	
	Sawada	[317]	**4.89**	5.18	3.77		20.7	27.9	17.8	
2011 [311][34]	Cho	[318]		3.09	2.54	1.19		26.6	26.3	29.7
	Nesta (1)	[319][37]		4.60	4.00	-		35.8	33.6	-
	Nesta (2)	[319]		**5.30**	4.38	-		**38.5**	**36.5**	-
	Ozerov and Vincent	[228][37]		3.30	2.13	-		32.7	29.5	-
	Sawada	[181]		-	-	**5.28**		-	-	31.1
2013 [312][35]	Cho	[320]		5.29	**4.69**	5.24		31.1	30.5	34.75
	Adiloglu, Kayser and Wang	[228], [231][37]		2.70	0.75	3.23		30.0	23.3	35.22
2015 [313][36]	Nguyen	[181][37]		5.06	4.01	4.27		36.2	32.6	**35.55**

perspective, there is no distinction between the two paradigms anymore. As a matter of fact, their equivalence was already observed in an early stage [322].

Concerning the spatial filter design criteria, array processing techniques traditionally rely on second-order statistics, while BSS techniques can either apply second- or higher-order statistics. However, this distinction is not always applicable. In [323] higher-order statistics were utilized to estimate the steering vectors of a beamformer. A structure, reminiscent of the GSC implementation, that utilizes higher-order statistics was proposed in [324], where the signals are first separated by applying a BSS method, then sorted according to their kurtosis, and finally the desired speech is further enhanced by means of an adaptive noise canceller. More recently, information theoretic criteria based on the TRINICON framework [184] were incorporated into an LCMV beamformer [325]. Conversely, the most widely-used BSS methods today rely on second-order statistics, as popularized by [110]–[112], [227].

Geometry-based information (e.g. microphone distances and DOA) is usually exploited by array processing. It can however also be incorporated into BSS criteria [70], [90], [326]. The use of priors [116], [182] breaks the "blindness" of BSS methods even further. Conversely, some array processing methods, e.g. [69], [151], are not using any spatial priors but rather rely on a specific activity pattern.

Concerning the application of the methods, the array processing paradigm traditionally addressed speech detection, parameter estimation and signal separation successively, while the BSS paradigm addressed them in parallel. This line is becoming blurred too, as certain BSS techniques relies on successive estimation [72], [85] while joint parameter estimation has been used [276], [277] and is becoming more popular in the array processing community [246], [247].

The scenario of several concurrently speaking speakers is well-studied in the BSS area, but more cumbersome in array processing. It was shown in [137] (and extended later in [95]), that the RTFs of the entire group of desired and the entire group of interference sources can be estimated, provided that the group activities do not overlap during the estimation period. These RTFs estimators are then utilized to construct a subspace-based LCMV beamformer. Recent contributions [90], [327] circumvent the requirement for disjoint activity of the desired and interference groups of signals, by incorporating concepts from the BSS paradigm into the beamformer design. Other array processing methods for multiple concurrent sources involve hypothesis testing for the activity of the sources in T-F bins, before applying the optimal beamformer [216], [233], [328].

Finally, classical beamformers do not handle under-determined problems. However, with the utilization of speech sparsity this becomes feasible for both paradigms [80], [150], [329].

We can therefore conclude that the array processing and BSS paradigms share many underling concepts, and will eventually converge to a point in which they will become indistinguishable.

B. Guidelines

We have surveyed a plethora of algorithms and methods for speech enhancement and separation. In this section we will not attempt to pick up the "best" algorithm, but rather give guidelines for selecting the most appropriate class of algorithms for a given scenario.

1) Number of sources and microphones: If there is only one source in noise, the natural choice would be the MWF/MVDR family of beamformers. If multiple sources of interest exist (either desired or interference) but their number remains smaller than the number of microphones, then LCMV beamformers or BSS methods can be considered. If the number of microphones is smaller than the number of sources of interest (under-determined problem), then the speech sparsity should be utilized, usually in conjunction with BSS methods, but also with modern beamformers.

2) Array geometry: If the specific array geometry is known, e.g., linear, differential or spherical, preference should be given to array processing or source separation methods which exploit this information. In entirely blind scenarios, BSS methods are commonly used, however some modern array processing methods based on ATFs or RTFs can also be used.

3) Prior information: If additional prior information is available, preference should be given to methods which exploit this information. For instance, information about the source DOA can be exploited both by array processing and source separation methods, while information about the nature of the sources and training data is more easily exploited by the latter.

C. Perspectives

In this section we explore some of the current and future trends in the field of multi-microphone array processing.

1) Learning-based spatial filters: The signal models we have reviewed in this article rely on limited prior information: the ATFs (or the respective RTFs) are assumed to be either unconstrained, or to satisfy universal constraints depending on the source positions [70], [81] and the room reverberation time [61], [116]. What if the exact shape and acoustic properties of the room were fixed? In this situation, one would know the exact ATFs/RTFs associated with all possible source and microphone positions. Source localization and ATF estimation would become identical problems, that would be much easier to solve. This ideal situation can be approximated in practice by acquiring ATFs for a finite set of source and microphone positions using mobile devices and interpolating them to find the ATFs in other positions, using the fact that the set of ATFs in a given room forms a (nonlinear) manifold. This idea has recently been the starting point for an increasing number of studies which model the manifold of ATFs using models based on sparsity and compressed sensing [56], [57], [60], [330]–[334] or locally linear embedding and manifold learning [62], [335]–[338]. A preliminary study of applying the latter concepts to construct a GSC beamformer can be found in [339]. Extending these approaches to work across a full, real-world room is an exciting perspective.

2) Deep learning-based parameter estimation: deep neural networks (DNNs) have emerged as a promising alternative to SPP estimation, EM, VB, or MM in the situations when large amounts (typically, hours) of source signals similar to those in the mixture to be separated, are available for training. DNNs model complex, high-dimensional functions by making efficient use of this wealth of data. They typically operate in the magnitude STFT domain, take several frames of the mixture as inputs, and output the SPP or the spectra of all sources in each time frame. Most work in this area has focused and still focuses on single-channel separation using spectral cues or channel-wise filtering using ILD and ITD cues [340] or pitch and azimuth [341] (using multi-layer perceptrons).

Yet, a few multichannel approaches have recently been proposed that use DNNs in a variety of ways: to estimate the SPP using a DNN and subsequently derive a beamformer [342], to alternately reestimate the source magnitude spectra using DNNs and the spatial covariance matrices in an EM-like fashion [343], or to learn beamformers directly in the time domain [344]. The integration of these data-driven techniques with the domain knowledge learned from the established techniques presented in this review is an open research direction.

3) Distributed algorithms for ad hoc microphone arrays: In classical microphone array processing, as explored in this survey, both the sensing and the processing of the acquired speech are concentrated in a single device, usually called a *fusion center*. In many scenarios, this approach cannot provide the required performance, since the acoustic scene may be spatially distributed, and a powerful fusion center may not be available. It is therefore reasonable to alleviate the performance drop by a large spatial deployment of inter-connected microphone sub-arrays (nodes), arranged in a wireless network, preferably equipped with local processors. Recent technological advances in the design of miniature and low-power electronic devices make such distributed microphone networks, often referred to as wireless acoustic sensor network (WASN), feasible. As a matter of fact, cellular phone, laptops and tablets are perfect candidates as nodes of such networks, as they are self-powered and equipped with multiple microphones (typically two to three), as well as powerful processors and various wireless communication modules. The large spatial distribution of WASNs increases the probability that a subset of the microphones is close to a relevant sound source and has the potential to yield improved performance as compared with classical, condensed, microphone arrays. However, the distributed and ad hoc nature of WASNs arises new challenges, e.g. transmission and processing constraints and intricate network topology, that should be addressed to fully exploit their potential.

