JULY 05 2022

A survey of sound source localization with deep learning methods

Pierre-Amaury Grumiaux; Srđan Kitić; Laurent Girin; Alexandre Guérin

J. Acoust. Soc. Am. 152, 107–151 (2022)
https://doi.org/10.1121/10.0011809

JASA REVIEW-ARTICLE

A survey of sound source localization with deep learning methods

Pierre-Amaury Grumiaux,[1,a)] Srđan Kitić,[2] Laurent Girin,[3] and Alexandre Guérin[2]
[1]*Nantes Université, École Centrale Nantes, CNRS, LS2N, 2 chemin de la Houssinière, F-44332 Nantes, France*
[2]*Orange Labs, 4 Rue du Clos Courtel, 35510 Cesson-Sévigné, France*
[3]*Univ. Grenoble Alpes, Grenoble-INP, GIPSA-lab, 11 Rue des Mathématiques, 38400 Saint-Martin-d'Hères, France*

ABSTRACT:
This article is a survey of deep learning methods for single and multiple sound source localization, with a focus on sound source localization in indoor environments, where reverberation and diffuse noise are present. We provide an extensive topography of the neural network-based sound source localization literature in this context, organized according to the neural network architecture, the type of input features, the output strategy (classification or regression), the types of data used for model training and evaluation, and the model training strategy. Tables summarizing the literature survey are provided at the end of the paper, allowing a quick search of methods with a given set of target characteristics. © *2022 Acoustical Society of America*. https://doi.org/10.1121/10.0011809

(Received 26 November 2021; revised 13 May 2022; accepted 6 June 2022; published online 5 July 2022)
[Editor: Peter Gerstoft] Pages: 107–151

I. INTRODUCTION

Sound source localization (SSL) is the problem of estimating the position of one or several sound sources relative to some arbitrary reference position, which is generally the position of the recording microphone array, based on the recorded multichannel acoustic signals. In most practical cases, SSL is simplified to the estimation of the sources' direction of arrival (DoA), i.e., it focuses on the estimation of azimuth and elevation angles, without estimating the distance to the microphone array (therefore, unless otherwise specified, in this article we use the terms "SSL" and "DoA estimation" interchangeably). SSL has numerous practical applications—for instance, in source separation (e.g., Chazan *et al.*, 2019), automatic speech recognition (ASR) (e.g., Lee *et al.*, 2016), speech enhancement (e.g., Xenaki *et al.*, 2018), human-robot interaction, (e.g., Li *et al.*, 2016a), noise control, (e.g., Chiariotti *et al.*, 2019), and room acoustic analysis (e.g., Amengual Garí *et al.*, 2017). In this paper, we focus on sound sources in the audible range (typically speech and audio signals) in indoor (office or domestic) environments.

Although SSL is a long-standing and widely researched topic (Argentieri *et al.*, 2015; Benesty *et al.*, 2008; Brandstein and Ward, 2001; Cobos *et al.*, 2017; DiBiase *et al.*, 2001; Gerzon, 1992; Hickling *et al.*, 1993; Knapp and Carter, 1976; Nehorai and Paldi, 1994), it remains a very challenging problem to date. Traditional SSL methods are based on signal/channel models and signal processing (SP) techniques. Although they have shown notable advances in the domain over the years, they are known to perform poorly in difficult yet common scenarios where noise, reverberation, and several simultaneously emitting sound sources may be present (Blandin *et al.*, 2012; Evers *et al.*, 2020). In the last decade, the potential of data-driven deep learning (DL) techniques for addressing such difficult scenarios has received an increasing interest. As a result, an increasing number of SSL systems based on deep neural networks (DNNs) have been proposed in recent years. Most of the reported works have indicated the superiority of DNN-based SSL methods over conventional (i.e., SP-based) SSL methods. For example, Chakrabarty and Habets (2017a) showed that, in low signal-to-noise ratio conditions, using a CNN led to a twofold increase in overall DoA classification accuracy compared to using the conventional method called steered response power with phase transform (SRP-PHAT) (see Sec. III). In Perotin *et al.* (2018b), the authors were able to obtain a 25% increase in DoA classification accuracy when using a convolutional recurrent neural network (CRNN) over a method based on independent component analysis (ICA). Finally, Adavanne *et al.* (2018) proved that employing a CRNN can reduce the average angular error by 50% in reverberant conditions compared to the conventional MUSIC algorithm (see Sec. III).

