Gaël Richard, Paris Smaragdis, Sharon Gannot, Patrick A. Naylor, Shoji Makino, Walter Kellermann, and Akihiko Sugiyama

Audio Signal Processing in the 21st Century

The important outcomes of the past 25 years

©SHUTTERSTOCK.COM/TRIFF

Audio signal processing has passed many landmarks in its development as a research topic. Many are well known, such as the development of the phonograph in the second half of the 19th century and technology associated with digital telephony that burgeoned in the late 20th century and is still a hot topic in multiple guises. Interestingly, the development of audio technology has been fueled not only by advancements in the capabilities of technology but also by high consumer expectations and customer engagement. From surround sound movie theaters to the latest in-ear devices, people love sound and soon build new audio technology into their daily lives as an essential and expected feature.

Some of the major outcomes of the research in audio and acoustic signal processing (AASP) prior to 1997 were summarized in a landmark paper published on the occasion of the 50th anniversary of the IEEE Signal Processing Society (SPS) [1]. At that time, the vast majority of the work was driven by the objective to build models that capture the essential characteristics of the analyzed audio signal and to represent it with a limited set of parameters and components. The field has now evolved beyond the essential characteristics explored in the past. For instance, a wide variety of speech/audio signal models have since been proposed and, in particular, around signal decomposition/factorization models and sparse signal representations. Nevertheless, the entire research domain covered by the IEEE Technical Committee (TC) on AASP is witnessing a paradigm shift toward data-driven methods based on machine learning and, especially, deep learning.

In many applications, such data-driven models obtain state-of-the-art results if appropriate data are available to train the models. This has accompanied sustained efforts to gather highly valuable and public data collections (and, in particular, annotated data), which are, in fact, essential for data-driven algorithms. Concurrently, to promote reproducible research and identify state-of-the-art methods, a number of challenges have arisen, for instance, in acoustic characterization of environments (ACE), reverberant speech processing (REVERB), acoustic source localization and tracking, source separation

Digital Object Identifier 10.1109/MSP.2023.3276171
Date of current version: 14 July 2023

(SiSEC), acoustic echo cancellation (AEC), deep noise suppression dedicated to single-microphone noise reduction, and the detection and classification of acoustic scenes and events (DCASE), which has been the subject of a yearly event since 2016 (SPS data challenges: https://signalprocessingsociety.org/publications-resources/data-challenges; REVERB Challenge: http://reverb2014.dereverberation.com; SiSEC challenge: https://sisec.inria.fr; and DCASE challenges: https://dcase.community/challenge2022).

Without aiming for exhaustiveness, the article provides a view of the important outcomes of the field in the past 25 years, also illustrating the emergence of purely data-driven models. In particular, the article covers the research addressed in signal models and representations; the modeling, analysis, and synthesis of acoustic environments and acoustic scenes; signal enhancement and separation; music information retrieval (MIR); and Detection and Classification of Acoustic Scenes and Events (DCASE).

The overall structure of the article is as follows. We discuss, in the "Advances and Highlights (Evolution and Breakthrough)" section, the main axes of progress and highlights of the domain underlining the evolution and breakthroughs of the field. We then focus, in the "Emerging Topics" section, on the new topics that have mostly emerged in the past 25 years, before suggesting some conclusions and perspectives.

Advances and highlights (evolution and breakthrough)

Building upon the achievements prior to 1997, already discussed in [1], we summarize, in this section, the key advances and highlights of recent years.

Modeling and representation

We first discuss the developments in audio coding and signal modeling, with a focus on multichannel audio channel coding. We then describe some of the important work pursued in modeling, analysis, and synthesis of acoustic environments, with specific highlights on room impulse response (RIR) analysis and synthesis.

Coding and signal modeling

Audio coding is a long-standing topic in the field and has led to several international standards. [The International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) audio coding standards in the following are accessible at https://www.iso.org/standards.html by providing the search window with the numbers and years in the parentheses.]

The field had its golden age in the 1990s, with the first international standard of audio coding, MPEG-1 Audio (11172-3:1993), and its extension to multichannel signals of up to five channels, MPEG-2 Audio (13818-3:1995). MPEG-2 Audio was developed for multichannel and multilingual applications, such as digital radio broadcasting in Europe, with backward compatibility with MPEG-1.

However, without the backward compatibility constraint, much higher subjective quality was successfully achieved with MPEG-2 Advanced Audio Coding (AAC) (13818-7:1997). It is still the foundation of today's audio coding algorithms and is employed in terrestrial TV broadcasting in Japan and Latin America. From a viewpoint of applications, MPEG-4 AAC (14496-3:2009) and MPEG-4 High-Efficiency (AAC HE-ACC) (14496-3:2009/Amd 7:2018) achieve sufficient audio quality at 64 kbit/s and 32 kbit/s, respectively, for mobile applications and are most widely used today.

One of the major improvements is brought by bandwidth extension (BWE), also known as *subband replication (SBR)*, which encodes only the low-frequency subband plus high-frequency power envelope information, thereby reducing the bitrate with inaudible quality degradation. The decoder copies the low-frequency spectrum to the high-frequency band and adjusts the envelope by the transmitted envelope information to reconstruct the full-band audio (see Figure 1). MPEG-4 AAC and HE-AAC are used in various consumer products, such as PCs, tablet PCs, mobile phones, and car navigation systems, to name a few.

The history of MPEG-1 Audio through MPEG-4 HE-AAC was to remove redundancy of the input audio in the frequency domain (transform coding), time domain (prediction), and spatial domain (multichannel coding). The next stage of MPEG Audio, MPEG Surround (MPS) (23003-1:2007), exploits further redundancy in the spatial domain, based on binaural cue coding [2]. A multichannel audio signal is decomposed into a monaural signal and additional spatial information in the form of the interaural level difference (ILD) and interaural time difference (ITD) in multiple time-frequency tiles (segments). The monaural data are encoded by MPEG-4 AAC, with a little side information representing the ILD and ITD. MPS achieves comparable quality to MPEG-4 AAC at one-third of the MPEG-4 AAC bitrate. The absolute subjective quality is transparent to the

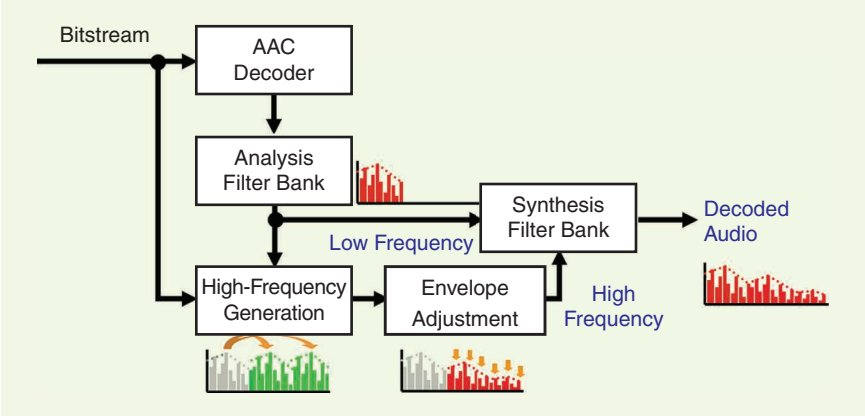

FIGURE 1. The BWE principle.

source signal, which is suitable for content delivery between geographically distributed studios and broadcasting stations. MPEG Spatial Audio Object Coding (SAOC) (23003-2:2010) removes the redundancy of the input audio, based on the composition of each audio object. The input audio signal consists of multiple audio objects, which are independent audio sources, such as individual musical instruments. Each audio object is expressed in multiple frequency tiles by object-level differences (OLDs) and interobject cross coherences (IOCs). The OLD is the relative energy to the energy of the downmix signal that is a combination of the audio objects. The IOC is the cross correlation to the downmix signal. The downmix signal of multiple objects is encoded by MPEG-4 AAC, whereas the OLD and IOC of each object are encoded as side information. The decoder recovers each object from the downmix signal, OLD, and IOC. A direct link to MPEG SAOC can also be made with the line of work developed simultaneously on (coding-based) informed source separation [3].

Until MPEG SAOC, speech-dominant audio signals and more general audio signals had been encoded with different algorithms. MPEG Unified Speech and Audio Coding (USAC) (14496-3:2009/Amd 3:2012) is the first audio coding framework that automatically switches between the speech-oriented algorithm and the audio-oriented algorithm, based on the input signal analysis result in multiple time-frequency tiles. The most recent member of the MPEG Audio family is MPEG-H (23008-3 2019), which is generic coding, including 3D audio (higher-order ambisonics or HoA).

The most successful application of audio coding is portable audio players, represented by Apple's iPod. The first prototype was the Silicon Audio, developed in 1994, which was a precursor of the iPod first put in the market in 2001. Audio players were later extended to include video data processing. The iPhone, released in 2007, was the first in the world and was combined with a large display to make a tablet PC or with a tiny display to make a smart watch. A history of these handy personal terminals can be found in [4]. Nevertheless, despite their immense success, audio players are now gradually being replaced by music streaming.