Several families of algorithms, that allow for optimal solutions without requiring the transmission of all signals to a central processor, but rather a compressed/fused version thereof, have been proposed [345]. One such family is the distributed adaptive node-specific signal estimation (DANSE) family of algorithms, which allow for a distributed implementation of several speech enhancement algorithms that were introduced in this survey (including SDW-MWF [346], [347] and LCMV beamforming [220]). Several network topologies can be implemented, e.g. fully-connected and tree-structure. Another family of algorithms exploits the special GSC structure to obtain recursive solutions that are proven to converge to the optimal beamforming criteria [348]. Theoretical performance bounds of such distributed microphone array algorithms can be derived [349]. Efficient adaptation mechanisms to changes in the number of available nodes and signals of interest can be found in [350].

Randomized gossip implementation of the DS beamformer is presented in [351]. A distributed algorithm for MVDR beamforming, based on message passing, is presented in [352]. A distributed MVDR beamformer based on the diffusion adaptation paradigm, that neither imposes conditions on the topology of the network nor requires knowledge of the noise covariance matrix, can be found in [353]. Near-field beamformers using smartphones forming an ad hoc network are described in [354]. Intra- and inter-node location features are

integrated in a clustering-based scheme for speech separation in [355].

In practical scenarios the microphone signals can have an arbitrary temporal offsets. A method to alleviate this problem can be found in [356]. In severe cases, identical sampling frequency across all nodes cannot be guaranteed. Method to re-synchronize the signals, either based on the communication link, on the speech signals, or on a combination thereof, can be found in [357]–[365].

Despite these advances in the field of distributed algorithms for ad hoc microphone arrays, the quest for a full-fledged solution, considering all challenges in reverberant and dynamic acoustic environment, is still far from reached.

4) Robustness: Two kinds of robustness can be attributed to beamformers, namely *numerical robustness* and *spatial robustness* [366]. Numerical robustness refers to the sensitivity of the array gain to mismatches in the microphone gains and phases and the beamformer weights. As was already explored in Section IV-A in the general context of (narrowband) array processing, numerical robustness is proportional to the WNG [118], [121] (see modification for structured arrays in [122]). It is further proposed in [121] to increase the robustness, trading-off array directivity, using diagonal loading. In [367] it is proposed to increase the robustness of a broadband array with an arbitrary directivity, by incorporating the probability density functions of the microphone gains and phases into the beamformer design criterion. The statistics of the microphone characteristics is also taken into account in the design of a robust superdirective beamformer (e.g. differential microphone arrays) [368].

The term spatial robustness refers to mismatches between the actual location of a desired source and the assumed location used to derive the beamformer weights. It is a widely-explored area in the field of array processing [369]–[371]. Widening the beampattern is a common practice to increase robustness to steering errors. This can be done by either adding derivative constraints [189], by methods borrowed from filtering design procedures [372], or by defining multiple constraints in an area surrounding the prospective source location [191]. Alternatively, a probabilistic framework for describing the errors in the steering vectors is proposed in [366] to design a robust beamformer based on the maximum signal to interference plus noise ratio (SINR) criterion.

The GSC structure is utilized to combine spatial robustness considerations (in the BM block) and numerical robustness considerations (in the adaptive noise canceller (ANC) block) [199], [200]. A GSC-type beamformer utilizing advanced BSS techniques, namely TRINICON, is also proposed for increasing robustness [373].

Although many methods for increasing the robustness of beamformers in speech applications can be found in the literature, designing a robust beamformer that takes room acoustics and speech properties into account, remains an open research question.

5) Dynamic scenarios and tracking: The application of beamforming and BSS techniques, explored in this survey, to dynamic scenarios is a cumbersome task, mainly due to the limited amount of data available for estimating the time-varying filters. Methods that utilizes instantaneous direction-of-arrival estimates to allow for fast adaptation of LCMV beamformers can be found in [163], [374]. A tracking mechanism for the RTFs of the desired and interference sources, based on subspace tracking [375], is described in [198]. The time-varying estimates of the RTF are utilized to design LCMV beamformer to extract the set of desired speakers. EM and VB frameworks were also proposed in [74], [114], [282]. The resulting algorithms employ a Kalman smoother to estimate time-varying mixing filters, that are subsequently utilized to construct Wiener filters for separating the sources in under-determined mixtures.

6) Binaural multi-microphone processing: The objective of a binaural noise reduction algorithm is not only to selectively extract the desired speaker and to suppress interfering sources and ambient background noise, but also to preserve the auditory impression, as perceived by the hearing aid user. Existing methods can be roughly categorized into three main families.

The first family is based on the concept of CASA [376]–[378], which aims at imitating the behavior of the human auditory system [25], [379].

The second family consists of BSS algorithms [380]–[382], which are based on the fundamental assumption of mutual statistical independence of the different source signals.

The third family is based on a binaural versions of the MMSE [92], MVDR and LCMV criteria. The binaural MWF inherently preserves the binaural cues of the desired source but distorts the binaural cues of the noise (i.e. the beamformer imposes the noise to be coherent and perceived as arriving from the same direction as the desired source). Several extensions of the binaural MWF have been introduced aiming to also preserve the binaural cues of the noise [93], [96], [97], [383]. By design, these methods suffer from some distortion imposed on the desired source component at the output. Alternatively, distortionless criteria, such as the MVDR and LCMV, can be used instead of MWF [95], [98], [384]–[386].

Hearing aids impose severe design constraints on the developed algorithm: short latency, fast adaption, small number of microphones, limited connectivity between the hearing devices and low-complexity, to name a few. Designing algorithms, satisfying these constraints, and still exhibiting high noise and interference reduction together spatial cues preservation, is still an ongoing research topic.

7) Audio-visual speech enhancement: Finally, although we have focused on audio-only algorithms above, one must bear in mind that microphones are quite often embedded in devices equipped with other sensors, e.g., cameras, accelerometers. In [387], cameras have been used to estimate speech statistics from visual (face and lips) features and noise statistics from visual voice activity detection. They have also been used to find the spatial location of the sources in [388]–[391]. An optimal integration of acoustic and visual information is obtained by joint inference in both modalities using the turbo-decoding framework [392]. In [393] an audio-visual voice activity detection is proposed, using dimensionality reduction.

The area of audio-visual speech processing remains largely understudied despite its great promise.

REFERENCES

[1] A. Boothroyd, K. Fitz, J. Kindred, S. Kochkin, H. Levitt, B. Moore, and J. Yanz, "Hearing aids and wireless technology," *Hearing Review*, vol. 14, no. 6, pp. 44, 2007.

[2] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. of Am.*, vol. 78, no. 5, pp. 1508–1518, 1985.

[3] H. Silverman, I. Patterson, W.R., J. Flanagan, and D. Rabinkin, "A digital processing system for source location and sound capture by large microphone arrays," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, Apr. 1997, vol. 1, pp. 251–254.

[4] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu, et al., "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Lang. Resources Eval.*, vol. 41, no. 3–4, pp. 389–407, 2007.

[5] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, "The Sheffield wargames corpus," in *Proc. Interspeech*, 2013, pp. 1116–1120.

[6] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. Int. Conf. on Lang. Res. and Eval.*, 2014.

[7] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Sig. Proc.*, vol. 87, no. 8, pp. 1933–1950, 2007.

[8] S. L. Gay and J. Benesty, Eds., *Acoustic signal processing for telecommunication*, Kluwer, 2000.

[9] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001.

[10] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, 2005.

[11] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[12] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech processing in modern communication: Challenges and perspectives*, Springer, 2010.

[13] P. O'Grady, B. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *Int. J. Imag. Sys. Tech.*, vol. 15, pp. 18–33, 2005.

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

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

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

[17] 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*, pp. 162–185. IGI Global, 2010.

