Audio source separation into the wild

Laurent Girin, Sharon Gannot, Xiaofei Li

Audio Source Separation into the Wild*

Laurent Girin[1,2], Sharon Gannot[3] and Xiaofei Li[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab

38402 Saint Martin d'Hères, France

[2] INRIA Grenoble Rhône-Alpes, Perception group

655 Avenue de l'Europe, 38330 Montbonnot-Saint-Martin, France

[3] Bar-Ilan University, Faculty of Engineering

5290002 Ramat-Gan, Israel

Corresponding: laurent.girin@gipsa-lab.grenoble-inp.fr

Abstract

This review chapter is dedicated to multichannel audio source separation in real-life environment. We explore some of the major achievements in the field and discuss some of the remaining challenges. We will explore several important practical scenarios, e.g. moving sources and/or microphones, varying number of sources and sensors, high reverberation levels, spatially diffuse sources, and synchronization problems. Several applications such as smart assistants, cellular phones, hearing aids and robots, will be discussed. Our perspectives on the future of the field will be given as concluding remarks of this chapter.

Keywords: Multichannel audio source separation, Beamforming for Audio signals, smart devices, hearing aids

1 Introduction

Source separation is a topic of signal processing that has been of major interest for decades. It consists of processing an observed mixture of signals so that to extract the elementary signals composing this mixture. In the context of audio processing, it refers to the extraction of signals simultaneously emitted by several sound sources, from the audio recordings of the resulting mixture signal. It has major applications, going from speech enhancement as a front-end for telecommunication systems and automatic speech recognition, to demixing and remixing of music. Despite a recent progress using deep learning techniques, single-channel multi-source recordings are still regarded particularly difficult to separate. In this chapter we deal with audio source separation in the wild, and

*In *Multimodal behavior analysis in the wild*, chapter 3, X. Alameda-Pineda, E. Ricci, N. Sebe editors, Academic Press, 2018, pages 53-78

we address multichannel recordings obtained using multiple microphones in a natural environment, as opposed to mixtures created by mixing software which generally do not match the acoustics of real environments, e.g. music production in studio. Typically, we discuss such problems as having to separate the speech signals emitted simultaneously by different persons sharing the same acoustic enclosure, considering also ambient noise and other interfering sources such as a domestic apparatuses.

Even when using multichannel recordings, source separation in general is a difficult problem that belongs to the general class of inverse problem. As such, it is often ill-posed, in particular when the number of sensors used to capture the mixture signals is lower than the number of emitting sources. Consequently, in the signal processing community in general, and in the audio processing community in particular, the source separation problem has often been addressed within quite controlled configurations, i.e. "laboratory" studies, that are carefully designed to allow a proper evaluation protocol and an in-depth inspection of the behavior of the proposed techniques. As we will detail in this chapter, this often comes in contrast to robust, into-the-wild configurations, where the source separation algorithms are confronted with the complexity of real-world data, and thus "do not work so well". In this chapter, we describe such "limitations" of multichannel audio source separation (MASS) and make a review of the studies that have been proposed to overcome those limitations, trying to make MASS techniques progressively go from laboratories into the wild.[1]

Research in speech enhancement and speaker separation has followed two convergent paths, starting with microphone array processing (also referred to as *beamforming*) and blind source separation, respectively. These communities are now strongly interrelated and routinely borrow ideas from each other. Hence, in this chapter we discuss the two paradigms interchangeably. We will explore several important practical scenarios, e.g. moving sources and/or microphones, varying number of sources and sensors, high reverberation levels, spatially diffuse sources, and synchronization problems. Several applications such as smart assistants, cellular phones, hearing devices and robots, which have recently gained a growing research and industrial interest, will be discussed.

This chapter is organized as follows. In Section 2, we briefly present the fundamentals of multichannel audio source separation. In Section 3 we list the current major limitations of MASS that prevent a large deployment of MASS technique into the wild, and we present approaches that have been proposed in the literature to overcome these limitations.

2 Multichannel audio source separation

In this section, we briefly present the fundamentals of multichannel audio source separation. This presentation is limited to the basic material that is necessary

[1] Of course, this is just a general view of the academic studies in the field as a whole. Some researchers in the field have considered with great attention the practical aspects of audio source separation techniques.

to understand the following discussion on MASS into the wild. Indeed, the goal of this chapter is not to extensively present the theoretical foundations and principles of source separation and beamforming, even limited to the audio context: Many publications have addressed this issue, including books [61, 33, 89, 13, 26] and overview papers [20, 106, 48].

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 [49, 18, 14, 84, 25], while MASS was later popularized by the machine learning community and it addressed "cocktail party" scenarios involving several sound sources mixed together [106, 89, 113, 26, 137, 136]. These two research tracks have converged in the last decade and they are hardly distinguishable today. 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.

The formalization of the MASS problem begins with the formalization of the mixture signal. The most general expression for a linear mixture of J source signals recorded by I microphones is:

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{y}_j(t) + \mathbf{b}(t) \ \in \mathbb{R}^I, \tag{1}$$

where $\mathbf{y}_j(t) \in \mathbb{R}^I$ is the multichannel image of the j-th source signal $s_j(t)$ [128], taking into account the effect of acoustic propagation from the position of the emitted source to the microphones (each entry $y_{ij}(t)$ of $\mathbf{y}_j(t)$ is the image of $s_j(t)$ at microphone i). $\mathbf{b}(t)$ is a sensor noise term. In most studies on MASS, the effect of acoustic propagation from source j to microphone i is modelled as a linear time-invariant filter of impulse response $a_{ij}(t)$, and we have:

$$\mathbf{x}(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{L_a-1} \mathbf{a}_j(\tau) s_j(t-\tau) + \mathbf{b}(t). \tag{2}$$

The vector $\mathbf{a}_j(\tau)$ contains all responses $a_{ij}(t)$ for $i \in [1, I]$, which are assumed to have the same length L_a for convenience. Depending on the application, the goal of MASS is to estimate either the source images $\mathbf{y}_j(t)$ or the (monochannel) source signals $s_j(t)$ from the observation of $\mathbf{x}(t)$.

State-of-the-art MASS methods generally start with a time-frequency (TF) decomposition of the temporal signals, usually by applying the short-time Fourier transform (STFT) [29]. This is for two main reasons. First, model-based approaches can take advantage of the very particular *sparse* structure of audio signals in the TF plane [115]: A small proportion of source TF coefficients have a significant energy. Source signals are thus generally much less overlapping in the TF domain than in the time-domain, naturally facilitating the separation.