This kind of result has further motivated the expansion of scientific papers on DL applied to SSL. In the meantime, there has been no comprehensive survey of the existing approaches, which would be very useful for researchers and practitioners in the domain. Although we can find reviews mostly focused on conventional methods, e.g., (Argentieri *et al.*, 2015; Cobos *et al.*, 2017; Evers *et al.*, 2020; Gannot *et al.*, 2019), to the best of our knowledge only a very few

[a)]Also at: Orange Labs, 4 Rue du Clos Courtel, F-35510 Cesson-Sévigné, France, and at Univ. Grenoble Alpes, Grenoble-INP, CNRS, GIPSA-lab, 11 Rue des Mathématiques, F-38400 Saint-Martin-d'Hères, France. Electronic mail: pierreamaury.grumiaux@gmail.com

have explicitly targeted SSL with DL methods. Ahmad *et al.* (2021) presented a short survey of several existing DL models and datasets for SSL before proposing a DL architecture of their own. Bianco *et al.* (2019) and Purwins *et al.* (2019) presented an interesting overview of machine learning applied to various problems in audio and acoustics. Nevertheless, only a short portion of each of these two reviews is dedicated to SSL with DNNs.

A. Aim of the paper

The goal of this paper is to fill this gap, and to provide a thorough survey of the SSL literature using DL techniques. More precisely, we examined and review 156 papers published from 2011 to 2021. We classify and discuss the different approaches in terms of characteristics of the employed methods and addressed configurations (e.g., single-source vs multi-source localization setup or neural network architecture; the exact list is given in Sec. I C). In other words, we present a taxonomy of the DL-based SSL literature published in the last decade. At the end of the paper, we present a summary of this survey in the form of four tables (one for the period 2011–2018, and one for each of the years 2019, 2020, and 2021). All of the methods that we reviewed are reported in these tables with a summary of their characteristics presented in different columns. This enables the reader to rapidly select the subset of methods having a given set of characteristics if they are interested in that particular type of method.

Note that in this survey paper, we do not aim to evaluate and compare the performance of the different systems. Due to the large number of DNN-based SSL papers and the diversity of configurations, such a contribution would be very difficult and cumbersome (albeit very useful), especially because the discussed systems are often trained and evaluated on different datasets. As we will see later, listing and commenting on these different datasets is, however, part of our survey effort. Note also that we do not consider SSL systems that exploit other modalities in addition to sound, e.g., audio-visual systems (Ban *et al.*, 2018; Masuyama *et al.*, 2020; Wu *et al.*, 2021c). Finally, we do consider DL-based methods for joint sound event localization and detection (SELD), which is a combination of sound event detection (SED; here detection actually means classification) and SSL, and in that case, we focus on the localization task. In particular, we include in the review the SELD methods presented to the DCASE Challenge (and/or to the corresponding DCASE Workshop) in 2019, 2020, and 2021 (see the DCASE Community, 2022). One of the tasks of this challenge is precisely dedicated to SELD, which has contributed to making the DL-based SSL (and SED) problem a popular research topic over the recent years.

B. General principle of DL-based SSL

The general principle of DL-based SSL methods and systems can be schematized with a simple pipeline, as illustrated in Fig. 1. A multichannel input signal recorded with a microphone array is processed by a feature extraction module to provide input features. These input features are fed into a DNN, which delivers an estimate of the source location or DoA. As discussed later in the paper, a recent trend is to skip the feature extraction module to directly feed the network with multichannel raw data. In any case, the two fundamental reasons behind the design of such SSL are the following.

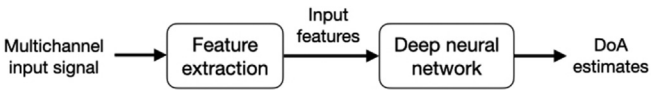

FIG. 1. General pipeline of a DL-based SSL system.

First, multichannel signals recorded with an array of I microphones distributed in space contain information about the location of the source(s). Indeed, when the microphones are close to each other compared to their distance to the source(s), the microphone signal waveforms, although appearing similar from a distance, exhibit more or less notable and complex differences in terms of delay and amplitude, depending on the experimental setup. These interchannel differences are due to distinct propagation paths from the source to the different microphones, for both the direct path (line of sight between source and microphone) and the numerous reflections that compose the reverberation in an indoor environment. In other words, a source signal $s_j(t)$ is convolved with different room impulse responses (RIRs) $a_{i,j}(t)$, which depend on the source position, microphone position and directivity (I denotes the microphone index in the array), and acoustic environment configuration (e.g., room shape):