[18] 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 Sig. Proc. Mag.*, vol. 31, no. 3, pp. 107–115, 2014.

[19] U. Zölzer, Ed., *DAFX: Digital Audio Effects*, Wiley, 2011.

[20] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Prague, May 2011, pp. 257–260.

[21] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet, "Linear mixing models for active listening of music productions in realistic studio conditions," in *Proc. Au. Eng. Soc. Conv.*, 2012.

[22] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A practical Approach*, Wiley, 2004.

[23] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[24] P. Divenyi, Ed., *Speech Separation by Humans and Machines*, Springer Verlag, 2004.

[25] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley, 2006.

[26] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing," *IEEE Sig. Proc. Mag.*, vol. 32, no. 2, pp. 125–144, 2015.

[27] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing and Speech Communication*. Springer, 2007.

[28] M. Wölfel and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.

[29] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, 2012.

[30] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 4, pp. 745–777, 2014.

[31] N. Ono, K. Miyamoto, H. Kameoka, J. Le Roux, Y. Uchiyama, E. Tsunoo, T. Nishimoto, and S. Sagayama, "Harmonic and percussive sound separation and its application to MIR-related tasks," in *Advances in Music Information Retrieval*. Springer, 2010.

[32] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Sp., Sig. Proc.*, vol. 24, no. 4, pp. 320–327, 1976.

[33] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP J. on Appl. Sig. Proc.*, vol. 2006, pp. 1–19, 2006.

[34] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *Proc. Joint Workshop on Hands-free Sp. Comm. and Mic. Arrays*. IEEE, 2008, pp. 69–72.

[35] H. Kuttruff, *Room acoustics*, Taylor & Francis, 2000.

[36] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, 1997.

[37] D. Marković, K. Kowalczyk, F. Antonacci, C. Hofmann, A. Sarti, and W. Kellermann, "Estimation of acoustic reflection coefficients through pseudospectrum matching," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 1, pp. 125–137, Jan. 2014.

[38] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[39] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 1998, vol. 4, pp. 1941–1944.

[40] S.-K. Lee, "Measurement of reverberation times using a wavelet filter bank and application to a passenger car," *Journal Au. Eng. Soc.*, vol. 52, no. 5, pp. 506–515, 2004.

[41] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. IEEE Int. Conf. on Dig. Sig. Proc.*, 2009, pp. 1–4.

[42] M. R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *Journal Au. Eng. Soc.*, vol. 35, no. 5, pp. 299–306, 1987.

[43] J.-D. Polack, "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics," *Appl. Acous.*, vol. 38, no. 2, pp. 235–244, 1993.

[44] M. R. Schroeder, "Frequency correlation functions of frequency responses in rooms," *J. Acoust. Soc. of Am.*, vol. 34, no. 12, pp. 1819–1823, 1963.

[45] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Sp. Au. Proc.*, vol. 11, pp. 791–803, 2003.

[46] R. Scharrer and M. Vorländer, "Sound field classification in small microphone arrays using spatial coherences," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 9, pp. 1891–1899, Sept. 2013.

[47] O. Schwartz, E. Habets, and S. Gannot, "Nested generalized sidelobe canceller for joint dereverberation and noise reduction," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Brisbane, Australia, Apr. 2015.

[48] H.-L. Nguyen Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Sig. Proc.*, vol. 45, no. 2, pp. 209–229, 1995.

[49] S. Choi and A. Cichocki, "Adaptive blind separation of speech signals: Cocktail party problem," in *Proc. IEEE Int. Conf. on Sig. Proc.*, 1997, pp. 617–622.

[50] F. Ehlers and H. G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment," *IEEE Trans. Sig. Proc.*, vol. 45, no. 10, pp. 2608–2612, 1997.

[51] R. H. Lambert and A. J. Bell, "Blind separation of multiple speakers in a multipath environment," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 1997, pp. I–423–I–426.

[52] H.-C. Wu and J. C. Príncipe, "Generalized anti-Hebbian learning for source separation," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 1999, pp. II–1073–II–1076.

[53] M. Ito, Y. Takeuchi, T. Matsumoto, H. Kudo, M. Kawamoto, T. Mukai, and N. Ohnishi, "Moving-source separation using directional microphones," in *Proc. Intl. Symp. on Sig. Proc. and Info. Tech. (ISSPIT)*, 2002, pp. 523–526.

[54] A. Aissa-El-Bey, K. Abed-Meraim, and Y. Grenier, "Blind separation of underdetermined convolutive mixtures using their time-frequency representation," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 15, no. 5, pp. 1540–1550, 2007.

[55] M. Gupta and S. C. Douglas, "Beamforming initialization and data prewhitening in natural gradient convolutive blind source separation of speech mixtures," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2007, pp. 462–470.

[56] Y. Lin, J. Chen, Y. Kim, and D. D. Lee, "Blind channel identification for speech dereverberation using ℓ1-norm sparse learning," in *Proc. Neural Info. Proc. Conf.*, 2007, pp. 921–928.

[57] P. Sudhakar, S. Arberet, and R. Gribonval, "Double sparsity: Towards blind estimation of multiple channels," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 571–578.

[58] M. Yu, W. Ma, J. Xin, and S. Osher, "Multi-channel ℓ_1 regularized convex speech enhancement model and fast computation by the split bregman method," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 20, no. 2, pp. 661–675, 2012.

[59] I. J. Kelly and F. M. Boland, "Detecting arrivals in room impulse responses with dynamic time warping," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 7, pp. 1139–1147, July 2014.

[60] Z. Koldovký, J. Malek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1335–1347, Aug. 2015.

[61] A. Benichoux, L. S. R. Simon, E. Vincent, and R. Gribonval, "Convex regularizations for the simultaneous recording of room impulse responses," *IEEE Trans. Sig. Proc.*, vol. 62, no. 8, pp. 1976–1986, Apr. 2014.

[62] B. Laufer, R. Talmon, and S. Gannot, "A study on manifolds of acoustic responses," in *Latent Variable Analysis and Independent Component Analysis (LVA-ICA)*, Liberec, Czech Republic, Aug. 2015.

[63] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 18, no. 7, pp. 1818–1829, 2010.

[64] J. Chen, J. Benesty, and Y. Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 16, no. 3, pp. 481–493, 2008.

[65] R. Crochiere and L. Rabiner, *Multi-Rate Signal Processing*, Englewood Cliffs, NJ: Prentice Hall, 1983.

[66] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Sp. Au. Proc.*, vol. 5, no. 5, pp. 425–437, Sept. 1997.

[67] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[68] S. Ikeda and N. Murata, "An approach to blind source separation of speech signals," in *Proc. Int. Conf. on Artificial Neural Networks (ICANN)*, 1998, pp. 761–766.

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

[70] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Sp. Au. Proc.*, vol. 10, no. 6, pp. 352–362, Sept. 2002.

[71] B. Albouy and Y. Deville, "Alternative structures and power spectrum criteria for blind segmentation and separation of convolutive speech mixtures," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2003, pp. 361–366.

[72] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2004, pp. 652–660.

[73] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Sp. Au. Proc.*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[74] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. P. Horaud, "A variational EM algorithm for the separation of moving sound sources," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, New Paltz, USA, Oct. 2015.

[75] 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. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2003, pp. 311–314.

[76] F. N. ans P. Svaizer and M. Omologo, "Convolutive BSS of short mixtures by ICA recursively regularized across frequencies," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 19, no. 3, pp. 624–639, 2011.

[77] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 15, no. 2, pp. 715–726, 2007.

[78] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Comp. Sp. and Lang.*, vol. 27, no. 3, pp. 726–745, 2013.

[79] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberation modeling," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, New Paltz, NY, United States, Oct. 2015.

[80] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Sig. Proc.*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[81] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components with estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 15, no. 5, pp. 1592–1604, 2007.