Acoustic environments modeling, analysis, and synthesis

Modeling and analysis of acoustic impulse responses

Sound propagation in acoustic enclosures is characterized by multiple reflections and the addition of noise, both associated with the acoustic environment. When an acoustic signal propagates in an echoic environment, it is reflected by the room facets and objects in the enclosure, resulting in the reverberation phenomenon. The acoustic impulse responses (AIRs) that relate sound sources and microphones are usually a few hundred milliseconds in duration, corresponding to a few thousand taps in discrete-time filtering at typical sampling rates. The decay rate of acoustic energy in an acoustic environment is measured by the reverberation time, T_{60}, the time it takes for the exponentially decaying power profile of the reverberation tail to decay by 60 dB from its initial value. Typical offices have a T_{60} around 300–400 ms, and larger rooms can approach 1 s, depending on the volume, shape, and materials. The perceived reverberation also depends on the ratio between the direct path (including the early reflections) and the power of the tail, denoted as the direct-to-reverberant ratio (DRR). In the same environment, distant sources will exhibit a lower DRR and be perceived as more reverberant.

Reverberation can degrade the quality of a speech signal and, in severe cases, particularly in noise, its intelligibility. The word error rate (WER) of automatic speech recognition (ASR) systems is usually severely impacted by high reverberation levels, especially for a low DRR.

An AIR encompasses the entire reflection pattern, consisting of the direct path, the early reflections (consisting of several distinguishable arrivals), and the late reflection tail, with an exponentially decaying power profile. The latter part is the main cause of the reverberation phenomenon.

Sound propagation in acoustic enclosures is characterized by multiple reflections and the addition of noise, both associated with the acoustic environment.

When an acoustic environment is a room, its AIR is referred to as an *RIR*. Room acoustics, even in mild reverberation conditions, should be taken into account when designing acoustic signal processing algorithms, and failing to do so may severely degrade their performance. Modeling and accurately analyzing the properties of the RIR is therefore of crucial importance.

Room simulators, RIR datasets, and sound field generators

Acoustic signal processing algorithms should be evaluated under reverberant conditions. This can be achieved either by using recorded RIRs or using room simulators. The outcome of such simulators may be less accurate, but using them allows researchers in the field to generate a vast number of examples. This has recently become extremely important with the emergence of machine learning algorithms that require a large volume and diversity of training data. The field has evolved from the pioneering work in acoustics by Schröder (frequency-domain modeling), Polack (time-domain modeling), and Allen and Berkely (the image method) [5]. Based on these models (especially the image method), many RIR generators were developed: the RIR generator (https://github.com/ehabets/RIR-Generator), PyRoomAcoustics (https://pyroomacoustics.readthedocs.io/en/pypi-release/pyroomacoustics.room.html), and gpuRIR (https://github.com/DavidDiazGuerra/gpuRIR). Using these generators, one can evaluate the performance of audio processing algorithms and also train data-driven methods. Recent advances improve the RIR generation using data-driven methods, usually generative adversarial networks (GANs).

Databases of real-world RIRs are also available, facilitating reliable evaluation of algorithms (https://www.dreams-itn.

eu/index.php/dissemination/science-blogs/24-rir-databases, https://github.com/RoyJames/room-impulse-responses, and https://asap.ite.tul.cz/downloads/mirage). In parallel, noise field generators were also proposed, including isotropic noise (https://github.com/ehabets/INF-Generator) and wind noise (https://github.com/ehabets/Wind-Generator).

Inference of room characteristics

The parameters characterizing the acoustic properties of an enclosure can be inferred from the AIR and the reverberant sound itself. These parameters can be used in the development of audio processing algorithms and also in rendering acoustic scenes. The reverberation time, T_{60}, and DRR were mentioned in the preceding. The coherent-to-diffuse power ratio (CDR) is another attribute of the sound field that determines the impact of reverberation and depends on the source–microphone distance and reverberation time. If the direct path and early reflections are dominant, the sound is perceived as more coherent, less diffuse, and less reverberant. The ACE Challenge (http://www.ee.ic.ac.uk/naylor/ACEweb) was dedicated to developing and benchmarking estimation procedures for the preceding room acoustic parameters. A recent database of RIRs with annotated reflections ("dEchorate") can be used to advance research further in this direction (https://zenodo.org/record/4626590#.Y1cMoOxByAQ).

Generation of artificial reverberation

Another thriving research direction is the generation of artificial reverberation, with the most popular method being feedback delay networks [6]. Traditionally (from the pioneering work of Schröder), these algorithms have been widely used in music production and now find applications in new fields, such as game audio, including virtual and augmented reality.

A different angle of research would rather consider geometric approaches, which rely on physics-based models. The image method remains intractable for modeling late reverberation, especially that of large rooms. The radiance transfer method (RTM) was introduced to overcome this limitation, as it can model the diffuse reflections and sound energy decay of the late reverberation [7]. Although complex, it was later shown that the RTM can be linked to feedback delay networks to build efficient geometry-based reverberators [8].

Analysis of acoustic scenes

Here, we explore the field of acoustic scene analysis, using microphone arrays that are either arranged in structured constellations (e.g., spherical and circular) or arbitrarily distributed in the acoustic enclosure. We discuss the localization of sound sources and basic concepts of data-independent spatial filtering. We further discuss wave domain representations using the cylindrical or spherical harmonics domain [9]. While originating from sound field rendering and microphone array beamforming, these representations are now frequently used for, e.g., source localization, echo cancellation, active noise control (ANC), and blind source separation (BSS), which are discussed in the following.

Acoustic sensor networks

Recent technological advances in the design of miniature and low-power devices enable the deployment of so-called wireless acoustic sensor networks (WASNs). A WASN consists of multiple (often battery-powered) microphone nodes, each of which is equipped with one or more microphones, a signal processing unit, and a wireless communication module. The large spatial distribution of such microphone constellations yields a large amount of spatial information and consequently increases the probability that a subset of the microphones (node) is close to a relevant sound source. Many daily life devices are now equipped with multiple microphones and considerable audio processing capabilities. These technological advancements significantly pushed the research forward. WASNs find applications in hearing devices, speech communication systems, acoustic monitoring, ambient intelligence, and more.

However, new challenges arise in these new ad hoc architectures. Typically, for a spatially extended network, the utility of sensors for a given task should be assessed, and for coherent signal processing of multiple sensor nodes, the signals must be synchronized. In particular, when data centralization is not possible, due either to the lack of a dedicated central processing device or to overly demanding transmission/processing requirements, one must rely on distributed processing, where nodes share only compressed/fused microphone signals with one another. The according modifications for the various algorithms, e.g., for beamforming, will be discussed along with their nondistributed versions in the following. First steps have also been taken to consider a moving robot as part of an acoustic sensor network.

Localization and tracking

Speaker localization algorithms, mainly time difference of arrival (TDoA) and direction of arrival (DoA) estimation, emerged in the 1970s, with solutions based on the normalized cross correlation between the signals received by a pair of microphones, the so-called generalized cross correlation, and were later extended to multimicrophone solutions, most notably the steered response power phase transform [10], which steers a beam toward all candidate directions. Especially for simultaneously localizing multiple sources, generic frequency estimation and direction-finding algorithms (such as MUSIC and ESPRIT) were also adapted to acoustic applications, most prominently to the cylindrical and spherical harmonics domain. While TDoA and DoA estimation dominate localization efforts, efficient range estimation based on sound field characteristics, e.g., the CDR, has been demonstrated and applied for position estimation in WASNs [11].

In later years, there were many attempts to incorporate statistical methods that can also facilitate the tracking of sources in dynamic scenarios, including Bayesian methods, e.g., nonlinear extensions of the Kalman filter, particle filters, and probability hypothesis density filters, and non-Bayesian methods, e.g., recursive expectation maximization (EM).

Acoustic reflections may degrade the performance of localization and tracking algorithms, especially in highly

reverberant environments and when multiple speakers are concurrently active. There are two paradigms in the literature to mitigate the effects of reverberation on localization accuracy. The first focuses on extracting the direct path of the sound propagation from the source to the microphones while trying to minimize the effects of the long AIR. Under the second paradigm, more general features are extracted from the microphone signals. These features characterize sound propagation. Then, a mapping from these high-dimensional features to the source location is learned. Manifold learning-based methods adopt this paradigm (see the 2019 European Signal Processing Conference tutorial at https://sharongannot.group/wp-content/uploads/2021/06/Speaker-Localization-on-Manifolds.pdf). This is part of the trend toward data-driven methods, specifically deep neural network (DNN)-based algorithms, that infer the source location from a feature vector [12]. A recent survey [13] explores many of these methods.

Under the same paradigm, simultaneous localization and mapping (SLAM) can be used in the acoustic domain (acoustic SLAM) to enable devices equipped with microphones, such as robots, to move within their environment to explore, adapt to, and interact with sound sources of interest [14].