$$\begin{aligned} x_i(t) &= a_{i,j}(t) \star s_j(t) + n_i(t) \\ &= \sum_{\tau=0}^{T-1} a_{i,j}(\tau) s_j(t-\tau) + n_i(t), \end{aligned} \tag{1}$$

where $x_i(t)$ denotes the resulting recorded signal at microphone i, $n_i(t)$ is the noise signal at microphone i (diffuse, "background" noise and possibly some sensor noise), and $\star$ denotes the convolution (note that we work with digital signals and t and τ are discrete time indexes; T is the effective length of the RIR). Therefore, the recorded signal contains information on the relative source-to-microphone array position. The microphone signals are often expressed in the time-frequency (TF) domain, using the short-term Fourier transform (STFT), where the convolution in Eq. (1) is assumed to transform into a product between the STFT of the source signal $S_j(f,n)$ and the acoustic transfer function (ATF) $A_{i,j}(f)$, which is the (discrete) Fourier transform of the corresponding RIR and is thus encoding the source spatial information (f denotes the frequency bin, and n is the STFT frame index) (Gannot *et al.*, 2017; Vincent *et al.*, 2018),

$$X_i(f,n) = A_{i,j}(f) S_j(f,n) + N_i(f,n). \tag{2}$$

When several, say J, sources are present, the recorded signal is the sum of their contribution (plus the noise),

https://doi.org/10.1121/10.0011809

$$x_i(t) = \sum_{j=1}^{J} a_{i,j}(t) \star s_j(t) + n_i(t). \quad (3)$$

This latter equation is often reformulated in the TF domain in matrix form,

$$\mathbf{X}(f,n) = \mathbf{A}(f)\mathbf{S}(f,n) + \mathbf{N}(f,n), \quad (4)$$

where $\mathbf{X}(f,n) = [X_1(f,n),\ldots,X_I(f,n)]^\top$ is the microphone signal vector, $\mathbf{A}(f)$ is the matrix gathering the ATFs, $\mathbf{S}(f,n) = [S_1(f,n),\ldots,S_J(f,n)]^\top$ is the source signal vector, and $\mathbf{N}(f,n) = [N_1(f,n),\ldots,N_I(f,n)]^\top$ is the noise vector. In that multi-source case, the difficulty of the SSL problem is that the contributions of the different sources generally overlap in time. SSL then requires to proceed to some kind of source clustering, which is generally easier to proceed in the frequency or TF domain due to the natural sparsity of audio sources in that domain (Rickard, 2002). In this paper, we do not describe the foundations of source-to-microphone propagation in more detail. They can be found in several references on general acoustics, e.g., (Jacobsen and Juhl, 2013; Rossing, 2007), room acoustics, e.g., (Kuttruff, 2016), array signal processing (e.g., Benesty *et al.*, 2008; Brandstein and Ward, 2001; Jarrett *et al.*, 2017; Rafaely, 2019), speech enhancement and audio source separation (e.g., Gannot *et al.*, 2017; Vincent *et al.*, 2018), and many papers on conventional SSL.

The second reason for designing DNN-based SSL systems is that even if the relationship between the information contained in the multichannel signal and the location of the source(s) is generally complex [especially in a multisource reverberant and noisy configuration, see Eqs. (3) and (4)], DNNs are powerful models that are able to automatically identify and exploit this relationship, given that they are provided with a sufficiently large number of representative training examples. This ability of data-driven DL methods to replace conventional methods based on a signal/channel model and SP techniques—or at least a part of them, since the feature extractor module can be based on conventional processing—makes them attractive for addressing problems such as SSL. While some conventional methods can adapt to the observed signals (e.g., Dvorkind and Gannot, 2005; Laufer-Goldshtein *et al.*, 2020; Li *et al.*, 2016a,b), they are all intrinsically based on certain (more or less plausible) modeling assumptions, which can limit their effectiveness when exposed to the complexity of real-world acoustics. Deep learning models do not *explicitly* impose any such assumptions, and instead they efficiently adapt to the presented training data. This is, however, also the major drawback of the DNN-based approaches, as they are less generic than traditional methods. A deep model designed for and trained in a given configuration (e.g., a given microphone array geometry) will not provide satisfying localization results if the setup changes (Le Moing *et al.*, 2021; Liu *et al.*, 2018), unless some relevant adaptation method can be used, which is still an open problem in DL in general.