[82] S. Stenzel, J. Freudenberger, and G. Schmidt, "A minimum variance beamformer for spatially distributed microphones using a soft reference selection," in *Proc. Joint Workshop on Hands-free Sp. Comm. and Mic. Arrays*, May 2014, pp. 127–131.

[83] A. Deleforge, S. Gannot, and W. Kellermann, "Towards a generalization of relative transfer functions to more than one source," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2015, pp. 419–423.

[84] X. Li, R. Horaud, L. Girin, and S. Gannot, "Local relative transfer function for sound source localization," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2015, pp. 399–403.

[85] O. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[86] M. Puigt and Y. Deville, "Time-frequency ratio-based blind separation methods for attenuated and time-delayed sources," *Mech. Syst. Signal Process*, vol. 19, pp. 1348–1379, 2005.

[87] T. Melia and S. T. Rickard, "Underdetermined blind source separation in echoic environments using DESPRIT," *EURASIP J. on Adv. in Sig. Proc.*, vol. 2007, pp. 1–19, 2007.

[88] C. Liu, B. C. Wheeler, W. D. O'Brien Jr., C. R. Lansing, R. C. Bilger, D. L. Jones, and A. S. Feng, "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," *J. Acoust. Soc. of Am.*, vol. 110, no. 6, pp. 3218–3231, Dec. 2001.

[89] J. Anemüller and B. Kollmeier, "Adaptive separation of acoustic sources for anechoic conditions: A constrained frequency domain approach," *Sp. Comm.*, vol. 39, no. 1–2, pp. 79–95, 2003.

[90] 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 App. of Sig. Proc. to Au. and Acous. (WASPAA)*, New Paltz, USA, Oct. 2013.

[91] A. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *J. Acoust. Soc. of Am.*, vol. 83, no. 4, pp. 1508–1516, 1988.

[92] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. Liu, Eds. Wiley-IEEE Press, 2010.

[93] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 18, no. 2, pp. 342–355, 2010.

[94] D. Marquardt, V. Hohmann, and S. Doclo, "Coherence preservation in multi-channel wiener filtering based noise reduction for binaural hearing aids," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2013, pp. 8648–8652.

[95] E. Hadad, S. Gannot, and S. Doclo, "Binaural linearly constrained minimum variance beamformer for hearing aid applications," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, Aachen, Germany, Sept. 2012.

[96] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel Wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, pp. 2384–2397, Dec. 2015.

[97] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 23, no. 12, pp. 2162–2176, Dec. 2015.

[98] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, pp. 2449–2464, Dec. 2015.

[99] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 24, no. 3, pp. 543–558, Mar. 2016.

[100] J. Shynk, "Frequency-domain and multirate and adaptive filtering," *IEEE Sig. Proc. Mag.*, vol. 9, no. 1, pp. 14–37, Jan. 1992.

[101] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Sp., Sig. Proc.*, vol. 28, no. 1, pp. 55–69, Feb. 1980.

[102] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, 2013.

[103] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. Sig. Proc.*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[104] Y. Avargel and I. Cohen, "Adaptive system identification in the short-time fourier transform domain using cross-multiplicative transfer function approximation," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 16, no. 1, pp. 162–173, Jan. 2008.

[105] H. Attias, "New EM algorithms for source separation and deconvolution with a microphone array," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2003, vol. V, pp. 297–300.

[106] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 17, no. 7, pp. 1420–1434, Sept. 2009.

[107] W. Kellermann and H. Buchner, "Wideband algorithms versus narrowband algorithms for adaptive filtering in the DFT domain," in *Proc. Asilomar Conf. on Sig., Syst. and Comp.*, 2003, vol. 2, pp. 1278–1282.

[108] C. Servière, "Separation of speech signals with segmentation of the impulse responses under reverberant conditions," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2003, pp. 511–516.

[109] S. Mirsamadi, S. Ghaffarzadegan, H. Sheikhzadeh, S. M. Ahadi, and A. H. Rezaie, "Efficient frequency domain implementation of noncausal multichannel blind deconvolution for convolutive mixtures of speech," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 20, no. 8, pp. 2365–2377, Oct. 2012.

[110] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, 2005, pp. 78–81.

[111] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2009, pp. 775 – 782.

[112] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[113] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[114] N. Q. K. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2011, pp. 205–208.

[115] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 3, pp. 727–739, Mar. 2014.

[116] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for Gaussian model based reverberant audio source separation," *EURASIP J. on Adv. in Sig. Proc.*, vol. 2013, pp. 149, Sept. 2013.

[117] R. Martin, *Freisprecheinrichtungen mit mehrkanaliger Echokompensation und Störgeräuschreduktion*, ABDN, Band 3. Verlag der Augustinus Buchhandlung, Aachen, 1995, (in German).

[118] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, vol. IV, Optimum Array Processing, Wiley, New York, Apr. 2002.

[119] D. Levin, E. Habets, and S. Gannot, "A generalized theorem on the average array directivity factor," *IEEE Signal Process. Lett.*, vol. 20, no. 9, pp. 877–880, July 2013.

[120] A. T. Parsons, "Maximum directivity proof for three-dimensional arrays," *J. Acoust. Soc. of Am.*, vol. 82, no. 1, pp. 179–182, 1987.

[121] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Sp., Sig. Proc.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[122] D. Levin, E. Habets, and S. Gannot, "Robust beamforming using sensors with nonidentical directivity patterns," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Vancouver, Canada, May 2013.

[123] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Sp. Comm.*, vol. 20, no. 3, pp. 229–240, 1996.

[124] J. Chen, J. Benesty, and C. Pan, "On the design and implementation of linear differential microphone arrays," *J. Acoust. Soc. of Am.*, vol. 136, no. 6, pp. 3097–3113, 2014.

[125] J. Benesty and J. Chen, *Study and Design of Differential Microphone Arrays*, Springer, 2013.

[126] J. Benesty, J. Chen, and I. Cohen, *Design of Third-Order Circular Differential Arrays*, Springer, 2015.

[127] J. Benesty, J. Chen, and C. Pan, *Fundamentals of Differential Beamforming*, Springer, 2016.

[128] H.-E. de Bree, P. Leussink, T. Korthorst, H. Jansen, T. S. Lammerink, and M. Elwenspoek, "The μ-flown: a novel device for measuring acoustic flows," *Sensors and Actuators A: Physical*, vol. 54, no. 1, pp. 552–557, 1996.

[129] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*. IEEE, 2002, vol. 2, pp. 1781–1784.

[130] G. W. Elko and J. M. Meyer, "Using a higher-order spherical microphone array to assess spatial and temporal distribution of sound in rooms," *J. Acoust. Soc. of Am.*, vol. 132, no. 3, pp. 1912–1912, Mar. 2012.

[131] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*. IEEE, 2002, vol. 2, pp. 1949–1952.

[132] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Sp. Au. Proc.*, vol. 13, no. 1, pp. 135–143, 2005.

[133] Z. Li and R. Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 15, no. 2, pp. 702–714, 2007.

[134] C. T. Jin, N. Epain, and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 1, pp. 193–204, Jan. 2014.

[135] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8, Springer, 2015.

[136] N. Ito, H. Shimizu, N. Ono, and S. Sagayama, "Diffuse noise suppression using crystal-shaped microphone arrays," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 19, no. 7, pp. 2101–2110, Sept. 2011.

[137] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

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

[139] E. Jan and J. Flanagan, "Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*. IEEE, 1996, vol. 2, pp. 917–920.

[140] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acous. Sp. and Sig. Proc. Mag.*, pp. 4–24, Apr. 1988.

[141] C. L. Dolph, "A current distribution for broadside arrays which optimizes the relationship between beam width and side-lobe level," *Proc. IRE*, vol. 34, no. 6, pp. 335–348, June 1946.