Second, it is common to consider that at each frequency, the time-domain convolutive mixing process (2) is transformed by the STFT into a simple product between the source STFT coefficients and the discrete Fourier transform (DFT) coefficients of the mixing filter, see e.g. [110, 147, 146, 4, 91, 107, 109] and many other studies:

$$\mathbf{x}(f,n) \approx \sum_{j=1}^{J} \mathbf{a}_j(f) s_j(f,n) + \mathbf{b}(f,n) = \mathbf{A}(f)\mathbf{s}(f,n) + \mathbf{b}(f,n), \tag{3}$$

where $\mathbf{x}(f,n)$, $s_j(f,n)$ and $\mathbf{b}(f,n)$ are the STFT of $\mathbf{x}(\tau)$, $\mathbf{b}(\tau)$ and $s_j(\tau)$, respectively, and $\mathbf{a}_j(f)$ gathers the DFT of the entries of $\mathbf{a}_j(\tau)$, known as the acoustic transfer functions (ATFs). The ATF vectors are concatenated in the matrix $\mathbf{A}(f)$ and the source signals $s_j(f,n)$ are stacked in the vector $\mathbf{s}(f,n)$. In many practical scenarios, $\mathbf{A}(f)$ is substituted by the respective relative transfer function (RTF) matrix, for which each column is normalized by its first entry. Under this normalization the first row of $\mathbf{A}(f)$ is all '1's and the source signals $s_j(f,n)$ are substituted by their image on the first microphone.[2]

As further discussed in Section 3.3, (3) is an approximation that is valid if the length of the mixing filters impulse responses is shorter than the length of the STFT window analysis. In the literature and in the following this approximation is referred to as the multiplicative transfer function (MTF) approximation [9] or the narrowband approximation [71].

MASS methods can then be classified into four (non-exclusive) categories [137, 48]. Firstly, separation methods based on independent component analysis (ICA) consist in estimating the demixing filters that maximize the independency of separated sources [26, 61]. TF-domain ICA methods have been largely investigated [127, 110, 103]. Unfortunately, ICA-based methods are subject to the well-known scale ambiguity and source permutation problems across frequency bins [3], which must generally be solved as a post-processing step [62, 120, 119]. In addition, these methods cannot be directly applied to underdetermined mixtures.

Secondly, methods based on sparse component analysis (SCA) and binary masking rely on the assumption that only one source is active at each TF point [118, 146, 4, 91] (most methods consider only one active source at each TF bin though in principle the SCA and ICA approaches can be combined by considering up to I active sources in each TF bin). These methods often rely on some sort of source clustering in the TF domain, generally based on spatial information extracted from prior mixing filter identification. Therefore, for this kind of methods, the source separation problem is often linked to the source localization problem (where are the emitting sources?).

Thirdly, more recent methods are based on probabilistic generative models in the STFT domain and associated parameter estimation and source inference

[2] One can select other microphones as the reference microphone or use other methods of normalization. Selecting the reference microphone might have an impact on the MASS performance. This topic is out of the scope of this chapter.

algorithms [137]. The latter are mostly based on the well-known expectation-maximization (EM) methodology [30] and the like (i.e. iterative alternating optimization techniques). One popular approach is to model the source STFT coefficients with a complex-valued local Gaussian model (LGM) [45, 37, 148, 82], often combined with a nonnegative matrix factorization (NMF) model [74] applied to the source PSD matrix [44, 107, 6, 109], which is reminiscent of pioneering works such as [12]. This allows one to drastically reduce the number of model parameters and (to some extent) to alleviate the source permutation problem. The sound sources are generally separated using Wiener filters constructed from the learned parameters. Such approach has been extended to super-Gaussian (or heavy-tailed) distributions to model the audio signal sparsity in the TF domain [98], as well as to a fully Bayesian framework by considering prior distributions for the (source and/or mixing process) model parameters [66]. Note that, as for SCA/binary masking methods, the estimation of mixing parameters and the source separation itself are two different steps, often processed alternately within the EM principled methodology.

Methods belonging to the fourth category can be broadly classified as beamforming methods, which is roughly equivalent to *linear spatial filtering*. A beamformer is a vector $\mathbf{w}(f) = [w_1(f), \ldots, w_I(f)]^T$ comprising one complex-valued weight per microphone, that is applied to $\mathbf{x}(n, f)$. The output $\mathbf{w}^H(f)\,\mathbf{x}(n, f)$ can be transformed back into the time-domain by an inverse STFT. Beamformers originally referred to spatial filters based on the direction of arrival (DOA) of the source signal and were only later generalized to any linear spatial filters. DOA-based beamformers are still widely-used, especially when simplicity of the implementation and its robustness are of crucial importance. However, the performance of these beamformers is expected to degrade in comparison with modern beamformers that take the entire acoustic path into account [46]. The beamformer weights are set according to a specific optimization criterion. Many beamforming criteria can be found in the general literature [134]. In the speech processing community, the minimum variance distortionless response (MVDR) beamformer [46], the maximum signal-to-noise ratio (MSNR) beamformer [142], the multichannel Wiener filter (MWF) [34], specifically its speech distortion weighted variant (SDW-MWF) [35], and the linearly constrained minimum variance (LCMV) beamformer [92], are widely used.

One may state that, so far, the beamforming approach led to more effective industrial real-world applications than the "generic" MASS approach. This lies in the difference between a) enhancing a spatially fixed dominant target speech signal from background noise (possibly composed of several sources) and b) clearly separating several source signals, all considered as signals of interest, that are mixed with a similar power (the cocktail party problem). The first problem is generally simpler to solve than the second one. In other words, the second one can be seen as an extension of the first one. Anyway, as we will see now, problems of into-the-wild processing remain challenging in each case.

3 Making MASS go from labs into the wild

3.1 Moving sources and sensors

In a realistic into-the-wild scenario, sound sources are often moving. Sometimes they are moving slightly, e.g. the small movements of the head of a sat speaker. Sometimes they are moving a lot, e.g. a person speaking while walking in a room. Sensors can also move, e.g. microphones embedded within a mobile robot. In the same vein, the acoustic environment can also change over time, e.g. we close a window, an object is placed in between a source and the microphones, or the separation system has to operate in another room. All those changes imply changes in the acoustic propagation of sources to microphones, i.e. changes in the mixing filters. Yet the vast majority of MASS methods described in the signal processing literature deals with the assumption of fixed sources, fixed sensors and fixed environment, i.e. technically speaking the mixing filters are considered as *time-invariant* (at least over the duration of the processed recordings), as expressed in (2). Into-the-wild MASS methods should consider the more realistic case of *time-varying* mixtures corresponding to source-to-microphone channels that can change over time, which would account for possible source or microphone motions and environment changes. For example, in many Human-robot interaction scenarios, there is a strong need to consider mixed speech signals emitted by moving speakers and perturbed by reverberation that can change over time and by non-stationary ambient noise.

Studies dealing with moving sources or moving sensors or changing environment actually exist, but they are quite sparse compared to the large number of studies with time-invariant filters. Early attempts addressing the separation of time-varying mixtures basically consisted in block-wise adaptations of time-invariant methods: An STFT frame sequence is split into blocks, and a time-invariant MASS algorithm is applied to each block. Hence, block-wise adaptations assume time-invariant filters within blocks. The separation parameters are updated from one block to the next and the separation result over a block can be used to initialize the separation of the next block. Frame-wise algorithms can be considered as particular cases of block-wise algorithms, with single-frame blocks, and hybrid methods may combine block-wise and frame-wise processing. Notice that, depending on the implementation, some of these methods may run online.