C. Outline of the paper

The remainder of the paper is organized as follows. In Sec. II, we specify the context and scope of the survey in terms of the considered acoustic environment and sound source configurations. In Sec. III, we briefly present the most common conventional SSL methods, for two reasons: first, they are often used as a baseline for the evaluation of DL-based methods; and second, we will see that several types of features extracted by conventional methods can be used in DL-based methods. Section IV aims to classify the different neural network architectures used for SSL. Section V presents the various types of input features used for SSL with neural networks. In Sec. VI, we explain the two output strategies employed in DL-based SSL: classification and regression. We then discuss in Sec. VII the datasets used for training and evaluating the models. In Sec. VIII, learning paradigms such as supervised or semi-supervised learning are discussed from the SSL perspective. Section IX provides the four summary tables and concludes the paper. Note that, due to the large number of acronyms used in this survey paper, we provide a list of these acronyms in Table I.

II. ACOUSTIC ENVIRONMENT AND SOUND SOURCE CONFIGURATIONS

SSL has been applied in different configurations, depending on the application. In this section, we specify the scope of our survey in terms of acoustic environment (noisy, reverberant, or even multi-room) and the nature of the considered sound sources (their type, number, and static/mobile status).

A. Acoustic environments

In this paper, we focus on SSL in an indoor environment, i.e., when the microphone array and the sound source(s) are present in a closed room, generally of moderate size, typically an office room or a domestic environment. This implies reverberation: in addition to the direct source-to-microphone propagation path, the recorded sound contains many other multi-path components of the same source. All of these components form the RIR, which is defined for each source position and microphone array position (including orientation) and for a given room configuration.

In a general manner, the presence of reverberation is seen as a notable perturbation that makes SSL more difficult compared to the simpler (but somewhat unrealistic) *anechoic* case, which assumes the absence of reverberation, as is obtained in the *free field* propagation setup. Another important adverse factor to take into account in SSL is noise. On the one hand, noise can come from interfering sound sources in the surrounding environment: TV, background music, pets, street noise passing through open or closed windows, etc. Often, noise is considered diffuse, i.e., it does not originate from a clear direction. On the other hand, the imperfections of the recording devices are another source of noise that are generally considered artifacts.

TABLE I. Table of acronyms.

ACCDOA	activity-coupled Cartesian direction of arrival
AE	autoencoder
ATF	acoustic transfer function
ASR	automatic speech recognition
BGRU	bidirectional gated recurrent unit
BIR	binaural impulse response
BRIR	binaural room impulse response
CC	cross correlation
CNN	convolutional neural network
CRNN	convolutional recurrent neural network
CPS	cross power spectrum
DCASE	Detection and Classification of Acoustic Scenes and Events
DIRHA	distant-speech interaction for robust home applications
DL	deep learning
DNN	deep neural network
DoA	direction of arrival
DP-RTF	direct-path relative transfer function
DRR	direct-to-reverberant ratio
EM	expectation maximization
ESPRIT	Estimation of Signal Parameters via Rotational Invariance Techniques
EVD	eigenvalue decomposition
FFNN	feed-forward neural network
FOA	first-order Ambisonics
GAN	generative adversarial network
GCC	generalized cross correlation
GLU	gated linear unit
GMM	Gaussian mixture models
GMR	Gaussian mixture regression
GPU	graphical processing unit
GRU	gated recurrent unit
HATS	head-and-torso simulator
HOA	higher-order Ambisonics
HRTF	head-related transfer function
ICA	independent component analysis
ILD	interaural level difference
IPD	interaural phase difference
ITD	interaural time difference
ISM	image source method
LSTM	long short-term memory
MHSA	multi-head self-attention
MLP	Multi-Layer Perceptron
MOT	multi-object tracking
MUSIC	MUltiple SIgnal Classification
NLP	natural language processing
NoS	number of sources
PHAT	PHAse Transform
RIR	room impulse response
RNN	recurrent neural network
RTF	relative transfer function
SA	self-attention
SCM	spatial covariance matrix
SED	sound event detection
SELD	sound event localization and detection
SH	spherical harmonics
SMIR	spherical microphone impulse response
SMN	sequence matching network
SP	signal processing
SPS	spatial pseudo-spectrum
SRP	steered power response
SSL	sound source localization
STFT	short-term Fourier transform
TCN	temporal convolutional network
TDoA	time difference of arrival
TF	time-frequency
VAD	voice activity detection
VAE	variational autoencoder
WDO	W-disjoint orthogonality