[142] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Sp. Au. Proc.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[143] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., chapter 3, pp. 39–60. Springer-Verlag, Berlin, 2001.

[144] W. Kellermann, "A self-steering digital microphone array," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 1991, pp. 3581–3584.

[145] D. B. Ward, R. A. Kennedy, and R. C. Williamson, "Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1023–1034, 1995.

[146] S. Doclo and M. Moonen, "Design of far-field and near-field broadband beamformers using eigenfilters," *Sig. Proc.*, vol. 83, no. 12, pp. 2641–2673, 2003.

[147] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel Wiener filter for multiple sources scenarios," in *Proc. IEEE Conv. of Elect. Electron. Eng. in Israel (IEEEI)*, Eilat, Israel, Nov. 2012, best student paper award.

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

[149] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. of Am.*, vol. 54, no. 3, pp. 771–785, Sept. 1973.

[150] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*. IEEE, 2007, vol. 1, pp. 41–44.

[151] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 15, no. 5, pp. 1529–1539, 2007.

[152] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, 3rd edition, Nov. 1996.

[153] E. Habets, J. Benesty, I. Cohen, and S. Gannot, "On a tradeoff between dereverberation and noise reduction using the MVDR beamformer," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*. IEEE, 2009, pp. 3741–3744.

[154] S. Doclo and M. Moonen, "Multimicrophone noise reduction using recursive GSVD-based optimal filtering with ANC postprocessing stage," *IEEE Trans. Sp. Au. Proc.*, vol. 13, no. 1, pp. 53–69, 2005.

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

[156] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, vol. 1, Springer, 2008.

[157] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.

[158] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. of Am.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[159] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, 2007, pp. 147–150.

[160] R. Gribonval, "Piecewise linear source separation," in *SPIE Wavelets: Applications in Signal and Image Processing*, 2003, pp. 297–310.

[161] J. P. Rosca, C. Borss, and R. V. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2004, pp. III–877–III–880.

[162] M. Togami, T. Sumiyoshi, and A. Amano, "Sound source separation of overcomplete convolutive mixtures using generalized sparseness," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2006.

[163] O. Thiergart, M. Taseska, and E. A. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.

[164] E. Vincent, "Complex nonconvex l_p norm minimization for underdetermined source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2007, pp. 430–437.

[165] M. Maazaoui, Y. Grenier, and K. Abed-Meraim, "Frequency domain blind source separation for robot audition using a parameterized sparsity criterion," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2011, pp. 1869–1873.

[166] S. Choi, A. Cichocki, and S. Amari, "Flexible independent component analysis," in *Neural Networks for Signal Processing (NNSP 8)*, 1998, pp. 83–92.

[167] R. Everson and S. Roberts, "Independent component analysis: a flexible nonlinearity and decorrelating manifold approach," *Neural Computation*, vol. 11, pp. 1957–1983, 1999.

[168] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 165–172.

[169] G. Bao, Z. Ye, X. Xu, and Y. Zhou, "A compressed sensing approach to blind separation of speech mixture based on a two-layer sparsity model," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 5, pp. 899–906, May 2013.

[170] R. C. Hendriks, J. S. Erkelens, J. Jensen, and R. Heusdens, "Minimum mean-square error amplitude estimators for speech enhancement under the generalized gamma distribution," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, 2006, pp. 1–6.

[171] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," *IEICE Technical Report*, vol. E86-A, no. 3, pp. 590–596, Mar. 2003.

[172] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Technical Report*, vol. E87-A, no. 8, pp. 1941–1948, 2004.

[173] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. on Appl. Sig. Proc.*, vol. 2005, pp. 1110–1126, 2005.

[174] J. I. Marin-Hurtado, D. N. Parikh, and D. V. Anderson, "Perceptually inspired noise-reduction method for binaural hearing aids," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 20, no. 4, pp. 1372–1382, May 2012.

[175] J.-F. Cardoso, "The three easy routes to independent component analysis; contrasts and geometry," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2001, pp. 1–6.

[176] P. Comon, "Independent component analysis, a new concept?," *Sig. Proc.*, vol. 36, no. 3, pp. 287–314, 1994.

[177] T. Lee, *Independent Component Analysis - Theory and Applications*, Norwell, 1998.

[178] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Adaptive and Learning Systems. Wiley, 1st edition, 2001.

[179] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. N. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 17, no. 5, pp. 994–1008, July 2009.

[180] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Sp. Au. Proc.*, vol. 12, no. 5, pp. 530–538, 2004.

[181] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 19, no. 3, pp. 516–527, 2011.

[182] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization," *EURASIP J. on Appl. Sig. Proc.*, vol. 2007, no. 1, pp. 81–81, 2007.

[183] S. Arberet, P. Vandergheynst, R. E. Carrillo, J.-P. Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 7, pp. 1391–1402, July 2013.

[184] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON-based blind system identification with application to multiple-source localization and separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds., pp. 101–147. Springer, Heidelberg, 2007.

[185] N. Mitianoudis and M. E. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. Sp. Au. Proc.*, vol. 11, no. 5, pp. 489–497, 2003.

[186] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2006, pp. 601–608.

[187] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 15, no. 1, pp. 70–79, 2007.

[188] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, 2011, pp. 189–192.

[189] M. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Trans. Acoust., Sp., Sig. Proc.*, vol. 31, no. 6, pp. 1378–1393, 1983.

[190] K. Buckley, "Spatial/spectral filtering with linearly constrained minimum variance beamformers," *IEEE Trans. Acoust., Sp., Sig. Proc.*, vol. 35, no. 3, pp. 249–266, Mar. 1987.

[191] Y. Zheng, R. Goubran, and M. El-Tanany, "Robust near-field adaptive beamforming with distance discrimination," *IEEE Trans. Sp. Au. Proc.*, vol. 12, no. 5, pp. 478–488, Sept. 2004.

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

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

[194] B. R. Breed and J. Strauss, "A short proof of the equivalence of LCMV and GSC beamforming," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 168–169, 2002.

[195] G. Strang, *Linear Algebra and its Application*, Academic Press, 2nd edition edition, 1980.

[196] B. Widrow, J. G. Jr., J. McCool, J. Kaunitz, C. Williams, R. Hearn, J. Zeider, E. D. Jr., and R. Goodlin, "Adaptive noise cancelling: Principals and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.

[197] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. Vehicular Tech.*, vol. 42, no. 4, pp. 514–518, 1993.

[198] S. Markovich-Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Dallas, Texas, USA, Mar. 2010, pp. 201–204.

[199] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Sig. Proc.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[200] W. Herbordt and W. Kellermann, "Computationally efficient frequency-domain robust generalized sidelobe canceller," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, Darmstadt, Germany, Sept. 2001, pp. 51–54.

[201] G. Reuven, S. Gannot, and I. Cohen, "Dual-source transfer-function generalized sidelobe canceller," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 16, no. 4, pp. 711–727, May 2008.

[202] S. Markovich-Golan, S. Gannot, and I. Cohen, "A sparse blocking matrix for multiple constraints GSC beamformer," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Kyoto, Japan, Apr. 2012, pp. 197–200.

[203] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 19, no. 7, pp. 1900–1912, Sept. 2011.

[204] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Sig. Proc.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[205] R. Martin and T. Lotter, "Optimal recursive smoothing of non-stationary periodograms," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2001, pp. 167–170.

[206] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Sp. Au. Proc.*, vol. 9, no. 5, pp. 504–512, 2001.

[207] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Sp. Au. Proc.*, vol. 11, no. 5, pp. 466–475, 2003.

[208] T. Gerkmann and R. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*. IEEE, 2011, pp. 145–148.

[209] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 16, no. 5, pp. 910–919, 2008.