Spatial filtering

Essentially all multichannel algorithms, implicitly or explicitly, use the spatial diversity of the sensor arrangement for spatially selective signal processing. Referring to later sections for the treatment of other spatial filtering methods, such as data-dependent beamforming and multichannel source separation and signal extraction, here, we limit the consideration to data-independent linear spatial filtering, which was portrayed as an active area of research in [1]. Since then, notable advances in this area include the exploitation of the spherical harmonics domain [9], [15] as well as differential microphone arrays [16], [17], due to their high directivity. These also included the introduction of polynomial beamforming for efficient and flexible beamsteering; the use of powerful optimization algorithms for noniterative designs of beamformers that meet robustness constraints, e.g., on white noise gain; and the incorporation of object-related transfer functions, e.g., head-related transfer functions (HRTFs), into the beamformer design. While these data-independent techniques were conceived for microphone array signal processing, they can also be used for sound reproduction by loudspeaker arrays. For the latter, more reproduction-specific techniques are discussed in the following.

Synthesis of acoustic scenes

Listener-centric binaural rendering

Binaural rendering usually refers to the process of spatial sound reproduction with headphones. One popular approach is based on the use of HRTF filters. Such filters contain all the cues that allow a listener to localize a sound source (and, in particular, spectral cues and interaural differences in time and intensity) [18]. The binaural signals are then obtained, for each ear, by filtering the input monophonic signal by the HRTF corresponding to a given position in space. The rendering for reverberant environments is more complex since it should superimpose different HRTFs for each direction of the early reflections. This approach is, however, facing major challenges: the difficulty to acquire large databases of HRTFs, the difficulty of obtaining generic and nonindividualized HRTFs, and the necessity to limit the computation complexity for high-quality rendering. These challenges have fueled extensive research in several complementary directions: 1) obtaining more generic HRTFs, 2) obtaining means to adapt generic HRTFs to individuals (for instance, by averaging sets of HRTFs, using anthropometric measurements, and resorting to physical models), and 3) selecting an appropriate set of HRTFs from a large database by, e.g., subjective tests [19].

Sound field rendering

Beyond the universal numerical methods based on finite elements and finite differences, the signal processing of sound fields started to take advantage of wave domain representations, especially using the cylindrical or spherical harmonics domain [9], and has now been applied to address many key challenges in sound field rendering.

An important class of sound rendering techniques relies on a specific setting of distributed loudspeakers surrounding the listening area. Specific formats were developed based on stereophonic principles for a variety of configurations: six channels, including an additional one for low frequencies (5.1); eight channels (7.1); 12 channels (10.2); and 24 channels (22.2). These formats are associated with directional sound field encoding, which imposes strict constraints on the loudspeaker positions. Also, in practice, the spatial illusion is correct only in a rather small area around the center of the room (called the *sweet spot*). Outside this sweet spot, the sound is perceived as coming from the closest loudspeaker. The approaches based on sound field reproduction, such as ambisonics, originally proposed by Gerson in 1973, and wave field synthesis, introduced in the 1980s by Berkhout, and, in a more general representation, the spatial frequency domain [20] solve some of these constraints by taking into account the actual position of the speakers and creating virtual speakers for each required direction. In practice, these approaches can rely on object-based coding and have a much wider sweet spot. Since their introduction, these methods have received much attention and led to many extensions for sound field reproduction with parametric and nonparametric methods, with potentially small-size microphone arrays for the recording to arbitrary loudspeaker layouts [21]. Once sound field rendering also accounts for the acoustic environment, room equalization techniques become necessary, which have been studied in [22].

Acoustic signal enhancement

In this section, we explore both single- and multimicrophone approaches for acoustic signal enhancement, addressing multiple sources of interference, namely, echo, feedback,

reverberation, noise, and competing signals. A generic view of an acoustic signal processing architecture together with sound field synthesis, which was discussed in the preceding, is depicted in Figure 2.

Echo cancellation

Echo cancellation emerged in the 1960s but has seen radical progress in the past 50 years. Many of the advances in the field of AEC were explored at the SPS 50th anniversary [1], including recursive least squares, affine projection, subband and frequency-domain adaptive filters, and double-talk detectors. AECs became the enabling technology of hands-free telecommunication systems, especially modern video conference systems.

Several important challenges were then tackled to take into account the nonlinearities of the reproduction system [23], [24], the latter also harnessing DNNs to improve performance. A global approach for combining (residual) echo cancellation, dereverberation, and noise reduction, usually by applying a postfiltering stage, was also a topic of extensive research. The classical spectral postfiltering may be substituted with modern structures, such as DNNs, to further improve performance. In multimicrophone settings with additive noise present, it is important to design the AECs and beamforming stages such that their cross interference is minimized. Step size control continued to develop from double-talk detection [25] to Kalman filter-based and, more recently, Kalman filter with deep learning-based step size optimization. Stereophonic AEC, as discussed in Sondhi's seminal work, was extended to the multichannel case [26] and multiple-input, multiple-output AEC in the wave domain.

Comprehensive surveys of the AEC field, its achievements, and remaining challenges can be found in [26] and [27]. The International Workshop on Acoustic Echo and Noise Control (https://www.iwaenc.org), begun in 1989 and held at two-years interval, was originally dedicated to AEC, but its scope was rapidly extended to other audio signal processing domains, and the name was accordingly changed to the International Workshop on Acoustic Signal Enhancement.

Acoustic feedback and ANC

Acoustic feedback occurs when a microphone signal is played back by a loudspeaker (e.g., in public announcement systems and hearing aids). This creates a closed loop that limits the amount of amplification that can be applied in the loop before the system becomes unstable and produces the howling effect [28]. This problem is well known to hearing aid wearers, who report it as one of the main drawbacks, especially for those requiring high gain due to moderate to severe hearing impairment. In the first step, a good "closed" fitting of a hearing aid can usually provide for a stable increase in useful gain. To go beyond this, adaptive processing was introduced in the 1990s to cancel the feedback components, and this approach has been advancing in recent years through the use of better models of the feedback path and better methods to control feedback-canceling algorithms. Usable gains have risen by as much as 10 dB in some cases, providing corresponding benefits to the hearing impaired.

ANC systems are based on microphones that capture the sound outside a volume and render "antisound" to create a quiet zone. Research in the field was boosted by commercial products, e.g., noise-canceling headphones and aircraft and automotive applications. Aside from just suppressing noise in a given zone, multizone rendering became a topic of significant, both theoretical and practical, interest [29]: here, in each zone, only one of multiple simultaneously active sources should be audible, i.e., forming a "bright" zone, whereas all others should be suppressed, i.e., forming a "dark" zone each. This technology finds applications in entertainment, business, and health applications. For example, the sound from multiple TVs in the same hospital room may be zoned separately to each patient's bed. Also, the sound level and rendering strategy of a movie may be zoned differently to different seats in the listening room, creating a "bright zone" and a "dark zone." Different languages for the dialogue may also be rendered in specific zones.

Note that as soon as the reference information on the undesired sound in a certain zone does not need to be acquired by microphones but can be estimated from an observable sound source and modeled and measured sound propagation path characteristics (e.g., impulse responses), the creation of dark and bright zones reduces to a spatial filtering task.

Dereverberation

Related to the objective of AEC, the topic of dereverberation has received growing attention due to the clear need to remove reverberation from audio signals, particularly in speech processing tasks. Dereverberation, as opposed to AEC, is a blind estimation problem, as no reference signal for the anechoic signal is available. While only a few dereverberation algorithms were available in the late 1990s, dereverberation has become a flourishing field of research and reached some level of maturity, as reflected by a dedicated and highly cited book summarizing a decade of intensive activity [30] and, later, by the community-wide REVERB Challenge. Both single- and

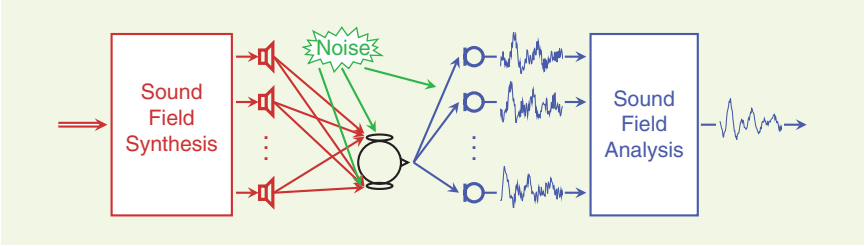

FIGURE 2. A typical multichannel sound system. On the analysis side, a spatially and/or spectrally selective acquisition is applied, including noise reduction, speaker separation (using either beamforming or independent component analysis), and dereverberation. Echo signals are also removed, and sources may be localized. On the synthesis side, a spatially selective rendering is applied, and noise can be actively canceled.

multimicrophone dereverberation algorithms have been proposed and evaluated. Statistical modeling of the decaying tail of the RIR has been used to derive spectral methods for single-microphone dereverberation [31].

In the multichannel case, dereverberation can be treated as a blind equalization problem. Hence, either the RIR coefficients or, alternatively, the inverse of a matrix of impulse responses should be estimated. Estimation procedures for the multichannel equalization system include subspace methods, i.e., extracting the RIRs from the null subspace of the spatial correlation matrix of the received microphone signals and least-sqaures methods for (partially) equalizing the multichannel RIRs and, consequently, the reverberation effects. The anechoic signal and (time-varying) RIRs can be also jointly estimated by applying a (recursive) EM algorithm in parallel to Kalman filtering.