Interestingly, most of the block-wise approaches use ICA, either in the temporal domain [70, 59, 1, 116] or in the Fourier domain [99, 100]. In addition to being limited to overdetermined mixtures, block-wise ICA methods need to account for the source permutation problem, not only across frequency bins, as usual, but across successive blocks as well. Examples of block-wise adaptation of binary-masking or LGM-based methods are more scarce. As for binary masking, a block-wise adaptation of [4] was proposed in [83]. This method performs source separation by clustering the observation vectors in the source image space. As for LGM, [126] describes an online block- and frame-wise adaptation of the general LGM framework proposed in [109]. One important

problem, common to all block-wise approaches, is the difficulty to choose the block size. Indeed, the block size must assume a good trade-off between local channel stationarity (short blocks) and sufficient data to infer relevant statistics (long blocks). The latter constraint can drastically limit the dynamics of either the sources or the sensors [83]. Other parameters such as the step-size of the iterative update equations may also be difficult to set [126]. In general, systematic convergence towards a good separation solution using a limited amount of signal statistics remains an open issue.

A more principled approach consists in modeling the mixing filter as a time-varying process and considering the MASS in the angle of an adaptive process, in the spirit of early works on adaptive filtering [145]. For example, an early iterative and sequential approach for speech enhancement in reverberant environment was proposed in [144]. This method used the EM framework to jointly estimate the desired speech signal and the required (deterministic) parameters, namely the speech auto-regressive coefficients, and the speech and noise mixing filters taps. Only the case of a 2×2 mixture was addressed. A subspace tracking recursive LCMV beamforming method for extracting multiple moving sources was proposed in [94]. This method is applicable only to over-determined mixture.

As for under-determined time-varying convolutive mixtures, a method using binary masking within a probabilistic LGM framework was proposed in [57]. The mixing filters were considered as latent variables following a Gaussian distribution with mean vector depending on the DOA of the corresponding source. The DOA was modeled as a discrete latent variable taking values from a finite set of angles and following a discrete hidden Markov model (HMM). A variational expectation-maximization (VEM) algorithm was derived to perform the inference, including forward-backward equations to estimate the DOA sequence.

In [65, 67], the transfer function of the mixing filters were considered as continuous latent variables ruled by a first-order linear dynamical system (LDS) with Gaussian noise [17], in the spirit of [47]. This model was used in combination with a source LGM-with-NMF model, still to process underdetermined time-varying convolutive mixtures. This approach may be seen as a generalization of [107] to moving sources/microphones. As in [57], a VEM algorithm was developed for the joint estimation of the model parameters and inference of the latent variables. Here, a Kalman smoother was used for the inference of the time-varying mixing filters, which were combined with estimated source PSDs to build separating Wiener filters. This model can be more effective than the discrete DOA-dependent HMM model of [57] in reverberant conditions, since the relationship between the transfer function and the source DOA can be quite complex, and Wiener filters are more general than binary masks.

A VEM approach for beamforming (hence, over-determined scenario) that is specifically designed for dynamic scenarios can be found in [90, 121, 73]. In these methods, the speech signal is modelled as a an LGM in the STFT domain and the RTFs as a first-order Markov model. The posterior distribution of the speech signal and the channel is recursively estimated in the E-step.

In comparison to the block-wise adaptation methodology described in [126],

explicit time-varying mixture models have the potential to exploit the information available within the whole sequence of input mixture frames. They were generally proposed in batch mode but they can be adapted to online processing, e.g. by replacing a Kalman smoother with a Kalman filter [144].

DOA estimation plays an important role in the design of beamforming methods. DOA tracking of multiple speakers based on a discrete set of angles and HMM is given in [117]. DOA estimation procedures for single source (or alternating sources) scenarios, which are based on variants of the recursive least squares (RLS) methodology, are presented in [39]. Recursive versions of the EM procedure are utilized for multiple speakers tracking in [123]. A simple, yet effective method for localizing multiple sources in reverberant environments is the steered-response-power with phase transform (SRP-PHAT) [31]. The MUltiple Signal Classification (MUSIC) algorithm [122] or more specifically, the root-MUSIC variant, allow for fast adaptation of LCMV beamformers by exploiting instantaneous DOA estimates [131, 132].

Finally, we can mention that robots, as a moving platforms, open new opportunities and challenges for the sound source localization and separation tasks. An open source robot audition system, named 'HARK', is described in [101]. The localization module is based on successive application of the MUSIC algorithm [122], while the separation stage uses a geometry-assisted MASS method [111]. Bayesian methods for tracking multiple sources are also gaining interest in the literature, e.g. using particle filters [133, 40] and probability hypothesis density (PHD) filters [41]. Two recent European projects, the "Ears"[3] project and the "Two!Ears"[4] project, explored new algorithms for enhancing the auditory capabilities of humanoid robots [85] and link them with decision and action [19].

3.2 Varying number of (active) sources

In a realistic into-the-wild scenario, sound sources are often intermittent, i.e. they do not emit sounds all the time. As an example of major importance we can mention a natural conversational regime between several speakers that includes speech turns. Depending on the context and content of the conversation, the speech signals can have very low to very strong time overlap. The sound sources may not even be present in the scene all the time, e.g. a person that goes in and out of a room, occasionally speaking or turning on and off a sounding device. Yet, the vast majority of MASS methods described in the signal processing literature deals with the assumption of a fixed number of sources over time. In addition, this fixed number of sound sources is often assumed to be known and all sources are assumed to be continuously active, i.e. they emit during all the time of the processed recording sequence. The situation of the literature with this respect is similar to the previous section: A few studies with the number of active sources varying in time do exist but they are largely outnumbered by the studies with fixed number of constantly active sources.

[3] `https://robot-ears.eu/`
[4] http://twoears.eu/project/

One straightforward manner to address this problem is to proceed to the estimation of the number of sources present in the scene and/or the number of active sources as a pre-processing step before going into the separation problem. A method based on speech sparsity in the STFT domain is presented in [5]. A variational EM approach for complex Watson mixture models is presented in [36]. In the beamforming context, identifying the number of speakers, as well as the number of available sensors, necessitates an update of the weights of the beamformer. An efficient implementation, based on low-rank update of correlation matrices, is presented in [95].

The detection of the number of active source and associating an "identity" (i.e. a label) to each detected source is related to the so-called *diarization* problem. Indeed, in speech processing, speaker diarization refers to the task of detecting who speaks and when in an audio stream [2, 135]. In many (dialog) applications, the speakers are assumed to take distinct speech turns, i.e. they speak one after the other, and speaker diarization thus amounts to signal segmentation and speaker recognition. In a source separation context, the automatic detection of the number and "identity" of simultaneously active sources can thus be considered as an additional multisource diarization task to be considered jointly with the separation task, or within the separation task. Indeed, both processes are complementary: Knowing the source diarization is assumed to ease the separation process, for example by enabling to adapt the separation system to the actual number of active sources and to the speaker characteristics; in turn diarization is easier using separated source signals than using mixed source signals.