[210] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Sp., Sig. Proc.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[211] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Sig. Proc.*, vol. 81, no. 11, pp. 2403–2418, 2001.

[212] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[213] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 18, no. 5, pp. 1072–1077, 2010.

[214] M. Taseska and E. A. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, 2012, pp. 1–4.

[215] M. Taseska and E. Habets, "Spotforming using distributed microphone arrays," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, Oct. 2013, pp. 1–4.

[216] M. Taseska and E. Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 7, pp. 1195–1207, 2014.

[217] M. Taseska and E. A. Habets, "Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 24, no. 7, pp. 1291–1304, 2016.

[218] M. Taseska, S. Markovich-Golan, E. Habets, and S. Gannot, "Near-field source extraction using speech presence probabilities for ad hoc microphone arrays," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*. IEEE, 2014, pp. 169–173.

[219] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Sp. Au. Proc.*, vol. 12, no. 5, pp. 451–459, 2004.

[220] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Trans. Sig. Proc.*, vol. 60, pp. 233–246, Jan. 2012.

[221] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Brisbane, Australia, Apr. 2015.

[222] M. Ito, S. Araki, and T. Nakatani, "Permutation-free clustering of relative transfer function features for blind source separation," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2015, pp. 409–413.

[223] M. Taseska and E. A. P. Habets, "Relative transfer function estimation exploiting instantaneous signals and the signal subspace," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2015, pp. 404–408.

[224] S. Meier and W. Kellermann, "Analysis of the performance and limitations of ICA-based relative impulse response identification," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2015, pp. 414–418.

[225] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Phys. Rev. Lett.*, vol. 72, no. 23, pp. 3634–3637, 1994.

[226] M. Z. Ikram and D. R. Morgan, "A beamformer approach to permutation alignment for multichannel frequency-domain blind source separation," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2002, pp. 881–884.

[227] D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity," in *Proc. Intl. Symp. on Sig. Proc. and its App. (ISSPA)*, 2003, pp. II–73–II–76.

[228] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[229] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2012, pp. 261–264.

[230] R. Sakanashi, S. Miyabe, T. Yamada, and S. Makino, "Comparison of superimposition and sparse models in blind source separation by multichannel Wiener filter," in *Proc. Asia-Pacific Sig. and Info. Proc. Assoc.*, 2012, pp. 1–6.

[231] K. Adiloğlu and E. Vincent, "A general variational Bayesian framework for robust feature extraction in multisource recordings," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2012, pp. 273–276.

[232] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 5, pp. 971–982, May 2013.

[233] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 9, pp. 1913–1928, 2013.

[234] J. Thiemann and E. Vincent, "A fast EM algorithm for Gaussian model-based source separation," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2013.

[235] N. Ito, E. Vincent, T. Nakatani, N. Ono, S. Araki, and S. Sagayama, "Blind suppression of nonstationary diffuse noise based on spatial covariance matrix decomposition," *J. Sig. Proc. Sys.*, vol. 79, no. 2, pp. 145–157, 2015.

[236] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational Bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 8, pp. 1320–1335, Aug. 2014.

[237] K. Adiloğlu and E. Vincent, "Variational Bayesian inference for source separation and robust feature extraction," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, to appear.

[238] A. P. Dempster, N. M. Laird, and D. B. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B (Methodological)*, vol. 39, pp. 1–38, 1977.

[239] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Sig. Proc.*, vol. 92, no. 8, pp. 1950–1960, 2012.

[240] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 15, no. 5, pp. 1564–1578, July 2007.

[241] M. Togami, "Online speech source separation based on maximum likelihood of local Gaussian modeling," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2011, pp. 213–216.

[242] L. S. R. Simon and E. Vincent, "A general framework for online audio source separation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel-Aviv, Israel, Mar. 2012.

[243] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. Royal Stat. Soc. B*, vol. 46, no. 2, pp. 257–267, 1984.

[244] O. Cappé and E. Moulines, "On-line expectation-maximization algorithm for latent data models," *J. Royal Stat. Soc. B*, vol. 71, no. 3, pp. 593–613, 2009.

[245] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*, M. I. Jordan, Ed., pp. 355–368. MIT Press, Cambridge, MA, 2009.

[246] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 2, pp. 392–402, 2014.

[247] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 23, no. 2, pp. 394–406, 2015.

[248] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[249] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, Feb. 2004.

[250] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, New-York, USA, Apr. 1988, pp. 2578–2581.

[251] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Munich, Germany, Apr. 1997, pp. 21–24.

[252] C. Marro, Y. Mahieux, and K. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Sp. Au. Proc.*, vol. 6, no. 3, pp. 240–259, May 1998.

[253] S. Leukimmiatis, D. Dimitriadis, and P. Maragos, "An optimum microphone array post-filter for speech applications," in *Interspeech - Int. Conf. on Spoken Lang. Proc.*, 2006, pp. 2142–2145.

[254] R. Balan and J. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *Proc. IEEE Workshop on Sensor Array and Multichannel Sig. Proc. (SAM)*. IEEE, 2002, pp. 209–213.

[255] Y. Ephraim and D. Mala, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Sp., Sig. Proc.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[256] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *EURASIP J. on Adv. in Sig. Proc.*, vol. 2003, pp. 1064–1073, Oct. 2003.

[257] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Sp. Au. Proc.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[258] C. Zheng, H. Liu, R. Peng, and X. Li, "A statistical analysis of two-channel post-filter estimators in isotropic noise fields," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 2, pp. 336–342, Feb. 2013.

[259] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2004, pp. 832–839.

[260] E. Hoffmann, D. Kolossa, and R. Orglmeister, "Time frequency masking strategy for blind source separation of acoustic signals based on optimally-modified log-spectral amplitude estimator," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2009, pp. 581–588.

[261] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 6, pp. 1240–1250, June 2013.

[262] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2000, vol. 5, pp. 2985–2988.

[263] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, pp. 181–197. Springer, 2005.

[264] J. Mouba and S. Marchand, "A source localization/separation/respatialization system based on unsupervised classification of interaural cues," in *Proc. Conf. on Dig. Aud. Eff.*, 2006, pp. 233–238.

[265] H.-M. Park and R. M. Stern, "Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero-crossings," *Sp. Comm.*, vol. 51, no. 1, pp. 15–25, 2009.

[266] A. Deleforge, F. Forbes, and R. Horaud, "Variational EM for binaural sound-source separation and localization," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2013, pp. 76–80.

[267] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proc. Conf. on Dig. Aud. Eff.*, 2003.

[268] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 18, no. 2, pp. 382–394, 2010.

[269] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Syst., Man, Cybern.*, vol. 34, no. 4, pp. 1763–1773, 2004.

[270] A. Shamsoddini and P. Denbigh, "A sound segregation algorithm for reverberant conditions," *Sp. Comm.*, vol. 33, no. 3, pp. 179–196, 2001.

[271] N. Roman, S. Srinivasan, and D. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. of Am.*, vol. 120, no. 6, pp. 4040–4051, 2006.

[272] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2004, pp. 327–332.

[273] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Mar. 2012, pp. 409–412.

[274] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Fast joint DOA and pitch estimation using a broadband MVDR beamformer," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, Marrakech, Morocco, Sept. 2013.

[275] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. Intl. Symp. on Sig. Proc. and its App. (ISSPA)*, 2010, pp. 1–4.

[276] B. Weinstein, A. V. Oppenheim, M. Feder, and J. R. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Trans. Sig. Proc.*, vol. 42, no. 4, pp. 846–859, 1994.

[277] M. Feder, A. V. Oppenheim, and E. Weinstein, "Maximum likelihood noise cancellation using the em algorithm," *IEEE Trans. Acoust., Sp., Sig. Proc.*, vol. 37, no. 2, pp. 204–216, 1989.