The weighted prediction error (WPE) method [32] realized blind dereverberation of time-varying colored audio sources, such as speech, based on multichannel linear prediction (MCLP). To enable MCLP to handle such a source, the WPE introduced two necessary extensions into it: a nonstationary Gaussian source model and a delayed prediction that protects inherent source correlation from being whitened by MCLP. The WPE established a new effective MCLP algorithm called *variance-normalized delayed linear prediction*. Several extensions to this method, including joint BSS and dereverberation and the incorporation of DNNs, were also proposed.

In recent years, several successful data-driven methods based on DNNs were proposed [33]. We believe that this research direction will continue, exploring aspects including the noisy and time-varying nature of real-world scenarios, probably combining model-based and data-driven paradigms.

There is a growing interest in developing "thin" models that can be deployed in edge devices, such as cellular phones, and even simpler devices that are used as nodes in WASNs.

Noise suppression

Noise reduction algorithms gained momentum in the late 1970s, with the pioneering single-channel spectral subtraction method published by Boll and by Berouti et al. A few years later, with the introduction of the seminal papers by Ephraim and Malah on the estimation of the spectral amplitude and the log-spectral amplitude (LSA), statistically optimal methods became dominant. Beyond the statistically optimal estimation under the Gaussian assumption on the speech spectral components, these papers also introduced novel concepts related to estimation under signal presence uncertainty as well as the decision-directed approach for the a priori signal-to-noise ratio (SNR) estimation. Extensions to other probability distributions, e.g., super-Gaussian, were later presented. Comprehensive surveys of the state of the art in the first decade of the 21st century can be found in [34] and [35].

While it was assumed for many years that the estimation of the phase is unimportant and that it is sufficient to estimate the amplitude spectrum of the speech and augment it with the noisy phase, recent findings have shown that it is beneficial to estimate the phase as well [36].

All-pole modeling of the speech signal, widely used in traditional speech compression algorithms, was adopted by Lim and Oppenheim to develop an iterative scheme, alternating between the estimation of the speech autoregressive coefficients and enhancing the speech signal using Wiener filtering. The same speech model was later used under the EM framework, with a Kalman filter substituting the Wiener filter.

An early data-driven model for speech enhancement was proposed in [37]. In this work, rather than using a specific model for the LSA of the speech, a mixture of Gaussians model is inferred in a training stage using the entire TIMIT database. In recent years, the field of single-microphone speech enhancement (including noise reduction) has been dominated by DNN-based algorithms. Many of these algorithms recast the noise reduction problem as a mask estimation. The ideal binary mask (IBM) determines for each time-frequency bin whether it is dominated by speech or noise. Another popular mask is the ideal ratio mask (IRM), which is a softer version of the IBM. A survey of many noise reduction algorithms can be found in [38], where other masks, e.g., the complex IRM, which is also sensitive to the phase, are explored and compared. Although already achieving remarkable results, there are still many challenges left. Many of the algorithms require huge amounts of speech and noise data for training, and the resulting models are usually very large. There is a growing interest in developing "thin" models that can be deployed in edge devices, such as cellular phones, and even simpler devices that are used as nodes in WASNs. Moreover, in most telecommunication applications, low latency is mandatory, rendering utterance-level algorithms inadequate. There are many challenging acoustic environments that require further algorithmic improvements. One example is busy cafés and bars, usually characterized by babble noise. Another example is factories and mines, characterized by extreme noise levels. A third example is transient noise, e.g., keyboard typing and wind noise.

Spatial filtering (beamforming)

The enhancement and separation capabilities offered by multichannel interfaces are usually greater than those of single-channel interfaces, although DNN-based single-microphone solutions now offer competitive performance. We have explored data-independent beamformers. This section is dedicated to data-dependent beamformers, namely, beamformers that adapt to the received microphone signals. Early multimicrophone speech enhancement and speaker separation solutions adopted beamforming techniques with free-field propagation models [1]. Early attempts to incorporate statistically optimal solutions in the beamformer design as well as advanced speaker localization algorithm are summarized in [39].

As discussed in the preceding, sound fields in acoustic enclosures are typically characterized by high-order multipath propagation. If the number of microphones is too small to form narrow beams, using only the direct path of the AIR may provide insufficient sound quality. It therefore became common to take into consideration the entire AIR in the beamformer design. The concept of designing a matched filter toward multiple reflections of the sound was first introduced by Jan and Flanagan in 1996, but without discussing AIR estimation procedures.

In [40], the acoustic transfer function (ATF) relating the speaker and a microphone array was estimated using a subspace tracking procedure and used in the design of a minimum variance distortionless response (MVDR) beamformer. The relative transfer function (RTF) was later introduced and used in the MVDR design as a substitute for the ATF. The RTF encompasses the relevant information regarding the acoustic propagation between the source and a pair of microphones. Multiple optimal design criteria were used in the literature of microphone arrays, namely, the MVDR, the multichannel Wiener filter (MWF) and its variant the speech distortion-weighted MWF [41], the maximum SNR, and the linearly constrained minimum variance (LCMV). The latter addresses the speaker extraction problem, which is closely related to (semi-) blind speaker separation, as discussed in the next section of this article. Here, we only briefly note that microphone array processing and BSS paradigms are now strongly interrelated and routinely borrow ideas from each other. Further elaboration on spatial processing algorithms can be found in [42] and [43], including spatial processing criteria and algorithms and the relation to blind speaker separation.

While general-purpose multimicrophone speech enhancement algorithms aim at selectively enhancing the desired speech source and suppressing interfering sources and ambient background noise, the objective of binaural algorithms is also to preserve the auditory impression of the acoustic scene. This can be achieved by preserving the so-called binaural cues of the desired speech source, interfering sources, and background noise such that the binaural hearing advantage of the auditory system can be exploited, and confusions due to a mismatch between acoustic and visual information are avoided. A range of multichannel filters to achieve this goal is surveyed in [43, Ch. 18].

All criteria discussed in the preceding were designed for centralized processing. In WASNs, when such processing becomes too expensive, either optimal or suboptimal distributed algorithms should be applied instead. The outcome of the optimal distributed algorithms should be identical to their centralized counterparts, while for suboptimal algorithms, some performance degradation may result. The advantage of the latter family of algorithms is reduced communication bandwidth and sometimes even a lower local computational load. The challenges typical to WASN processing, several important applications, and several efficient node fusion schemes can be found in [44]. Distributed versions of many of the preceding criteria can be found in the literature. In WASN processing, sampling rate synchronization may be crucial for guaranteeing the proper operation of the system. Multiple resynchronization schemes can be found in the literature.

A large number of DNN-based spatial processing algorithms were proposed in recent years. Three main trends can be found in the current literature. In the first line of work, the DNN is used for estimating the building blocks of the statistically optimal beamformers. In the second line of work, e.g., in [45], the DNN directly estimates the multichannel weights of the beamformer. The advantage of the latter is the ability to go beyond the conventional second-order statistics and implement a beamformer with perceptually more meaningful cost functions (or with the WER as a loss function in ASR applications). However, this may not be as robust as the DNN-controlled beamformers. In the third line of work, the DNN is directly applied to the multichannel data, and the beamformer structure is not preserved.

Audiovisual signal enhancement

The visual modality can clearly support the enhancement task. As an example, focusing on the face of the speaker, and particularly the lips, can be used to extract the desired speaker from background noise and competing speakers [46].

Signal separation

Source separation and blind source separation (BSS) were topics of growing interest in the mid-1990s and gradually moved from determined and overdetermined cases to the more challenging underdetermined case, where there are potentially more sources than observed mixtures [47]).

Determined case

BSS started as an application of independent component analysis (ICA). A series of ICA conferences began in 1999 and were held in 1.5-year intervals, playing an important role in promoting the field. Audio signals are, due to TDoAs of the source signals arriving at different sensors and reverberation, convolutively mixed in a room. Because a convolutive mixture in the time domain can be converted to instantaneous mixtures in the frequency domain, the frequency-domain ICA approach converts time-domain signals into the time-frequency domain by using a short-time Fourier transform (STFT). ICA theory inherently includes two ambiguities: output order (permutation) and output amplitude (scaling). Both become serious problems in frequency-domain ICA. To solve the permutation problem, spatial information and spectral information of the sources are key information. It was further shown that ICA-based BSS forms a null directivity pattern toward the interfering source and suppresses it [48].

An interesting framework for multichannel blind signal processing for convolutive mixtures, known as *Triple-N ICA for Convolutive Mixtures* [49], defines an information-theoretic cost function and enables the utilization of three fundamental signal properties, namely, nonwhiteness, non-Gaussianity, and nonstationarity. Nonnegative matrix factorization (NMF) [50]

separates sources by using common frequency patterns as frequency bases. Independent low-rank matrix analysis [51] separates sources by using spatial information of ICA and spectral information of NMF. As in most fields of audio processing, deep learning methods are now widely used, and some of them are improved variants of classical algorithms. For instance, the multichannel variational autoencoder (VAE) [52] combines spatial information of ICA and spectral information of DNNs. Audio source separation methods and algorithms are surveyed in [43] and [53].