Processing of speech intermittency for MASS appears in [26] for the instantaneous mixing case. For convolutive mixtures, [108, 58] presented a framework for joint processing of MASS and diarization, where factorial Hidden Markov models were used to model the activity of the sources. Unfortunately, due to its factorial nature, the model does not account for correlations on the activity of different sources, i.e. the activities of the different sources are assumed independent with each other, which is questionable for natural conversations for example. Very recently, joint processing of the two tasks have been proposed in [64] (for over-determined mixtures) and in [68, 69]. The models in [68] and [69] combine a diarization state model (that encodes the combination of active sources within a given set of maximum size N) with the multichannel LGM+NMF model of [107] and with the full-rank spatial covariance matrix model of [37], respectively. In contrast to [108, 58], modelling the activity of all sources jointly using a diarization state enables to exploit the potential correlations on speaker activity.

Note that estimating the number of active sources present in the scene is a good example of problem common to source separation and source localization. Strategies have been developed for the automatic detection of the number of sources within a (probabilistic) source localization framework, e.g. [42, 141, 81]. Obviously, such strategies may be exploited in or extended to source separation, as already explored in [118, 117].

3.3 Spatially diffuse sources and long mixing filters

The vast majority of current state-of-the-art MASS methods considers convolutive mixtures of sources, as expressed by (2): each source image within the mixture signal recorded at the microphones is assumed to be the result of the convolution of a small "point" source signal with the impulse response of the source-to-microphone acoustic path. This formulation implies that each sound source is assumed to be spatially concentrated at a single point of the acoustic space. This is fine to some extent for speech signals, but this is more questionable for "large" sound sources such as wind, trucks, or large musical instruments, which emit sound in a large region of space. In that later case, each source image is better considered as a spatially distributed source.

Moreover, if the source signal propagates in highly reverberant enclosures, the late tail of the room impulse response (RIR) is perceived as arriving from all directions. If the reverberation time T_{60}, defined as the elapsed time until the reverberation level has decreased by 60 dB from its initial value, the reverberant sound field is said to be diffuse, homogenous and isotropic. The normalized spatial correlation between the received signal at two different microphones i and i', and at frequency f, is given in closed-form for spherical symmetry [28, 72]:

$$\begin{aligned}\Omega_{ii'}(f) &= \frac{\mathbb{E}^{\text{spat}}(r_{ij}(f)r^*_{i'j}(f))}{\sqrt{\mathbb{E}^{\text{spat}}(|r_{ij}(f)|^2)}\sqrt{\mathbb{E}^{\text{spat}}(|r_{i'j}(f)|^2)}} \\ &= \frac{\sin(2\pi f \ell_{ii'}/c)}{2\pi f \ell_{ii'}/c} \end{aligned} \tag{4}$$

where $\mathbb{E}^{\text{spat}}$ denotes *spatial expectation* over all possible absolute positions of the sources and of the microphone array in the room, and $\ell_{ii'}$ denotes the distance between the microphones. A closed-form result also exists for cylindrical symmetry [27]. A simulator of both sound fields can be found in [53].

To address this problem, the authors of [37] proposed to use a full-rank (FR) spatial covariance matrix (SCM) for characterizing the spatial distribution of the source images (across channels), instead of the rank-1 matrix corresponding to the MTF model [107]. This FR-SCM model is assumed to represent diffuse sources better than the MTF approximation and it is compliant with the vision of a diffuse source as a sum of subsources with identical PSD distributed in a large region of the physical space. The model parameters are estimated using an expectation-maximization (EM) algorithm. Note that this approach does not attempt to explicitly model the mixture process but rather focuses on the properties of the resulting source images. The FR-SCM model was further used and improved in [6, 38].

Moreover, even for point sources, the processing of convolutive mixtures in the STFT domain is confronted to a severe limitation with respect to into-the-wild scenarios: The length of room impulse responses (RIRs) is in general (much) larger than the length of the STFT analysis window, which can severely question the validity of the approximation (3). Typical values for the STFT window length for speech mixtures are within 16-64 ms, in order to adapt to

the global non-stationarity / local stationarity of speech signals. At the same time, the typical T_{60} reverberation time of usual home/office rooms is within 200-600 ms, and large meeting rooms or auditorium can have a T_{60} larger than 1 s. The ratio between RIR length and STFT window length can thus easily be within 10-50 instead of being lower than 1. Therefore, (3) can be a quite poor approximation, even for moderately reverberant environments, if the sources are positioned further away from the microphones.[5] The MTF approximation is still widely used to address convolutive mixtures problems due to its practical interest: The fact that a time-domain convolutive mixture becomes an independent instantaneous mixture at each frequency bin f facilitates the technical development of solutions to the separation problem. While this can be a reasonable choice of model, we stress that the validity of the MTF approximation should be verified prior to its application. In some cases, a mixed MTF and full-rank models should be considered [37].

Here again, compared to the impressive amount of papers on MASS methods for convolutive mixtures based on (3), only a few have addressed solutions to overcome the limitation of the MTF approximation. Although they are only a few, these solutions can be classified into two general approaches: Methods mixing time-domain (TD) and TF-domain processing, and methods that totally remain in the TF-domain.

As for the first approach, the method of [71] consists in modeling the sources in the TF domain while keeping a TD representation of the convolutive mixture using the inverse transform expression. The TD source signals are estimated using a Lasso optimization technique (hence the method is called W-Lasso for wideband Lasso) with a regularization term on STFT source coefficients to take into account source sparsity in the TF domain. In [7] an improved W-Lasso with a re-weighted scheme is presented. W-Lasso achieves quite good source separation performance in reverberant environments, at the price of a tremendous computation time. Also, only semi-blind separation with known mixing filters was addressed in [71], which is poorly satisfying with regards of going towards separation into the wild. A similar hybrid TD/TF approach was recently followed in [76, 77]. The source signals were represented using either the modified discrete cosine transform (MDCT), which is real-valued and critically sampled, or the odd-frequency STFT (OFSTFT). A probabilistic LGM + NMF model was used for the source coefficients, which were inferred from TD mixture observations using a VEM algorithm. This led to very interesting results, at the price of huge computation. Here also, most experiments were conducted in a semi-blind setup with known mixing filters since the joint estimation of the filters' impulse responses remains difficult.