[278] X. Sun and S. Douglas, "A natural gradient convolutive blind source separation algorithm for speech mixtures," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2001.

[279] M. Reyes-Gomez, B. Raj, and D. Ellis, "Multi-channel source separation by factorial HMMs," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2003.

[280] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, 2009.

[281] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[282] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 24, no. 8, pp. 1408–1423, Aug. 2016.

[283] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. Int. Conf. on Lang. Res. and Eval.*, 2000.

[284] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Sig. Proc.*, vol. 92, pp. 1928–1936, 2012.

[285] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. on Adv. in Sig. Proc.*, vol. 2009, pp. 6, 2009.

[286] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2013, pp. 126–130.

[287] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, Antibes - Juan les Pins, France, Sept. 2014.

[288] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 20, no. 5, pp. 1421–1447, Jul. 2012.

[289] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. of Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[290] D. R. Campbell, K. J. Palomäki, and G. J. Brown, "Roomsim, a MATLAB simulation of "shoebox" room acoustics for use in teaching and research," *J. of Comp. Info. Sys.*, vol. 9, no. 3, pp. 48–51, 2005.

[291] D. Jarrett, E. Habets, M. Thomas, and P. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *J. Acoust. Soc. of Am.*, vol. 132, no. 3, pp. 1462–1472, 2012.

[292] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (SMARD)," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, 2014, pp. 40–44.

[293] J. Le Roux and E. Vincent, "A categorization of robust speech processing datasets," Tech. Rep., Mitsubishi Electric Research Laboratories, Aug. 2014.

[294] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings: The AMI and AMIDA projects," in *Proc. Joint Workshop on Hands-free Sp. Comm. and Mic. Arrays*, 2008, pp. 115–118.

[295] A. Brutti, L. Cristoforetti, W. Kellermann, L. Marquardt, and M. Omologo, "WOZ acoustic data collection for interactive TV," in *Proc. Int. Conf. on Lang. Res. and Eval.*, 2008.

[296] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, "The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments," *Comp. Sp. and Lang.*, vol. 26, no. 1, pp. 52–66, 2011.

[297] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Automatic Sp. Recognition and Understanding*, 2015.

[298] J. Le Roux, E. Vincent, J. R. Hershey, and D. P. W. Ellis, "MICbots: collecting large realistic datasets for speech and audio research using mobile robots," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2015.

[299] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) — A new method for speech quality assessment of telephone networks and codecs," *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, vol. 2, pp. 749 – 752, 2001.

[300] R. Huber and B. Kollmeier, "PEMO-Q — A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[301] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[302] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 19, no. 7, pp. 2046–2057, Sept. 2011.

[303] ITU, "ITU-T Recommendation P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," 2003.

[304] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms," in *Proc. UK ICA Research Network Workshop*, 2006.

[305] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Comp. Sp. and Lang.*, vol. 27, no. 3, pp. 621–633, 2013.

[306] K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, B. Raj, R. Singh, and I. Tashev, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *Proc. Asia-Pacific Sig. and Info. Proc. Assoc.*, 2012, pp. 1–10.

[307] J. Thiemann and E. Vincent, "An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement," in *IEEE Int. Workshop on Machine Learning for Sig. Proc. (MLSP)*, 2013.

[308] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation*, 2007.

[309] E. Vincent, S. Araki, and P. Bofill., "The 2008 signal separation evaluati on campaign: A community-based approach to large-scale evaluation," in *Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 734–741.

[310] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Q. K. Duong, "The 2010 signal separation evaluation campaign (SiSEC 2010): Audio source separation," in *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 114–122.

[311] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011): - Audio source separation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, 2012, pp. 414–422.

[312] N. Ono, Z. Koldovský, S. Miyabe, and N. Ito, "The 2013 signal separation evaluation campaign," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, UK, Sept. 2013.

[313] N. Ono, D. Kitamura, Z. Rafii, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, Aug. 2015.

[314] M. I. Mandel and D. P. W. Ellis, "EM localization and separation using interaural level and phase cues," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2007.

[315] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Proc. Neural Info. Proc. Conf.*, 2007, pp. 953–960.

[316] Z. El Chami, A. Pham, Servière, and G. A. C., "A new model based underdetermined source separation," in *Proc. IWAENC*, 2008.

[317] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, Oct. 2007, pp. 139–142.

[318] J. Cho, J. Choi, and C. D. Yoo, "Underdetermined convolutive blind source separation using a novel mixing matrix estimation and MMSE-based source estimation," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2011)*, 2011.

[319] F. Nesta and M. Omologo, "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal correlation," in *Proceedings LVA/ICA*, 2012.

[320] J. Cho and C. Yoo, "Underdetermined convolutive BSS: Bayes risk minimization based on a mixture of super-gaussian posterior approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 828 – 839, 2015.

[321] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Sp. Au. Proc.*, vol. 1, no. 4, pp. 405–413, 1993.

[322] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP J. on Appl. Sig. Proc.*, vol. 11, pp. 1157–1166, 2003.

[323] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proceedings F (Radar and Signal Processing)*, vol. 140, no. 6, pp. 362–370, 1993.

[324] S. Y. Low and S. Nordholm, "A hybrid speech enhancement system employing blind source separation and adaptive noise cancellation," in *Proceedings of the 6th Nordic Signal Processing Symposium (NORSIG)*, 2004.

[325] K. Reindl, S. Meier, H. Barfuss, and W. Kellermann, "Minimum mutual information-based linearly constrained broadband signal extraction," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 6, pp. 1096–1108, June 2014.

[326] H. Buchner, "A systematic approach to incorporate deterministic prior knowledge in broadband adaptive mimo systems," in *ASILOMAR*, 2010, pp. 461–468.

[327] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Combined LCMV-TRINICON beamforming for separating multiple speech sources in noisy and reverberant environments," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, 2016, to be published.

[328] S. Araki, M. Okada, T. Higuchi, A. Ogawa, and T. Nakatani, "Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Mar. 2016, pp. 385–389.

[329] S. Araki and T. Nakatani, "Hybrid approach for multichannel source separation combining time-frequency mask with multi-channel Wiener filter," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, May 2011, pp. 225–228.

[330] A. Asaei, M. E. Davies, H. Bourlard, and V. Cevher, "Computational methods for structured sparse component analysis of convolutive speech mixtures," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2012, pp. 2425–2428.

[331] Z. Koldovský, J. Málek, P. Tichavský, and F. Nesta, "Semi-blind noise extraction using partially known position of the target source," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 10, pp. 2029–2041, Oct. 2013.

[332] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 11, pp. 2013–2312, Nov. 2013.

[333] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 3, pp. 620–633, Mar. 2014.

[334] R. Mignot, G. Chardon, and L. Daudet, "Low frequency interpolation of room impulse responses using compressed sensing," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 1, pp. 205–216, Jan. 2014.

[335] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International journal of neural systems*, vol. 25, no. 1, 2015.

[336] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 24, no. 8, pp. 1393–1407, Aug. 2016.

[337] B. Laufer, R. Talmon, and S. Gannot, "Manifold-based Bayesian interference for semi-supervised source localization," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Shanghai, China, Mar. 2016.

[338] Laufer-Goldshtein, Bracha, R. Talmon, and S. Gannot, "Semi-supervised source localization on multiple-manifolds with distributed microphones," *arXiv preprint arXiv:1610.04770*, 2016.

[339] R. Talmon and S. Gannot, "Relative transfer function identification on manifolds for supervised GSC beamformers," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, Marrakech, Morocco, Sept. 2013.

[340] Y. Jiang, D. L. Wang, R. S. Liu, and Z. M. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 12, pp. 2112–2121, Dec. 2014.

[341] J. Woodruff and D. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 4, pp. 806–815, Apr. 2013.

[342] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Workshop Automatic Sp. Recognition and Understanding*, 2015.