Early works on using neural networks for DoA estimation most often considered direct-path propagation only (the anechoic setting) (e.g., El Zooghby *et al.*, 2000; Falong *et al.*, 1993; Goryn and Kaveh, 1988; Jha *et al.*, 1988; Jha and Durrani, 1989, 1991; Rastogi *et al.*, 1987; Southall *et al.*, 1995; Yang *et al.*, 1994), though a model of the acoustical environment was used to generate simulated data to train the neural network of Datum *et al.* (1996). Most of these works are from the pre-deep-learning era, using "shallow" neural networks with only one or two hidden layers (Goodfellow *et al.*, 2016). We do not detail these works in our survey, although we acknowledge them as pioneering contributions to the neural network-based DoA estimation problem. A few more recent works based on more "modern" neural network architectures also focused on anechoic propagation only or did not consider sound sources in the audible bandwidth (Bialer *et al.*, 2019; Choi and Chang, 2020; Elbir, 2020; Liu *et al.*, 2018; Ünleren and Yaldiz, 2016).

B. Source types

In the SSL literature, a great proportion of systems focuses on localizing speech sources because of their importance in related tasks such as speech enhancement or speech recognition. Examples of speaker localization systems can be found in papers by Chakrabarty and Habets (2019b); Grumiaux *et al.* (2021b); Hao *et al.* (2020); He *et al.* (2021a). In such systems, the neural networks are trained to estimate the DoA of speech sources so that they are somewhat specialized in this type of source. Other systems, in particular those participating in the DCASE Challenge, consider a variety of sound source types (Politis *et al.*, 2020b). Depending on the challenge task and its corresponding dataset, these methods are capable of localizing alarms, crying babies, crashes, barking dogs, female/male screams, female/male speech, footsteps, knockings on doors, ringings, phones, and piano sounds. Note that the localization of such sources, even if they overlap in time, is not necessarily a more difficult problem than the localization of several overlapping speakers, since the former usually have distinct spectral characteristics that neural models may exploit for better detection and localization.

https://doi.org/10.1121/10.0011809

C. Number of sources

The number of sources (NoS) in a recorded mixture signal is an important parameter for SSL. In the SSL literature, the NoS might be considered as known (as a working hypothesis). Alternatively, it can be estimated along with the source location, in which case the SSL problem is a combination of detection and localization. Examples of conventional (non-deep) SSL works including NoS estimation can be found in papers by Arberet *et al.* (2009) and Landschoot and Xiang (2019).

Many DNN-based works have considered only one source to localize, as it is the simplest scenario to address (e.g., Bologni *et al.*, 2021; Liu *et al.*, 2021; Perotin *et al.*, 2018b). We refer to this scenario as *single-source* SSL. In this case, the networks are trained and evaluated on datasets with only at most one active source (a source is said to be active when emitting sound and inactive otherwise). In terms of NoS, we thus have here either 1 or 0 active source. The activity of the source in the processed signal, which generally contains background noise, can be artificially controlled, i.e., the knowledge of source activity is a working hypothesis. This is a reasonable approach at training time when using synthetic data, but it is quite unrealistic at test time on real-world data. Alternatively, the source activity can be estimated, which is a more realistic approach at test time. In the latter case, there are two ways of dealing with the source activity detection problem. The first is to employ a source detection algorithm beforehand and then apply the SSL method only on the signal portions with an active source. For example, a voice activity detection (VAD) technique has been used in the SSL systems of Chang *et al.* (2018); Kim and Hahn (2018); Li *et al.* (2016c); Sehgal and Kehtarnavaz (2018). The other way is to detect the activity of the source at the same time as the localization algorithm. For example, an additional neuron was added by Yalta *et al.* (2017) to the output layer of their DNN, which outputted 1 when no source was active (in that case, all other localization neurons were trained to output 0), and 0 otherwise.

Multi-source localization is a much more difficult problem than single-source SSL. Current state-of-the-art DL-based methods address multi-source SSL in adverse environments. In this survey, we consider multi-source localization the scenario in which several sources overlap in time (i.e., they are simultaneously emitting), regardless of their type (e.g., there could be several speakers or several distinct sound events). The specific case of a multi-speaker conversation with or without speech overlap is strongly connected to the *speaker diarization* problem ("who speaks when?") (Anguera *et al.*, 2012; Park *et al.*, 2021b; Tranter and Reynolds, 2006). Speaker localization, diarization, and (speech) source separation are intrinsically connected problems, as the information retrieved from solving each of them can be useful for addressing the others (Jenrungrot *et al.*, 2020; Kounades-Bastian *et al.*, 2017; Vincent *et al.*, 2018). An investigation of these connections is beyond the scope of this survey.