Monophonic separation

Although multichannel separation provided a way to invert mixing, the case in which the input mixture is presented in a single channel only, known as *monophonic separation*, posed a new challenge. Techniques that emerged in this area utilized either generative modeling or variations of masking approaches to recover the intended source. This problem also brought into the spotlight the idea of trained separation algorithms as opposed to blind methods.

An early successful approach along these lines came from models based on NMF [50]. These models were pretrained using sound examples, learned a target-specific spectral dictionary, and were able to isolate and reconstruct such a target from an input mixture. Variations of this approach included multichannel versions, convolutional models, models trained on a variety of spectrotemporal representations, Markov models, probabilistic formulations, and more [54], [55].

Although generative models performed well at the time, an alternative approach came from a technique that was first used for multichannel separation. W-disjoint orthogonality [56] took advantage of sparsity in the time-frequency representation of most sounds to directly apply a binary mask on a spectrogram and isolate the desired sound. First formulated for stereo recordings, this idea became a cornerstone for approaches based on NNs and resulted in a discriminative approach to solving the separation problem, where each time-frequency point is classified as useful or not. A popular NN model that made use of this idea was deep clustering [57], which projected mixtures in a space where time-frequency bins could be clustered and labeled accordingly as belonging to independent sources. Other NN models dispensed with the clustering step, thereby losing some generality, and directly attempted to predict a mask given just an input mixture [38]. The latter approach has dominated the source separation research of late, providing many approaches with impressive-sounding results, ranging in their application from small and efficient on-device speech enhancers that are commonly used for most voice communication today to larger high-quality offline models, such as those used for the award-winning restorations of historical Beatles recordings. Models along these lines have explored many of the new neural architectures (the U-net, transformers, and so on) and span a wealth of extensions, such as the use of soft masks, models that learn a latent space as opposed to using an STFT [58], models that resolve ambiguity in the order of output sources (permutation-invariant training, conditional models that are guided toward a target by a user, models that directly optimize perceptual metrics, and more). In Figure 3, several examples of approaches for monophonic separation are given.

A special case of these models has had a significant impact on music processing. The release of easy-to-use music-oriented source separation models (https://research.deezer.com/projects/spleeter) has resulted in a wealth of free and commercial software that allows users to decompose a music recording into its constituent instrument tracks and freely remix and manipulate. Aside from being a very useful tool, this has enhanced the way we interact with recorded music and opened new avenues of media interactivity that are still being explored.

Although discriminative models offer superior performance with relative ease of use, their downside as compared to generative methods is that they are prone to overspecialization and cannot be easily extended and redeployed for alternative uses. Some open questions still remain on how to make universal separators, learn with limited training data, extend a trained model to work out-of-distribution, and so on. Despite the impressive-sounding demos, there is still a lot of work to be done in this space.

Emerging topics

Another viewpoint of the evolution and breakthrough discussed in the preceding is the emergence of new topics, almost absent in the 1990s and that today are among the most popular fields.

Objective evaluation

Objective evaluation of speech and audio quality has emerged as a highly relevant topic in the past 25 years. If the ultimate means for speech/audio quality evaluation and intelligibility assessment is a human perceptual test, it is also known that it is costly and tedious to organize. This has motivated the community to develop objective metrics for sound quality that are better correlated with perception. For instance, led by the speech coding community, several speech quality metrics were developed (and standardized), including Perceptual Evaluation of Speech Quality, Perceptual Objective Listening Quality Assessment, and Virtual Speech Quality Objective Listener. An overview of objective perceptual measures is provided in [59]. There is also widespread adoption of speech intelligibility measures for hearing aids, such as Short-Time Objective Intelligibility (STOI) together with binaural extensions: modified binaural STOI. These measures are the de facto standard for assessing the impact of speech enhancement algorithms in human interface devices. Similarly, several metrics were proposed to evaluate audio quality (such as Perceptual Evaluation of Audio Quality and Perception Model-Based Quality) and the performance of an audio source separation algorithm (the scale-invariant signal-to-distortion ratio, signal-to-artifact ratio, and signal-to-interference ratio) [60]. Other interesting objective measures were also proposed, in particular for hearing-impaired listeners (see [61] for an overview).

More recently, we have also seen the incorporation of trained models that output perceptual scores [62]. These models can be

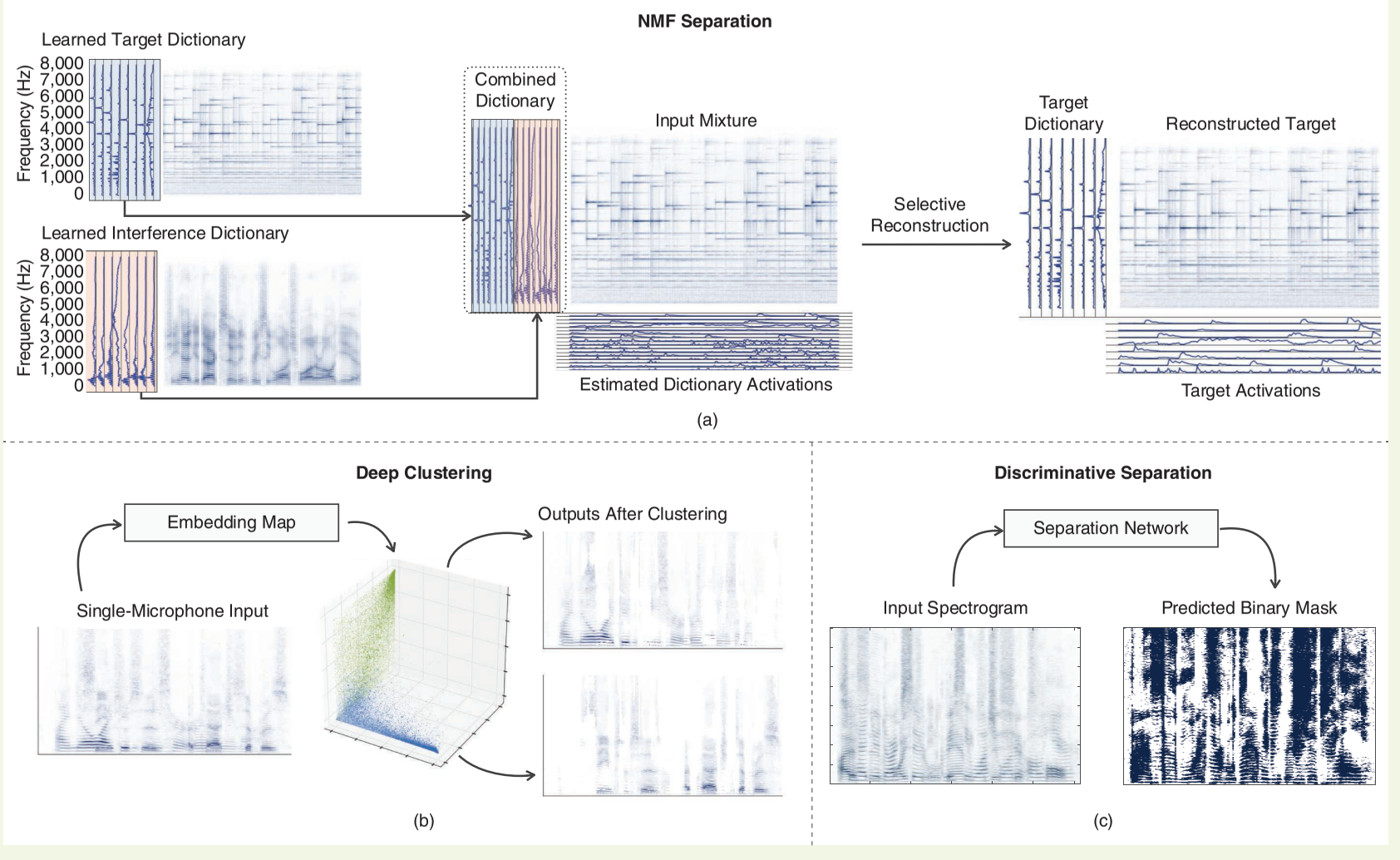

FIGURE 3. Approaches for monophonic separation. (a) NMF models decompose inputs based on trained dictionaries and then use that information to reconstruct selected parts of the input. (b) Deep clustering projects the time-frequency points of a mixture to a latent space adapted such that different sources cluster separately, and it then uses the cluster labels to reconstruct each source. (c) Finally, discriminative separation, mostly based on NNs, predicts masking functions directly from input signals such that it can mute interference and isolate target sounds.

trained on audio inputs to directly predict user responses and provide a rapid alternative to listener tests and otherwise slow-to-compute evaluation methods. When used with differentiable models, these evaluation methods can also be directly incorporated into algorithm optimization, providing new possibilities for training perceptually relevant systems.

Finally, when any of the approximations in the preceding are not deemed sufficient, audio algorithm designers can resort to modern crowdsourcing tools that can reach thousands of listeners and conduct experiments with unprecedented sample sizes. The ability to do this has revolutionized how audio products are evaluated today and provides stronger statistical results than ever before.