As for the approaches that work totally in the TF domain, let us first mention that the authors of [37] have shown that, in addition to modeling diffuse sources, their method is able to circumvent *to some extent* the discussed limitations of the MTF approximation. Other methods have investigated TF mixture

[5] Actually, the ratio between the coherent direct-path and the reverberation tail, the so-called direct-to-reverberant ratio (DRR) plays an important role examining the validity of MTF assumption.

models more accurate than (3). Fundamentally, the time-domain convolution can be exactly represented as a two-dimensional filtering in the TF domain [50]. This representation was used for linear system identification in [10] as an alternative to MTF, under the name of cross-band filters (CBFs). Using CBFs, a source image STFT coefficient is represented as a summation over frequency bins of multiple convolutions between the input source STFT coefficients and the TF-domain filter impulse response, along the frame axis. This exact representation becomes an approximation when we limit the number of bins either in the frequency-wise summation or in the frame-wise convolution. In particular, considering only the current frequency bin, i.e. a unique convolution along the STFT frame axis, is a reasonable practical approximation, referred to as the convolutive transfer function (CTF) model [130]. Using this CTF model, the mixture model (2) writes in the STFT domain:

$$\mathbf{x}(f,n) = \sum_{j=1}^{J} \sum_{n'=0}^{Q_a-1} \mathbf{a}_j(f,n') s_j(f,n-n') + \mathbf{b}(f,n). \tag{5}$$

Here, the i-th entry of $\mathbf{a}_j(f,n)$, denoted $a_{ij}(f,n)$, is not the DFT of $a_{ij}(t)$ (nor is it its STFT), but it is its CTF. The CTF contains several STFT-frame-wise filter taps and its expression is a bit more complicated than the DFT, though easily computable from $a_{ij}(t)$, see [10].

The full CBF representation was considered for solving the MASS problem in [11], in combination with a high-resolution NMF (HR-NMF) model of the source signal. A variational EM (VEM) algorithm was proposed to estimate the filters and infer the source signals. Unfortunately, due to the model complexity, this method was observed to perform well only in an oracle setup where both filters and source parameters are initialized from the individual source images. Therefore, in the current state-of-knowledge, the CBFs seem difficult to integrate into a realistic MASS framework and one has to resort to CTF-like approximations.

An MVDR beamformer, implemented in a generalized sidelobe canceller (GSC) structure, that utilizes the CTF model was proposed in [129]. It was shown to outperform the GSC beamformer which uses the MTF approximation [46]. It can be noted that, as opposed to the full-rank model, the CTF approximation allows for coherent processing and can therefore implement an almost perfect null towards a point interfering source. The ability of the FR model to suppress the interference signal is limited by the number of microphones and their constellation and cannot exceed I^2 [112] for fully diffuse signal.

Interestingly, the pioneering work [8] combined an STFT-domain convolutive model very similar to (5) with a Gaussian mixture model (GMM) of source signals. In this paper, the STFT-domain convolution was intuited from empirical observations and was referred to as "subband filters" (no theoretical justification nor references were provided). Because of the overall complexity of the model (especially the large number of GMM components that is necessary to accurately represent speech signals), the author resorted to a VEM algorithm for parameters estimation. In [56], an STFT-domain convolutive model also very

similar to (5) was used together with an HMM on source activity. However, the optimization method used to estimate the parameters and infer the source signal is quite complex.

A Lasso-type optimization applied to the MASS problem was considered in [79] within the CTF framework. More specifically, the ℓ_2-norm model fitting term of Lasso was defined at each frequency bin with the STFT-domain convolutive mixture (5) instead of the TD convolutive mixture (2) as done in [71]. In parallel, the ℓ_1-norm regularizer of Lasso was kept so as to exploit the sparsity of TF-domain audio signals. Because the number of filter frames Q_a in (5) is much lower than the length L_a of the TD filter impulse response in (2), the computation cost in [79] is drastically reduced compared to [71]. This was obtained at the price of quite moderate loss in separation performance, showing the good accuracy of the CTF approximation. However, as for [71], this was done only in a semi-blind setup with known filters. To address the use of CTF in a fully blind scenario, the mixture model (5) was plugged into a probabilistic framework with an LGM for source STFT coefficients in [80]. An exact EM algorithm was developed for joint estimation of the parameters (source parameters and CTF filter coefficients) and inference of the source STFT coefficients. The joint estimation of source STFT coefficients and CTF mixing filter coefficients was recently addressed as a pure optimization problem in [43].

Another attempt to address the problem of long filters is presented in [124]. In this work, the RTF is split into an early part which is coherently processed and a late part which is treated as an additive noise. This noise is reduced by a combination of an MVDR beamformer and a postfilter. In [125], a nested GSC scheme was proposed that treats the long RIRs jointly as a coherent phenomenon, using CTF modelling, and as a diffused sound field. Different blocks of the proposed scheme use the different RIR models.

In parallel with the above attempts to model long mixing filters, as briefly mentioned in the previous sections, several more or less recent studies have considered the modeling of the mixing filters/process as latent variables, possibly in a fully Bayesian framework, either to better account for uncertainty in filter estimation or to introduce prior knowledge on these filters (for example the approximate knowledge of source DOA or the specific structure of room acoustic impulse responses). Because of room limitation, we do not describe those works and only add [21, 75, 52] to the already cited references [109, 57, 38, 67].

It is clear from the above discussion that many models of the mixing filters are used in the literature. It is still in open question, which of the models is the most appropriate. Most probably, the answer to this question depends on the scenario.

3.4 Ad hoc microphone arrays

Classical microphone arrays usually consist of a condensed set of microphone mounted on a single device. Establishing wireless acoustic sensor networks (WASNs), comprising multiple cooperative devices (e.g., cellphone, tablet, hearing aid, smart watch) may increase the chances to find a subset of the micro-

phones that is close to a relevant sound source. Consequently, WASNs may demonstrate higher separation capabilities than a single-device solution. The wide-spread availability of devices equipped with multiple microphones makes the vision closer to reality.

The distributed and ad hoc nature of WASNs arises new challenges, e.g. transmission and processing constraints, synchronization between nodes and dynamic network topology. Addressing these new challenges is a prerequisite for fully exploiting the potential of WASNs.

Several families of distributed algorithms, that only require the transmission of a fused version of the signals received by each node was proposed in [93]. The distributed adaptive node-specific signal estimation (DANSE) family of algorithms consists of distributed version of SDW-MWF [15] and LCMV beamformers [16]. A distributed version of the GSC beamformer is presented in [96]. A randomized gossip implementation of the delay and sum beamformer is presented in [149], and a diffusion adaptation method for distributed MVDR beamformer in [105]. The problem of synchronizing the clock drifts in several nodes is addressed in e.g. [143, 114, 140, 24].

The full potential of ad hoc microphone arrays to separate source in the wild is yet to be explored.

4 Conclusions and Perspectives

In this review, we have presented several ways to make MASS and beamforming techniques go from laboratories to real-life scenarios. Laboratory studies are often based on a set of assumptions on the source signals and/or the mixture process that may not be totally realistic (e.g. static, point sources, and spatially stable microphone constellation). In this section, we will briefly explore a few families of devices that already work in real-life scenarios. We will conclude this section and the entire chapter by a perspective on the future of the research in the field.

In recent years, we have witnessed the penetration, in an accelerating pace, of smart audio devices to the consumer electronics market. These devices, designed to work in adverse conditions, include personal assistants embedded in smartphones, portable computers and most notably, smart loudspeakers, e.g. Amazon Echo ("Alexa"),[6] Microsoft Invoke (with "Cortana"),[7] Apple HomePod (with "Siri")[8] and Google Home [78].