[343] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, to appear.

[344] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *Proc. IEEE Workshop Automatic Sp. Recognition and Understanding*, 2015.

[345] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Sig. Proc.*, vol. 107, pp. 4–20, 2015.

[346] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 17, no. 1, pp. 38–51, Jan. 2009.

[347] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks – part I: sequential node updating," *IEEE Trans. Sig. Proc.*, vol. 58, pp. 5277–5291, 2010.

[348] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 2, pp. 343–356, Feb. 2013.

[349] S. Markovich-Golan, S. Gannot, and I. Cohen, "Performance of the SDW-MWF with randomly located microphones in a reverberant enclosure," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 7, pp. 1513–1523, 2013.

[350] S. Markovich-Golan, S. Gannot, and I. Cohen, "Low-complexity addition or removal of sensors/constraints in LCMV beamformers," *IEEE Trans. Sig. Proc.*, vol. 60, no. 3, pp. 1205–1214, Mar. 2012.

[351] Y. Zeng and R. Hendriks, "Distributed delay and sum beamformer for speech enhancement via randomized gossip," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 1, pp. 260–273, Jan. 2014.

[352] R. Heusdens, G. Zhang, R. C. Hendriks, Y. Zeng, and W. B. Kleijn, "Distributed MVDR beamforming for (wireless) microphone networks using message passing," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, 2012, pp. 1–4.

[353] M. O'Connor and W. B. Kleijn, "Diffusion-based distributed MVDR beamformer," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*. IEEE, 2014, pp. 810–814.

[354] N. D. Gaubitch, J. Martinez, W. B. Kleijn, and R. Heusdens, "On near-field beamforming with smartphone-based ad-hoc microphone arrays," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, Sept. 2014, pp. 94–98.

[355] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, "Location feature integration for clustering-based speech separation in distributed microphone arrays," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 2, pp. 354–367, Feb. 2014.

[356] P. Pertilä, M. S. Hämäläinen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.

[357] S. Wehr, I. Kozintsev, R. Lienhart, and W. Kellermann, "Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation," in *Proc. IEEE Intl. Symp. on Multimedia Software Eng.* IEEE, 2004, pp. 18–25.

[358] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, Aachen, Germany, Sept. 2012, Final list for best student paper award.

[359] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A gossiping approach to sampling clock synchronization in wireless acoustic sensor networks," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, May 2014, pp. 7575–7579.

[360] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enchantment," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, Antibes - Juan les Pins, France, Sept. 2014.

[361] Y. Zeng, R. Hendriks, and N. Gaubitch, "On clock synchronization for multi-microphone speech processing in wireless acoustic sensor networks," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Apr. 2015, pp. 231–235.

[362] D. Cherkassky, S. Markovich-Golan, and S. Gannot, "Performance analysis of MVDR beamformer in WASN with sampling rate offsets and blind synchronization," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, Nice, France, Aug. 2015.

[363] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A combined hardware–software approach for acoustic sensor network synchronization," *Sig. Proc.*, vol. 107, pp. 171–184, 2015.

[364] L. Wang and S. Doclo, "Correlation maximization based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, 2016.

[365] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, Aug. 2016, accepted for publication with mandatory minor revisions.

[366] C. Anderson, P. Teal, and M. Poletti, "Spatially robust far-field beamforming using the von Mises(-Fisher) distribution," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 23, no. 12, pp. 2189–2197, Dec. 2015.

[367] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE Trans. Sig. Proc.*, vol. 51, no. 10, pp. 2511–2526, Oct. 2003.

[368] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Acoust., Sp., Sig. Proc.*, vol. 15, no. 2, pp. 617–631, Feb. 2007.

[369] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Sig. Proc.*, vol. 51, no. 7, pp. 1702–1715, July 2003.

[370] S. Vorobyov, A. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Sig. Proc.*, vol. 51, no. 2, pp. 313–324, Feb. 2003.

[371] R. Lorenz and S. Boyd, "Robust minimum variance beamforming," *IEEE Trans. Sig. Proc.*, vol. 53, no. 5, pp. 1684–1696, May 2005.

[372] S. Nordebo, I. Claesson, and S. Nordholm, "Adaptive beamforming: spatial filter designed blocking matrix," *IEEE J. Ocean. Eng.*, vol. 19, no. 4, pp. 583–590, Apr. 1994.

[373] C. A. Anderson, S. Meier, W. Kellermann, P. D. Teal, and M. A. Poletti, "TRINICON-BSS system incorporating robust dual beamformers for noise reduction," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*. IEEE, 2015, pp. 529–533.

[374] O. Thiergart, M. Taseska, and E. A. Habets, "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*. IEEE, 2013, pp. 659–663.

[375] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Sig. Proc.*, vol. 43, no. 1, pp. 95–107, Jan. 1995.

[376] B. Kollmeier, J. Peissig, and V. Hohmann, "Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain," *Scandinavian Audiology. Supplementum*, vol. 38, pp. 28, 1993.

[377] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Sp. Comm.*, vol. 39, no. 1, pp. 111–138, 2003.

[378] J. Li, M. Akagi, and Y. Suzuki, "Extension of the two-microphone noise reduction method for binaural hearing aids," in *Proc. Inter. Conf. on Au., Lang. and Image Proc.*, 2008, pp. 97–101.

[379] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1994.

[380] S. Wehr, M. Zourub, R. Aichner, and W. Kellermann, "Post-processing for BSS algorithms to recover spatial cues," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, Paris, France, Sep. 2006.

[381] R. Aichner, H. Buchner, M. Zourub, and W. Kellermann, "Multi-channel source separation preserving spatial information," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, Honolulu HI, USA, Apr. 2007, pp. 5–8.

[382] K. Reindl, Y. Zheng, and W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," in *Proc. Intl. Symp. on Control, Commmunications and Signal Processing*, Mar. 2010, pp. 1–6.

[383] S. Doclo, R. Dong, T. Klasen, J. Wouters, S. Haykin, and M. Moonen, "Extension of the multi-channel Wiener filter with ITD cues for noise reduction in binaural hearing aids," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, 2005, pp. 70–73.

[384] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. on Adv. in Sig. Proc.*, pp. 175–175, Jan. 2006.

[385] S. Markovich-Golan, S. Gannot, and I. Cohen, "A reduced bandwidth binaural MVDR beamformer," in *Proc. Intl. Workshop Acoust. Sig. Enh. (IWAENC)*, Tel-Aviv, Israel, Sept. 2010.

[386] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, Dec. 2015, accepted for publication.

[387] I. Almajai and B. Milner, "Visually derived Wiener filters for speech enhancement," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 19, no. 6, pp. 1642–1651, Aug. 2011.

[388] V. Khalidov, F. Forbes, and R. Horaud, "Conjugate mixture models for clustering multimodal data," *Neural Computation*, vol. 23, no. 2, pp. 517–557, 2011.

[389] M. S. Khan, S. M. Naqvi, A. ur Rehman, W. Wang, and J. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 21, no. 9, pp. 1900–1912, Sept. 2013.

[390] I.-D. Gebru, X. Alameda-Pineda, R. Horaud, and F. Forbes, "Audio-visual speaker localization via weighted clustering," in *IEEE Int. Workshop on Machine Learning for Sig. Proc. (MLSP)*, Reims, France, Sept. 2014, pp. 1–6.

[391] A. Deleforge, R. Horaud, Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 23, no. 4, pp. 718–731, Apr. 2015.

[392] S. Zeiler, H. Meutzner, A. H. Abdelaziz, and D. Kolossa, "Introducing the Turbo-Twin-HMM for audio-visual speech enhancement," *Proc. Interspeech*, pp. 1750–1754, 2016.

[393] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 23, no. 4, pp. 732–745, 2015.