MIR

MIR is defined as a field that covers all the research topics involved in the understanding and modeling of music and that uses information processing methodologies (see the MIR road map at http://www.mires.cc/wiki/index1a1d.html?title=Roadmap&oldid=2137). It is, in essence, an interdisciplinary domain involving machine learning, signal processing, and/or musicology. The nature of the processed music can also be very diverse, including the raw audio signal, a symbolic representation of the music score or recording (for example, in the musical instrument digital interface format), an image (for example, as a scanned version of the music score), and even as 3D trajectory movements (for example, as gestures of performers). If the MIR domain has initially focused on symbolic music processing, some early studies have paved the way for many subsequent works on raw audio signals, for example, in speech/music discrimination, beat tracking [63], and music analysis and recognition [64], to name a few. The early approaches often took inspiration from speech recognition methods, mostly using mel-frequency cepstral coefficients (MFCCs) as features, with statistical models such as Gaussian mixture models (GMMs), hidden Markov models (HMMs), support vector machines (SVMs), and more. Similarly, for underdetermined source separation, major progress was made in using dedicated low-rank and sparse decomposition, such as based on NMF and matching pursuit and its variants. With the exception of some early papers that exploited NNs (see, for example, [65] for multipitch estimation), the advent of deep learning is rather recent (see Figure 4). Today, the major trend is to consider deep learning for nearly all applications, with remarkable achievements in polyphonic music source separation, music transcription (estimation of melody, harmony, rhythm, lyrics, and so on), music style transfer, and music synthesis, for instance, [66]. As in speech recognition, the field has also received a great interest toward end-to-end deep learning approaches, which even replace the traditional feature extraction step with a data-driven representation learning paradigm.

The variety and complexity of music signals also motivate the development of new tailored methods for representation learning and unsupervised learning to avoid the particularly cumbersome stage of music signal annotation. A particularly interesting approach was recently introduced for self-supervised pitch estimation [67]. Besides the main historic domains of MIR, music synthesis is becoming a stronger field with impressive results, especially around new generative models. In recent years, we have witnessed the emergence of approaches at the crossroads of DNNs and classical generative models in so-called deep generative models. Some of the most popular models include different forms of autoencoders (including VAEs, autoregressive models, and GANs). A concurrent trend, especially for music generation, revisits the use of classic audio signal models, such as, for instance, the source–filter model of speech production and the harmonic + noise model. In fact, such models have great potential in hybrid neural architectures integrating audio models under the form of differentiable signal processing blocks [68]. Hybrid architectures are indeed particularly attractive and already show great promise. For instance, the use of differentiable source generative models opens the path to data-efficient fully unsupervised music source separation paradigms [69].

DCASE

Nevertheless, the most recent and strongest growth has been in the field of DCASE [70]. This growing interest is visible

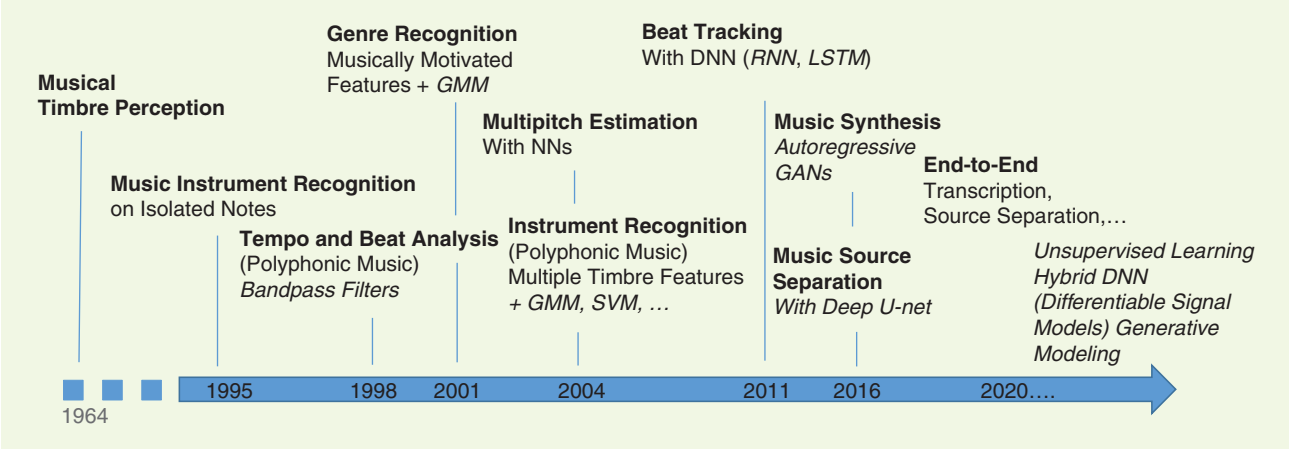

FIGURE 4. MIR: a rather early adoption of DNNs. RNN: recurrent NN; LSTM: long short-term memory.

in the increase of the DCASE community and the success of its DCASE workshop, a series launched in 2016 (with attendance growing from 68 in 2016 to 201 in 2019, with an average of 50% from the industry), and its companion international challenge (with continuous growth of the number of submitted systems, from 84 in 2016 to 470 in 2020). (Note, though, that the very first DCASE challenge was organized in 2013, but it became an annual event only from 2016.) This steady increase of interest is clearly visible in the number of submissions to ICASSP: in 2022, DCASE was by far the field with the highest number of submissions, with up to 23.5% of all submissions in audio. Although very important work on the perception of sound objects was reported by Schaeffer in his treatise on musical objects in the 1960s, one often refers to computational auditory scene analysis and the work on acoustic scene analysis by Bregman in the early 1990s as the most emblematic initial work in DCASE.

As illustrated in Figure 5, this field has seen a similar (although much faster) evolution from speech recognition-inspired methods to fully data-driven deep learning methods, with a particularly strong axis on weakly supervised approaches [71].

With the notable exception of work by Sawhney and Maes in 1997, which exploited NNs, most of the studies until 2015 relied on more traditional clustering and machine learning paradigms, for instance, based on the SVM, GMM, and HMM. Also, similar to the domains of audio source separation and MIR at the dawn of the 21st century, many works have exploited approaches to obtain compact and informative audio signal representations. Sparse decomposition methods, image-based features, and NMFs have been particularly popular. Then, since 2014, deep learning has gained strong momentum and very rapidly become the mainstream architecture. In the DCASE 2016 challenge, all submitted systems for acoustic scene classification but four involved NNs, even if they were not yet defining the state of the art. Two years later, in the 2018 challenge, the top 30 performing systems were DNN-based, confirming the indisputable supremacy of NNs for such a task. Although DCASE often refers to a single domain, it considers, in practice, multiple applications, which have their own specifics and constraints. In acoustic scene recognition, a more mature application, numerous approaches were proposed to operate at low complexity, and in that regard, the use of network compression, pruning, and knowledge distillation, for instance, exploiting teacher–student frameworks, are among the most successful developments. For the task of acoustic events detection and localization, there is easy access to huge weakly annotated databases. This has obviously accompanied the emergence of an anthology of weakly supervised and few-shot learning approaches, for instance, around prototypical networks and mean teacher architectures, which are particularly efficient for few-shot learning, weakly supervised learning, and domain adaptation. Finally, it is worth mentioning the wide use of data augmentation techniques, which have proved, in many domains, to be very efficient to reduce model overfitting. Popular data augmentation techniques include SpecAugment (with feature warping and time-frequency masking), pitch shifting, time stretching, mix-up and channel confusion in the case of multichannel recordings, random noise addition, and many more.

Powerful consumer electronics devices and fast Internet connections

Finally, recent years have witnessed a very fast deployment of powerful consumer electronics devices with audio processing capabilities and usually with more than a single microphone. Example devices are laptops, tablet PCs, cellular phones, smartphones and smart watches, smart speakers, hearing devices and hearables, smart loudspeakers (Amazon Echo, Apple HomePod, and Google Home), and virtual and augmented reality glasses. Dedicated multimicrophone hardware, e.g., spherical microphone arrays, is also available (see the Eigenmike at https://mhacoustics.com).