Basically, these smart loudspeakers demonstrate that tremendous progress has already been made in middle-range devices, capable of executing automatic speech recognition (ASR) engines in noisy environments. Smart loudspeakers are equipped with several microphones (six for Apple Homepod, seven for Ama-

[6] https://www.slideshare.net/AmazonWebServices/designing-farfield-speech-processing-systems-with-intel-and-amazon-alexa-voice-service-alx305-reinvent-2017

[7] https://news.harman.com/releases/harman-reveals-the-harman-kardon-invokeTM-intelligent-speaker-with-cortana-from-microsoft

[8] https://www.apple.com/homepod/

zon Echo and Microsoft Invoke, and only two microphones for Google Home). Algorithmically, these devices consist of a denoising (mostly using a steered beamformer), dereverberation and echo cancellation stages. The devices usually employ localization (or DOA estimation) algorithms to provide the steering direction, as an important prerequisite to the application of the beamformer. The acquired localization information is also used for indicating the direction of the detected source with respect to the device. As smart loudspeakers acquire and enhance a speech signal in a noisy and reverberant environment, they provide a living example of into-the-wild beamforming. Yet, their performance may be still limited to home scenarios with a predominant and reasonably spatially stable speaker relatively close to the device (as opposed to the above-mentioned adverse scenarios with several active and moving sources with low DRR).

Hearing aids [32] are another example of successful application of MASS / beamforming technologies, aiming at speech quality and intelligibility improvement, as well as enhancing the spatial awareness of the hearing aid wearer (in binaural setting). Hearing devices impose severe real-time constraints on the applied algorithms (latency shorter than 10 ms). Moreover, robustness and reliability are of major importance to prevent potential hearing damage to the hearing impaired person. Binaural cue preservation can be obtained by calculating a common gain to both hearing devices [63] or by applying a beamformer that incorporates binaural information into the optimization criterion, e.g. MWF [97] or LCMV [54]. Beamforming-based binaural processing is usually regarded computationally more expensive than the common gain approach. An important issue in designing a binaural enhancement algorithm is to determine the source of interest. In many cases, the beamformer is steered towards the look direction of the hearing aid wearer.

Most cellular phones are nowadays equipped with multiple microphones (3-4) and they usually work in adverse conditions demonstrating reasonable performance. A few systems already employ microphone networks, e.g. smart home and smart cities.

The quest for realistic solutions, capable of processing a large amount of sound sources in real-life environments and in dynamical scenarios of various character, still continues. Many of the theoretical and practical questions are still open and there are performance gaps to be filled for many scenarios such as under-determined mixtures with many simultaneously active sources, multiple moving sources and moving sensors (e.g. robots, cellular phones), high-power and non-stationary noise (e.g. from heavy machinery and drilling noise in mines), and binaural hearing (for both hearing impaired people and robots imitating the Human auditory system [87]).

Recent years have witnessed a revolution in MASS techniques. Nowadays, *deep learning* solutions seem to be the new El Dorado for audio processing. Still, most studies deal with single-channel denoising / enhancement / separation algorithms [102, 22, 51, 55, 139, 86]. More recently, multichannel processing solutions that employ deep learning [104, 23], as well as robust ASR systems [60], have been proposed. Deep learning has also influenced the hearing aid industry [138] and the development of binaural algorithms [150]. An improved

localization strategy that utilizes active head movements and deep learning is proposed in [88]. Despite the impressive performance gains obtained by deep learning based speech processing approaches, the field is still in its infancy and major breakthroughs are expected in the foreseeable future.

As an outcome from this review, it is evident that a significant progress is still required for obtaining robust and reliable source separation in difficult real-life scenarios, especially under severe online constraints. We anticipate that future solutions will combine ideas from both the array processing / source separation and machine learning paradigms. As always, only such combined solutions, together with practical knowhow, are capable of advancing the already established solutions towards comprehensive audio source separation methods that work into-the-wild.

References

References

[1] R. Aichner, H. Buchner, S. Araki, and S. Makino. On-line time-domain blind source separation of nonstationary convolved signals. In *Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, Nara, Japan, 2003.

[2] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–371, 2012.

[3] 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 Transactions on Speech and Audio Processing*, 11(2):109–116, 2003.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino. Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Processing*, 87(8):1833–1847, 2007.

[5] S. Arberet, R. Gribonval, and F. Bimbot. A robust method to count and locate audio sources in a multichannel underdetermined mixture. *IEEE Transactions on Signal Processing*, 58(1):121–133, 2010.

[6] S. Arberet, A. Ozerov, N. Q. K. 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 *IEEE International Symposium on Signal Processing and Its Applications (ISSPA)*, Kuala Lumpur, Malaysia, 2010.

[7] S. Arberet, P. Vandergheynst, R. Carrillo, J-P. Thiran, and Y. Wiaux. Sparse reverberant audio source separation via reweighted analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1391–1402, 2013.

[8] H. Attias. New EM algorithms for source separation and deconvolution with a microphone array. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003.

[9] Y. Avargel and I. Cohen. On multiplicative transfer function approximation in the short-time Fourier transform domain. *IEEE Signal Processing Letters*, 14(5):337–340, 2007.

[10] Y. Avargel and I. Cohen. System identification in the short-time Fourier transform domain with crossband filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1305–1319, 2007.

[11] R. Badeau and M.D. Plumbley. Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(11):1670–1680, 2014.

[12] L. Benaroya, L.M. Donagh, F. Bimbot, and R. Gribonval. Non negative sparse representation for Wiener based source separation with a single sensor. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

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

[14] J. Benesty, S. Makino, and J. Chen, editors. *Speech Enhancement*. Springer, 2005.

[15] A. Bertrand and M. Moonen. Distributed adaptive node-specific signal estimation in fully connected sensor networks – part I: sequential node updating. *IEEE Transactions on Signal Processing*, 58:5277–5291, 2010.

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

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

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

[19] G. Bustamante, P. Danès, T. Forgue, A. Podlubne, and J. Manhes. An information based feedback control for audio-motor binaural localization. *Autonomous Robots*, 42(2):477–490, 2018.

[20] J.F. Cardoso. Blind signal separation: Statistical principles. *Proceedings of the IEEE*, 9(10):2009–2025, 1998.

[21] A.T. Cemgil, C. Févotte, and S. Godsill. Variational and stochastic inference for Bayesian source separation. *Digital Signal Processing*, 2007(17):891–913, 2007.

[22] S. E. Chazan, J. Goldberger, and S. Gannot. A hybrid approach for speech enhancement using MoG model and neural network phoneme classifier. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12), December 2016.

[23] S. E. Chazan, J. Goldberger, and S. Gannot. DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming. In *IEEE International Conference on Audio and Acoustic Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018.

[24] D. Cherkassky and S. Gannot. Blind synchronization in wireless acoustic sensor networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(3):651–661, March 2017.

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

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

[27] R. K. Cook, R.V. Waterhouse, R.D. Berendt, S. Edelman, and M.C. Thompson Jr. Measurement of correlation coefficients in reverberant sound fields. *The Journal of the Acoustical Society of America*, 27(6):1072–1077, 1955.