In the multi-source scenario, the source detection problem transposes to a source counting problem, but the same considerations as in the single-source scenario hold. In some works, the knowledge of the NoS is a working hypothesis (e.g., Bohlender *et al.*, 2021; Fahim *et al.*, 2020; Grumiaux *et al.*, 2021a; Grumiaux *et al.*, 2021b; He *et al.*, 2019a; Ma *et al.*, 2015; Perotin *et al.*, 2019b) and the sources' DoA can be directly estimated. If the NoS is unknown, one can apply a source counting system beforehand, e.g., with a dedicated DNN (Grumiaux *et al.*, 2020). For example, Tian (2020) trained a separate neural network to estimate the NoS in the recorded mixture signal, after which he used this information along with the output of the DoA estimation neural network. Alternatively, the NoS can be estimated alongside the DoAs, as in the single-source scenario, based on the SSL network output. When using a classification paradigm, the network output generally predicts the probability of the presence of a source within each discretized region of the space (see Sec. VIII). One can thus set a threshold on this estimated probability, which implicitly provides source counting.[1] Otherwise, the ground-truth or estimated NoS is typically used to select the corresponding number of classes having the highest probability.

Finally, several DNN-based systems were purposefully designed to estimate the NoS alongside the DoAs. For example, the method proposed by Nguyen *et al.* (2020a) uses a neural architecture with two output branches: the first branch is used to estimate the NoS (up to four sources; the problem is formulated as a classification task), while the second branch is used to classify the azimuth into several regions. In the same spirit, we can mention the numerous systems presented at the DCASE Challenge, in which the SED task, jointly conducted with SSL, intrinsically provides an estimate of the NoS. Note that many DCASE Challenge candidate systems will be reviewed at the core of this survey.

D. Moving sources

Source tracking is the problem of estimating the evolution of the sources' position(s) over time, especially when the sources are mobile. In this survey paper, we do not address the problem of tracking on its own, which is usually done in a separate algorithm using the sequence of DoA estimates obtained by applying SSL on successive time windows (Vo *et al.*, 2015). Still, several DL-based SSL systems have been shown to produce more accurate localization of moving sources when they were trained on a dataset that includes this type of source (Adavanne *et al.*, 2019b; Diaz-Guerra *et al.*, 2021b; Guirguis *et al.*, 2020; He *et al.*, 2021b). In other cases, as the number of real-world datasets with moving sources is limited and the simulation of signals with moving sources is cumbersome, a number of systems trained on static sources have been shown to retain fair to good performance for moving sources, e.g., (Grumiaux *et al.*, 2021a; Opochinsky *et al.*, 2021; Sundar *et al.*, 2020).

III. CONVENTIONAL SSL METHODS

Before the advent of DL, a set of signal processing techniques were developed to address SSL. A detailed review of these techniques was made by DiBiase *et al.* (2001). A review in the specific robotics context was made by Argentieri *et al.* (2015). In this section, we briefly present the most common conventional SSL methods. As briefly stated in the introduction, the reason for this is twofold: first, conventional SSL methods are often used as baselines for DL-based methods; and second, many DL-based SSL methods use input features extracted with conventional methods (see Sec. V).

When the geometry of the microphone array is known, DoA estimation can be performed by estimating the time-difference of arrival (TDoA) of the sources between the microphones (Xu *et al.*, 2013). The generalized cross correlation (CC) with phase transform (GCC-PHAT) is one of the most employed method when dealing with a 2-microphone array (Knapp and Carter, 1976). It is computed as the inverse Fourier transform of a weighted version of the cross-power spectrum (CPS) between the signals of the two microphones:

$$r_{1,2}(\tau) = \sum_{f=0}^{F-1} \frac{X_1(f)X_2(f)^*}{|X_1(f)X_2(f)^*|} e^{j2\pi(f\tau/N)}, \tag{5}$$

where $X_i(f)$ are the N-point Fourier transform of the microphone signals $x_i(t)$, and $X_1(f)X_2(f)^*$ is the CPS (* denotes the complex conjugate). The TDoA estimate is then obtained by finding the time delay between the microphone signals that maximizes the GCC-PHAT function,

$$\hat{\tau} = \arg\max_{\tau} r_{1,2}(\tau). \tag{6}$$

The GCC approach has been extended to arrays with more than two microphones, showing in particular that the localization could be improved by taking advantage of the multiple microphone pairs (Benesty *et al.*, 2008; DiBiase *et al.*, 2001).