Concurrently, the rapid deployment of fast Internet connections, specifically with data over the cellular network, dramatically changed the way we communicate. Rather than

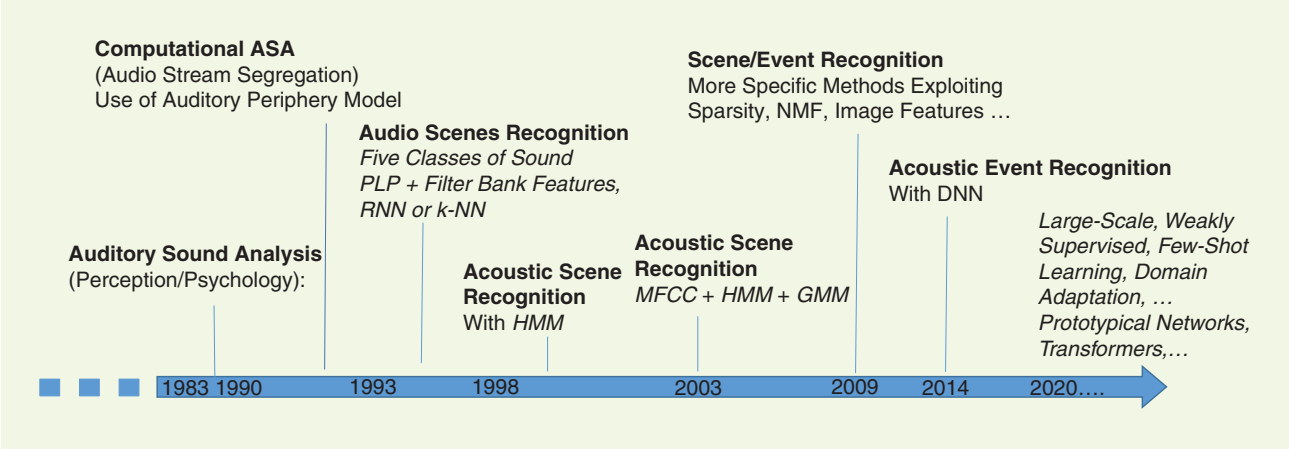

FIGURE 5. DCASE: from perceptual auditory sound analysis to large-scale deep learning algorithms. ASA: auditory scene analysis; PLP: perceptual linear predictive; *k*-NN: *k*-nearest neighbors.

communicating over the wired telephone network and, later, the cellular network, we now widely use voice over Internet Protocol (VoIP) as a cheap and reliable alternative. Moreover, teleconferencing tools, e.g., Google Meet, Skype, and Zoom, have become very popular, as recently demonstrated during the COVID-19 pandemic, allowing everyone to work from home and remotely communicate with colleagues and coworkers. The VoIP technology promoted research on audio coding, packet loss concealment, and echo cancellation over IP. Similarly, the widespread use of the Internet has revolutionized the consumption of music through new applications, such as audio and music retrieval and music identification [e.g., the popular Shazam service (https://www.shazam.com)], and around streaming services with automatic recommendation and playlist generation.

Conclusions and perspectives

The domain of AASP is clearing experiencing growing interest, with a broad range of specific and interdisciplinary research and development. This growth has been accompanied by the AASP TC, whose "mission is to support, nourish and lead scientific and technological development in all areas of audio and acoustic signal processing." Over the years, and especially recently, the domain has shifted toward more data-driven methods for nearly all speech and audio applications. In some cases, the methods developed are pure end-to-end approaches, where all the "knowledge" is extracted from data. We believe that this is a very strong trend that will be further developed in the future but probably with a different angle. In fact, pure end-to-end deep neural approaches are complex, often overparametrized, and, in many cases, remain rather unexplainable. There is thus an interest to go toward more frugal data-driven and interpretable and controllable systems. A potential path is to combine the strength of data-driven paradigms with efficient signal models to build new model-based (and hybrid) deep neural architectures. For example, in MIR, it is possible to associate differentiable sound production models and deep learning architectures to design interpretable, more frugal, and yet efficient methods. This may be one of the future paths toward developing new algorithms and technologies that will be in accordance with sustainable and ecological development and compliant with high ethical standards, which we believe will become general concerns of major importance.

Another future research direction that should receive growing interest in audio processing is federated (or collaborative) learning [72]. In fact, massive amounts of data are now stored on devices. As a result, more models can now be directly trained on devices (often referred to as *on the edge*). This allows us to better take into account privacy concerns (recorded data are not stored centrally) but also brings a number of challenges for audio applications, particularly in global optimization with communication constraints, learning with heterogeneous data (audio data recorded from diverse and heterogeneous recording devices), and learning with partial and missing data. Federated learning, which gathers techniques for machine learning and statistical signal processing using multiple distributed devices, then, appears as a particularly promising framework for future audio processing applications. Stronger edge devices, with more powerful processing units and faster communication capabilities, will certainly support this trend.

We also expect that multimodal processing will become more prominent and that we will witness, in the near future, more algorithms that utilize vision to support speaker localization and separation. Beyond audiovisual processing, other modalities will be more extensively used, e.g., brain-informed speech separation using electroencephalography signals [73].

Authors

Gaël Richard (gael.richard@telecom-paris.fr) received his State Engineering degree from Telecom Paris, France, in 1990, and his Ph.D. degree from the University of Paris-Saclay, in 1994. He is now a full professor of audio signal processing at Telecom-Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France, and the scientific codirector of the Hi! PARIS interdisciplinary center on artificial intelligence and data analytics. He is the past chair of the IEEE Signal Processing Society Technical Committee for Audio and Acoustic Signal Processing. He received, in 2020, the IMT–National Academy of Science grand prize for his research contribution in sciences and technologies. In 2022, he was awarded an advanced European Research Council grant for a project on hybrid deep learning for audio. His research interests include machine learning, audio and music signal processing. He is a Fellow of IEEE.

Paris Smaragdis (smaragdis@ieee.org) received his Ph.D. degree from the Massachusetts Institute of Technology in 2001. He is now a professor of computer science at the University of Illinois Urbana-Champaign, Champaign, IL 61801, USA. He was an IEEE Distinguished Lecturer (2016–2017) and has previously chaired the IEEE Data Science Initiative, IEEE Audio and Acoustics Signal Processing Technical Committee, and IEEE Machine Learning for Signal Processing Technical Committee. He has also served as a member of the IEEE Signal Processing Society Board of Governors. He is currently the editor-in-chief of *IEEE Transactions on Audio, Speech, and Language*. His research interests include machine learning and signal processing. He is a Fellow of IEEE.

Sharon Gannot (sharon.gannot@biu.ac.il) received his Ph.D. degree in electrical engineering from Tel-Aviv University, Israel in 2000. He is currently a full professor in the Faculty of Engineering, Bar-Ilan University, Ramat Gan 5290002, Israel. He currently serves as a senior area chair for *IEEE Transactions on Audio, Speech, and Language Processing*; a member of the senior editorial board of *IEEE Signal Processing Magazine*; and the chair of the IEEE Signal Processing Society Data Science Initiative. He also served as the chair of the IEEE Audio and Acoustic Signal Processing Technical Committee in 2017–2018. He was the general cochair of the 2010 International Workshop on Acoustic Signal Enhancement and 2013 IEEE Workshop on

Applications of Signal Processing to Audio and Acoustics. He is a recipient of the 2022 European Association for Signal Processing Group Technical Achievement Award. His research interests include statistical signal processing and machine learning methods for single- and multi-microphone arrays, with applications to speech enhancement, noise reduction and speaker separation, and diarization, dereverberation, speaker localization, and tracking. He is a Fellow of IEEE.

Patrick A. Naylor (p.naylor@imperial.ac.uk) received his Ph.D. degree from Imperial College London in 1990. He is now a professor of speech and acoustic signal processing at Imperial College London, SW7 2AZ London, U.K. He has served on the IEEE Signal Processing Society Board of Governors, as chair of the IEEE Audio and Acoustic Signal Processing Technical Committee, as an associate editor of *IEEE Signal Processing Letters*, and as a senior area editor of *IEEE Transactions on Audio, Speech, and Language Processing*. He is a past president of the European Association for Signal Processing. His research interests include microphone array signal processing, speaker diarization and localization, and multichannel speech enhancement for application to binaural hearing aids. He is a Fellow of IEEE.

Shoji Makino (s.makino@ieee.org) received his Ph.D. degree from Tohoku University in 1993. He is currently a professor at Waseda University, Kitakyushu 808-0135 Japan. He has served on the IEEE Signal Processing Society (SPS) Board of Governors, SPS Technical Directions Board, SPS Awards Board, and SPS Fellow Evaluation Committee. He has received 30 awards, including the SPS Leo L. Beranek Meritorious Service Award in 2022, SPS Best Paper Award in 2014, IEEE Machine Learning for Signal Processing Competition Award in 2007, and ICA Unsupervised Learning Pioneer Award in 2006. His research interests include adaptive filtering technologies, acoustic signal processing, and machine learning for speech and audio applications. He is an SPS Distinguished Lecturer and a Fellow of IEEE.

Walter Kellermann (walter.kellermann@fau.de) received his Dr.-Ing. degree from the Technical University Darmstadt in 1988. He is currently a professor at the University of Erlangen-Nürnberg, 91058 Erlangen, Germany. He was a Distinguished Lecturer and a Vice President Technical Directions of the IEEE Signal Processing Society. He is a corecipient of ten best paper awards and a recipient of the Group Technical Achievement Award of the European Association for Signal Processing (EURASIP). His main research interests focus on physical model-based and data-driven multichannel methods for acoustic signal processing and speech enhancement. He is a Fellow of EURASIP and a Life Fellow of IEEE.

Akihiko Sugiyama (a.sugiyama@ieee.org) received his Dr. Eng. from Tokyo Metropolitan University in 1998. He is currently working for Yahoo Japan Corporation, Tokyo 102-8282, Japan. He served as the chair of the IEEE Audio and Acoustic Signal Processing Technical Committee, an associate editor of *IEEE Transactions on Signal Processing*, and a member of IEEE Fellow Committee. He was a technical program chair for ICASSP 2012, past IEEE Signal Processing Society (SPS) Distinguished Industry Speaker, and past SPS and IEEE Consumer Technology Society Distinguished Lecturer. His research interests include audio coding and interference/noise control. He is a Fellow of IEEE.