[28] H. Cox. Spatial correlation in arbitrary noise fields with application to ambient sea noise. *The Journal of the Acoustical Society of America*, 54(5):1289–1301, 1973.

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

[30] A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

[31] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein. Robust localization in reverberant rooms. In *Microphone Arrays*, pages 157–180. Springer, 2001.

[32] H. Dillon. *Hearing Aids*. Thieme, 2012.

[33] P. Divenyi, editor. *Speech separation by Humans and machines*. Springer Verlag, 2004.

[34] S. Doclo and M. Moonen. GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing*, 50(9):2230–2244, 2002.

[35] 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, pages 199–228. Springer, Berlin, 2005.

[36] L. Drude, A. Chinaev, D.H. Tran Vu, and R. Haeb-Umbach. Source counting in speech mixtures using a variational EM approach for complex Watson mixture models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[37] N. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1830–1840, 2010.

[38] N. Duong, E. Vincent, and R. Gribonval. Spatial location priors for Gaussian model based reverberant audio source separation. *EURASIP Journal on Advances in Signal Processing*, 2013(149), 2013.

[39] T.G. Dvorkind and S. Gannot. Time difference of arrival estimation of speech source in a noisy and reverberant environment. *Signal Processing*, 85(1):177–204, 2005.

[40] C. Evers, Y. Dorfan, S. Gannot, and P.A. Naylor. Source tracking using moving microphone arrays for robot audition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, 2017.

[41] C. Evers, A.H. Moore, and P.A. Naylor. Localization of moving microphone arrays from moving sound sources for robot audition. In *European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary, 2016.

[42] M.F. Fallon and S.J. Godsill. Acoustic source localization and tracking of a time-varying number of speakers. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1409–1415, 2012.

[43] F. Feng. *Séparation aveugle de sources: de l'instantané au convolutif*. Ph.D. thesis, Université Paris Sud, 2017.

[44] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[45] C. Févotte and J.-F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models. In *IEEE Workshop Applicat. Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NJ, 2005.

[46] S. Gannot, D. Burshtein, and E. Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626, 2001.

[47] S. Gannot and M. Moonen. On the application of the unscented Kalman filter to speech processing. In *IEEE Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 2003.

[48] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25(4):692–730, 2017.

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

[50] A. Gilloire and M. Vetterli. Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation. *IEEE Transactions on Signal Processing*, 40(8):1862–1875, 1992.

[51] E. Girgis, G. Roma, A. Simpson, and M. Plumbley. Combining mask estimates for single channel audio source separation using deep neural networks. *Conference of the International Speech Communication Association (INTERSPEECH)*, 2016.

[52] L. Girin and R. Badeau. On the use of latent mixing filters in audio source separation. In *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Grenoble, France, 2017.

[53] E. Habets and S. Gannot. Generating sensor signals in isotropic noise fields. *The Journal of the Acoustical Society of America*, 122:3464–3470, 2007.

[54] E. Hadad, S. Doclo, and S. Gannot. The binaural LCMV beamformer and its performance analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):543–558, 2016.

[55] J.R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016.

[56] T. Higuchi and H. Kameoka. Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014.

[57] T. Higuchi, N. Takamune, N. Tomohiko, and H. Kameoka. Underdetermined blind separation and tracking of moving sources based on DOA-HMM. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[58] T. Higuchi, H. Takeda, N. Tomohiko, and H. Kameoka. A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden Markov models. In *Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, 2014.

[59] K.E. Hild II, D. Erdogmus, and J.C. Principe. Blind source separation of time-varying, instantaneous mixtures using an on-line algorithm. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, Florida, 2002.

[60] T. Hori, Z. Chen, H. Erdogan, J.R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe. Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced NN/RNN backend. *Computer Speech & Language*, 46:401–418, 2017.

[61] A. Hyvärinen, J. Karhunen, and E. Oja, editors. *Independent Component Analysis*. Wiley and Sons, 2001.

[62] M. Z. Ikram and D. R. Morgan. A beamformer approach to permutation alignment for multichannel frequency-domain blind source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, Florida, 2002.

[63] A. H. Kamkar-Parsi and M. Bouchard. Instantaneous binaural target PSD estimation for hearing aid noise reduction in complex acoustic environments. *IEEE Transactions on Instrumentation and Measurments*, 60(4):1141–1154, 2011.

[64] B. Kleijn and F. Lim. Robust and low-complexity blind source separation for meeting rooms. In *Int. Conf. on Hands-free Speech Communication and Microphone Arrays*, San Francisco, CA, 2017.

[65] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud. A variational EM algorithm for the separation of moving sound sources. In *IEEE Workshop Applicat. Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NJ, 2015.

[66] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud. An inverse-Gamma source variance prior with factorized parameterization for audio source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016.

[67] 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 Transactions on Audio, Speech, and Language Processing*, 24(8):1408–1423, 2016.

[68] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud. An EM algorithm for joint source separation and diarization of multichannel convolutive speech mixtures. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, 2017.

[69] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud. Exploiting the intermittency of speech for joint separation and diarization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NJ, 2017.

[70] A. Koutras, E. Dermatas, and G. Kokkinakis. Blind speech separation of moving speakers in real reverberant environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000.

[71] M. Kowalski, E. Vincent, and R. Gribonval. Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1818–1829, 2010.

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

[73] Y. Laufer and S. Gannot. A Bayesian hierarchical model for speech enhancement. In *IEEE International Conference on Audio and Acoustic Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018.

[74] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[75] S. Leglaive, R. Badeau, and G. Richard. Multichannel audio source separation with probabilistic reverberation priors. *IEEE Transactions on Audio, Speech and Language Processing*, 24(12), 2016.

[76] S. Leglaive, R. Badeau, and G. Richard. Multichannel audio source separation: Variational inference of time-frequency sources from time-domain observations. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New-Orleans, Louisiana, 2017.

[77] S. Leglaive, R. Badeau, and G. Richard. Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2017.

[78] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K. C. Sim, R. J. Weiss, K. W. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon. Acoustic modeling for Google Home. In *Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden, 2017.

[79] X. Li, L. Girin, and R. Horaud. Audio source separation based on convolutive transfer function and frequency-domain Lasso optimization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, 2017.

[80] X. Li, L. Girin, and R. Horaud. An EM algorithm for audio source separation based on the convolutive transfer function. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2017.

[81] X. Li, L. Girin, R. Horaud, and S. Gannot. Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1007–2012, 2017.

[82] A. Liutkus, B. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155–3167, 2011.

[83] B. Loesch and B. Yang. Online blind source separation based on time-frequency sparseness. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

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

[85] H. Löllmann, A. Moore, P. Naylor, B. Rafaely, R. Horaud, A. Mazel, and W. Kellermann. Microphone array signal processing for robot audition. In *IEEE Int. Conf. on Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, 2017.

[86] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani. Deep clustering and conventional networks for music separation: Stronger together. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Lousiana, 2017.

[87] R. F. Lyon. *Human and Machine Hearing: Extracting Meaning from Sound*. Cambridge University Press, 2017.