Building an acoustic power map $P(\mathbf{x})$, with $\mathbf{x}$ the spatial coordinates, usually a regular grid, is another way to retrieve the DoA of one or multiple sources, as local maxima of this map mainly correspond to the sources' DoA. The Steered-Response Power (SRP) map has been extensively used: it consists in pointing delay and sum beamformers towards each of the candidate grid positions and measuring the energy that arises from these directions. Its PHAT version, which reveals more robust to reverberation, is certainly the most popular. Practically, it can be derived from the average of the GCC-PHAT computed on all microphone pairs (DiBiase *et al.*, 2001),

$$P(\mathbf{x}) = \sum_{m_1=1}^{M} \sum_{m_2=m_1+1}^{M} r_{1,2}(\tau_{m_1,m_2}(\mathbf{x})), \tag{7}$$

where $\tau_{m_1,m_2}(\mathbf{x})$ is the delay between the microphones m_1 and m_2 associated with the spatial position $\mathbf{x}$.

An alternative to building the SRP-based acoustic map—which happens to be computationally expensive as it usually amounts to a grid search—is localization by exploiting the sound intensity. The use of sound intensity for source localization has a long history (e.g., Basten *et al.*, 2008; Hickling *et al.*, 1993; Jarrett *et al.*, 2010; Nehorai and Paldi, 1994; Raangs and Druyvesteyn, 2002; Tervo, 2009). In favorable acoustic conditions, sound intensity is parallel to the direction of the propagating sound wave (see Sec. V E), and hence the DoA can be efficiently estimated. Unfortunately, its accuracy quickly degrades in the presence of acoustic reflections (Daniel and Kitić, 2020).

Subspace methods are another classical family of localization algorithms. These methods rely on the computation of the (time-averaged) CPS matrix $\mathbf{R}(f)$ defined by

$$\mathbf{R}(f) = \sum_{n=1}^{N} \mathbf{X}(f,n)\mathbf{X}(f,n)^H, \tag{8}$$

where $\mathbf{X}(f,n)$ is the STFT (or more generally a local discrete Fourier transform) of the multichannel signal vector defined in Eq. (4) (^{H} denotes the Hermitian operator), and its eigenvalue decomposition (EVD). Assuming that the target source signals and noise are uncorrelated, the multiple signal classification (MUSIC) method (Schmidt, 1986) applies EVD to estimate the signal and noise subspaces. After Eq. (4), the signal subspace bases are assumed to correspond to the columns of the mixing matrix $\mathbf{A}(f)$, which are the multichannel ATFs of the sources (often referred to as *steering vectors* in this context). The signal or noise subspace bases are then used to probe a given direction for the presence of a source, i.e., apply *spatial filtering* or *beamforming* (Benesty *et al.*, 2008; Van Veen and Buckley, 1988). This time-demanding search can be relaxed using the Estimation of Signal Parameters *via* Rotational Invariance Technique (ESPRIT) algorithm (Roy and Kailath, 1989), which exploits the structure of the source subspace to directly infer the source DoA. However, this often comes at the cost of producing less accurate predictions than MUSIC (Mabande *et al.*, 2011). MUSIC and ESPRIT assume narrowband signals, although wideband extensions have been proposed (e.g., Dmochowski *et al.*, 2007; Hogg *et al.*, 2021). Subspace methods are robust to noise and can produce highly accurate estimates, but they are sensitive to reverberation.

Methods based on probabilistic generative mixture models have been proposed by, e.g., Dorfan and Gannot (2015); Li *et al.* (2017); Mandel *et al.* (2009); May *et al.* (2011); Roman and Wang (2008); Schwartz and Gannot (2013); Woodruff and Wang (2012). Typically, the models are variants of Gaussian mixture models (GMMs), with one Gaussian component per source to be localized or per candidate source position. In very few papers (e.g., May *et al.*, 2011), the model is trained offline with a dedicated training dataset. But most often, the model parameters are directly estimated "at test time," that is using the multichannel signal containing the sources to localize. This is done by maximizing the data likelihood function with histogram-based or

https://doi.org/10.1121/10.0011809

expectation-maximization (EM) algorithms exploiting the sparsity of sound sources in the TF domain (Rickard, 2002), which can be computationally intensive. A GMM variant functioning directly in regression mode, i.e., a form of Gaussian mixture regression (GMR), was proposed for single-source localization by Deleforge and Horaud (2012) and later extended to multi-source localization (and possibly separation) (Deleforge *et al.*, 2013, 2015). The GMR is locally linear but globally non-linear and the estimation of the model parameters is done offline on training data. Hence the spirit is close to DNN-based SSL. White noise signals convolved with synthetic RIRs were used for training. The method was shown to generalize well to speech signals, which are sparser than noise in the TF domain, thanks to the use of a latent variable modeling the signal activity in each TF bin.