References

[1] M. Kahrs et al., "The past, present and future of audio signal processing," *IEEE Signal Process. Mag.*, vol. 14, no. 5, pp. 30–57, Sep. 1997, doi: 10.1109/MSP.1997.1179708.

[2] F. Baumgarte and C. Faller, "Binaural cue coding-part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 509–519, Nov. 2003, doi: 10.1109/TSA.2003.818109.

[3] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 8, pp. 1699–1712, Aug. 2013, doi: 10.1109/TASL.2013.2260153.

[4] A. Sugiyama and M. Iwadare, "The origin of digital information devices: The silicon audio and its family," *APSIPA Trans. Signal Inf. Process.*, vol. 7, no. 1, Jan. 2018, Art. no. e1, doi: 10.1017/ATSIP.2017.16.

[5] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979, doi: 10.1121/1.382599.

[6] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "More than 50 years of artificial reverberation," *J. Audio Eng. Soc.*, Jan. 2016.

[7] E. Nosal, M. Hodgson, and I. Ashdown, "Improved algorithms and methods for room sound-field prediction by acoustical radiosity in arbitrary polyhedral rooms," *J. Acoust. Soc. Amer.*, vol. 116, no. 2, pp. 970–980, Sep. 2004, doi: 10.1121/1.1772400.

[8] H. Bai, G. Richard, and L. Daudet, "Late reverberation synthesis: From radiance transfer to feedback delay networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2260–2271, Dec. 2015, doi: 10.1109/TASLP.2015.2478116S.

[9] B. Rafaely, *Fundamentals of Spherical Array Processing*. Berlin, Germany: Springer-Verlag, 2015.

[10] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Berlin, Germany: Springer, 2001, pp. 157–180.

[11] A. Brendel and W. Kellermann, "Distributed source localization in acoustic sensor networks using the coherent-to-diffuse power ratio," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 61–75, Mar. 2019, doi: 10.1109/JSTSP.2019.2900911.

[12] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, Mar. 2019, doi: 10.1109/JSTSP.2019.2901664.

[13] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Amer.*, vol. 152, no. 1, pp. 107–151, Jul. 2022, doi: 10.1121/10.0011809.

[14] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018, doi: 10.1109/TASLP.2018.2828321.

[15] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Cham, Switzerland: Springer, 2017.

[16] G. W. Elko, "Differential microphone arrays," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Boston, MA, USA: Springer, 2004, pp. 11–65.

[17] J. Benesty and C. Jingdong, *Study and Design of Differential Microphone Arrays*. Berlin, Germany: Springer-Verlag, 2012.

[18] D. Begault and L. Trejo, "3-D sound for virtual reality and multimedia," Nat. Aeronaut. Space Admin., Washington, DC, USA, NASA/TM-2000-209606, 2000.

[19] D. Poirier-Quinot and B. F. Katz, "On the improvement of accommodation to non-individual HRTFs via VR active learning and inclusion of a 3d room response," *Acta Acoust.*, vol. 5, no. 25, pp. 1–17, Jun. 2021, doi: 10.1051/aacus/2021019.

[20] J. Ahrens and S. Spors, "Sound field reproduction using planar and linear arrays of loudspeakers," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 2038–2050, Nov. 2010, doi: 10.1109/TASL.2010.2041106.

[21] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 852–866, Aug. 2015, doi: 10.1109/JSTSP.2015.2415762.

[22] S. Cecchi, A. Carini, and S. Spors, "Room response equalization - A review," *Appl. Sci.*, vol. 8, no. 1, 2018, Art. no. 16, doi: 10.3390/app8010016.

[23] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Process.*, vol. 80, no. 9, pp. 1747–1760, Sep. 2000, doi: 10.1016/S0165-1684(00)00085-2.

[24] M. M. Halimeh, C. Huemmer, and W. Kellermann, "A neural network-based nonlinear acoustic echo canceller," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1827–1831, Dec. 2019, doi: 10.1109/LSP.2019.2951311.

[25] T. Gänsler and J. Benesty, "The fast normalized cross-correlation double-talk detector," *Signal Process.*, vol. 86, no. 6, pp. 1124–1139, Jun. 2006, doi: 10.1016/j.sigpro.2005.07.035.

[26] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Berlin, Germany: Springer-Verlag, 2008.

[27] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. New York, NY, USA: Wiley, 2005.

[28] T. Van Waterschoot and M. Moonen, "Fifty years of acoustic feedback control: State of the art and future challenges," *Proc. IEEE*, vol. 99, no. 2, pp. 288–327, Feb. 2011, doi: 10.1109/JPROC.2010.2090998.

[29] Y. J. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 6, pp. 1711–1720, Aug. 2011, doi: 10.1109/TASL.2010.2097249.

[30] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. London, U.K.: Springer-Verlag, 2010.

[31] E. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, Sep. 2009, doi: 10.1109/LSP.2009.2024791.

[32] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010, doi: 10.1109/TASL.2010.2052251.

[33] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015, doi: 10.1109/TASLP.2015.2416653.

[34] R. C. Hendriks, T. Gerkmann, and J. Jensen, "DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," in *Synthesis Lectures Speech Audio Processing*, vol. 9. San Rafael, CA, USA: Morgan & Claypool, 2013, pp. 1–80.

[35] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[36] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015, doi: 10.1109/MSP.2014.2369251.

[37] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 341–351, Oct. 2002, doi: 10.1109/TSA.2002.803420.

[38] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018, doi: 10.1109/TASLP.2018.2842159.

[39] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[40] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 425–437, Sep. 1997, doi: 10.1109/89.622565.

[41] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel wiener filtering techniques for noise reduction," in *Speech Enhancement*. Berlin, Germany: Springer-Verlag, 2005, pp. 199–228.

[42] 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, Language Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017, doi: 10.1109/TASLP.2016.2647702.

[43] E. Vincent, T. Virtanen, and S. Gannot, Eds. *Audio Source Separation and Speech Enhancement*. Hoboken, NJ, USA: Wiley, 2018.

[44] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. IEEE Symp. Commun. Veh. Technol. Benelux (SCVT)*, 2011, pp. 1–6, doi: 10.1109/SCVT.2011.6101302.

[45] X. Xiao, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2016, pp. 5745–5749, doi: 10.1109/ICASSP.2016.7472778.

[46] A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," 2018, *arXiv:1804.03619*.

[47] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. New York, NY, USA: Academic, 2010.

[48] 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. Speech Audio Process.*, vol. 11, no. 2, pp. 109–116, Apr. 2003, doi: 10.1109/TSA.2003.809193.

[49] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, Feb. 2005, doi: 10.1109/TSA.2004.838775.

[50] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007, doi: 10.1109/TASL.2006.876726.

[51] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016, doi: 10.1109/TASLP.2016.2577880.

[52] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Comput.*, vol. 31, no. 9, pp. 1891–1914, Sep. 2019, doi: 10.1162/neco_a_01217.

[53] S. Makino, *Audio Source Separation*. Cham, Switzerland: Springer, 2018.

[54] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 125–144, Mar. 2015, doi: 10.1109/MSP.2013.2288990.

[55] P. Smaragdis, C. Févotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 66–75, May 2014, doi: 10.1109/MSP.2013.2297715.

[56] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2002, vol. 1, pp. I–529–I–532, doi: 10.1109/ICASSP.2002.5743771.

[57] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2016, pp. 31–35, doi: 10.1109/ICASSP.2016.7471631.

[58] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019, doi: 10.1109/TASLP.2019.2915167.

[59] M. Torcoli, T. Kastner, and J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1530–1541, Mar. 2021, doi: 10.1109/TASLP.2021.3069302.

[60] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006, doi: 10.1109/TSA.2005.858005.

[61] T. H. Falk et al., "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015, doi: 10.1109/MSP.2014.2358871.

[62] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 631–635, doi: 10.1109/ICASSP.2019.8683175.

[63] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, Jan. 1998, doi: 10.1121/1.421129.

[64] J. Foote, "An overview of audio information retrieval," *Multimedia Syst.*, vol. 7, no. 1, pp. 2–10, Jan. 1999, doi: 10.1007/s005300050106.

[65] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 439–449, Jun. 2004, doi: 10.1109/TMM.2004.827507.

[66] G. Peeters and G. Richard, "Deep learning for audio and music," in *Multi-Faceted Deep Learning: Models and Data*, J. Benois-Pineau and A. Zemmari, Eds. Cham, Switzerland: Springer, 2021, pp. 231–266.

[67] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirović, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1118–1128, Mar. 2020, doi: 10.1109/TASLP.2020.2982285.

[68] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.

[69] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau, "Unsupervised audio source separation using differentiable parametric source models," 2022, *arXiv:2201.09592*.

[70] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015, doi: 10.1109/TMM.2015.2428998.

[71] T. Virtanen, D. Ellis, and M. Plumbley, Eds. *Computational Analysis of Sound Scenes and Events*. Cham, Switzerland: Springer, 2018.

[72] J. Konecný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, *arXiv:1610.02527*.

[73] E. Ceolini et al., "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *Neuroimage*, vol. 223, Dec. 2020, Art. no. 117282, doi: 10.1016/j.neuroimage.2020.117282.

SP