[88] N. Ma, T. May, and G. J. Brown. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2444–2453, 2017.

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

[90] S. Malik, J. Benesty, and J. Chen. A Bayesian framework for blind adaptive beamforming. *IEEE Transactions on Signal Processing*, 62(9):2370–2384, 2014.

[91] M. Mandel, R. J. Weiss, and D.P.W. Ellis. Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):382–394, 2010.

[92] S. Markovich, S. Gannot, and I. Cohen. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1071–1086, 2009.

[93] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot. Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks. *Signal Processing*, 107:4–20, 2015.

[94] S. Markovich-Golan, S Gannot, and I. Cohen. Subspace tracking of multiple sources and its application to speakers extraction. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, 2010.

[95] S. Markovich-Golan, S. Gannot, and I. Cohen. Low-complexity addition or removal of sensors/constraints in LCMV beamformers. *IEEE Transactions on Signal Processing*, 60(3):1205–1214, 2012.

[96] S. Markovich-Golan, S. Gannot, and I. Cohen. Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):343–356, 2013.

[97] 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 Transactions on Audio, Speech, and Language Processing*, 23(12):2384–2397, 2015.

[98] N. Mitianoudis and M.E. Davies. Audio source separation of convolutive mixtures. *IEEE Transactions on Speech and Audio Processing*, 11(5):489–497, 2003.

[99] R. Mukai, H. Sawada, S. Araki, and S. Makino. Robust real-time blind source separation for moving speakers in a room. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[100] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino. Sound source separation of moving speakers for robot audition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[101] K. Nakadai, T. Takahashi, H. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and implementation of robot audition system 'HARK'-Open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[102] A. Narayanan and D. Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[103] F. Nesta, P. Svaizer, and M. Omologo. Convolutive BSS of short mixtures by ICA recursively regularized across frequencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):624–639, 2011.

[104] A. Nugraha, A. Liutkus, and E. Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664, 2016.

[105] M. O'Connor and W. B. Kleijn. Diffusion-based distributed MVDR beamformer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[106] P. O'Grady, B. A. Pearlmutter, and S. Rickard. Survey of sparse and non-sparse methods in source separation. *Int. Journal of Imaging Systems and Technology*, 15(1):18–33, 2005.

[107] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010.

[108] A. Ozerov, C. Févotte, and M. Charbit. Factorial scaled hidden markov model for polyphonic audio representation and source separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009.

[109] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, 2012.

[110] L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3):320–327, 2000.

[111] L.C. Parra and C.V. Alvino. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352–362, 2002.

[112] A.T. Parsons. Maximum directivity proof for three-dimensional arrays. *Journal of the Acoustical Society of America*, 82(1):179–182, 1987.

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

[114] P. Pertilä, M. S. Hämäläinen, and M. Mieskolainen. Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2393–2402, 2013.

[115] M. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M.E. Davies. Sparse representations in audio and music: From coding to source separation. *Proceedings of the IEEE*, 98(6):995–1005, 2010.

[116] R. E. Prieto and P. Jinachitra. Blind source separation for time-variant mixing systems using piecewise linear approximations. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PN, 2005.

[117] N. Roman and D. Wang. Binaural tracking of multiple moving sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):728–739, 2008.

[118] N. Roman, D. Wang, and G.J. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114(4):2236–2252, 2003.

[119] H. Sawada, S. Araki, R. Mukai, and S. Makino. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1592–1604, 2007.

[120] 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 Transactions on Speech and Audio Processing*, 12(5):530–538, 2004.

[121] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin. Variational bayesian inference for multichannel dereverberation and noise reduction. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 22(8):1320–1335, 2014.

[122] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[123] O. Schwartz and S. Gannot. Speaker tracking using recursive EM algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):392–402, 2014.

[124] O. Schwartz, S. Gannot, and E. Habets. Multi-microphone speech dereverberation and noise reduction using relative early transfer functions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):240–251, 2015.

[125] O. Schwartz, S. Gannot, and E.P. Habets. Nested generalized sidelobe canceller for joint dereverberation and noise reduction. In *IEEE International Conference on Audio and Acoustic Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

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

[127] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1):21–34, 1998.

[128] 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. Convention of the Audio Engineering Society (AES)*, Budapest, Hungary, 2012.

[129] R. Talmon, I. Cohen, and S. Gannot. Convolutive transfer function generalized sidelobe canceler. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1420–1434, 2009.

[130] R. Talmon, I. Cohen, and S. Gannot. Relative transfer function identification using convolutive transfer function approximation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):546–555, 2009.

[131] O. Thiergart, M. Taseska, and E. Habets. An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[132] O. Thiergart, M. Taseska, and E. Habets. An informed parametric spatial filter based on instantaneous direction-of-arrival estimates. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):2182–2196, 2014.

[133] J.-M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3):216–228, 2007.

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

[135] D. Vijayasenan, F. Valente, and H. Bourlard. Multistream speaker diarization of meetings recordings beyond MFCC and TDOA features. *Springer handbook on speech processing and speech communication*, 54(1), 2012.

[136] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot. From blind to guided audio source separation: How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine*, 31(3):107–115, 2014.

[137] E. Vincent, M.G. Jafari, S.A. Abdallah, M.D. Plumbley, and M.E. Davies. Probabilistic modeling paradigms for audio source separation. *Machine Audition: Principles, Algorithms and Systems*, pages 162–185, 2010.

[138] D. Wang. Deep learning reinvents the hearing aid. *IEEE Spectrum*, 54(3):32–37, 2017.

[139] D. Wang and J. Chen. Supervised speech separation based on deep learning: an overview. *arXiv preprint arXiv:1708.07524*, 2017.

[140] L. Wang and S. Doclo. Correlation maximization-based sampling rate offset estimation for distributed microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):571–582, 2016.

[141] L. Wang, T.-K. Hon, J.D. Reiss, and A. Cavallaro. An iterative approach to source counting and localization using two distant microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(6):1079–1093, 2016.

[142] E. Warsitz and R. Haeb-Umbach. Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1529–1539, 2007.

[143] 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 *IEEE Int. Symposium on Multimedia Software Engineering*, Miami, FL, 2004.

[144] E. Weinstein, A.V. Oppenheim, M. Feder, and J.R. Buck. Iterative and sequential algorithms for multisensor signal enhancement. *IEEE Transactions on Signal Processing*, 42(4):846–859, 1994.

[145] B. Widrow, J.R. Glover, J.M. McCool, J. Kaunitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, J.R.E. Dong, and R.C. Goodlin. Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63(12):1692–1716, 1975.

[146] 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 Journal on Applied Signal Processing*, 2007(1):81–81, 2007.

[147] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

[148] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. Okuno. Blind separation and dereverberation of speech mixtures by joint optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):69–84, 2011.

[149] Y. Zeng and R.C. Hendriks. Distributed delay and sum beamformer for speech enhancement via randomized gossip. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):260–273, 2014.

[150] X. Zhang and D. Wang. Deep learning based binaural speech separation in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):1075–1084, 2017.