Mixture models are strongly connected to Bayesian inference, which considers the posterior distribution of model parameters given the observed data (hence involving both the likelihood function and a prior distribution of the model parameters). Escolano *et al.* (2014) considered applying Bayesian inference on a Laplacian source mixture model, using GCC-PHAT features in a two-microphone array set-up. Interestingly, they used two levels of Bayesian inference: one for the estimation of the NoS (which is an hyper-parameter of the model), using Bayesian model selection, and one for the estimation of the model parameters (and thus the corresponding source DoAs), using posterior distribution evaluation. In this work, the evaluation of the involved distributions was done with sampling techniques, e.g., Markov Chain Monte Carlo (MCMC) methods. The same methodology was further applied by Bush and Xiang (2018) with a coprime array consisting of two superimposed, spatially undersampled, uniform linear arrays (Vaidyanathan and Pal, 2010), and by Landschoot and Xiang (2019) in the spherical harmonics (SH) domain using a spherical microphone array (see Sec. V).

Compressive sensing and sparse recovery methods are widely used in acoustics for different purposes (Gerstoft *et al.*, 2018; Xenaki *et al.*, 2014), including SSL (Yang *et al.*, 2018). The main premise is that many high-dimensional signals admit a low-dimensional representation, which can be viewed through, e.g., *sparse synthesis* (Candes *et al.*, 2006) or *sparse analysis* (Nam *et al.*, 2013) model. Concerning the SSL problem, the sparsity assumption is usually assumed in the spatial (or spatial beam) domain (e.g., Chardon and Daudet, 2012; Fortunati *et al.*, 2014; Gerstoft *et al.*, 2016; Kitić *et al.*, 2014; Noohi *et al.*, 2013), and the resulting problem is addressed by convex optimization, greedy or Bayesian methods (e.g., Foucart and Rauhut, 2013; Gerstoft *et al.*, 2018). This concept has led to prominent localization methods achieving remarkable performance. Nonetheless, despite their strong theoretical guarantees, compressive sensing methods suffer from two drawbacks. For one, it is usually required that the sources coincide with points of some pre-defined grid, although grid-free methods have been proposed in some specific cases (e.g., Xenaki and Gerstoft, 2015; Yang and Xie, 2015). The second issue is shared with other conventional methods, i.e., the strong modeling assumptions reflected in, for example, the known structure of the (sub-Gaussian) dictionary matrix. Dictionary learning techniques have been proposed to alleviate the latter problem to some extent (e.g., Hahmann *et al.*, 2021b; Wang *et al.*, 2018; Zea and Laudato, 2021). Sparse Bayesian learning (SBL) is a combination of the Bayesian framework with the principles of sparse representations and compressed sensing. It usually involves using sparse arrays such as the coprime array mentioned previously and nested arrays (Pal and Vaidyanathan, 2010). SBL has been used for SSL by, e.g., Gerstoft *et al.* (2016); Liu *et al.* (2012); Nannuru *et al.* (2018); Ping *et al.* (2020); Xenaki *et al.* (2018); Zhang *et al.* (2014).

Finally, ICA is a class of algorithms aimed at retrieving the different source signals comprising a mixture by assuming and exploiting their mutual statistical independence. ICA has most often been used in audio processing for blind source separation, but it has also proven to be useful for multi-source SSL (Sawada *et al.*, 2003). As briefly stated before, in the multi-source scenario, SSL is closely related to the source separation problem, since localization can help separation, and separation can help localization (Gannot *et al.*, 2017; Vincent *et al.*, 2018).

IV. NEURAL NETWORK ARCHITECTURES FOR SSL

In this section, we discuss the neural network architectures that have been proposed in the literature to address the SSL problem. However, we do not present the basics of these neural networks since they have been extensively described in the general DL literature (e.g., Chollet, 2017; Goodfellow *et al.*, 2016; LeCun *et al.*, 2015). The design of DNNs for a given application often requires investigating (and possibly combining) different architectures and tuning their hyperparameters. This was the case for SSL over the last decade, and the evolution of DL-based SSL techniques has followed the general evolution of DNNs toward more and more complex architectures or new efficient models adopted by the DL and SP communities at large, i.e., largely beyond the SSL problem (e.g., attention models). In other words, the DNN architectures used in SSL are often inherited from other works in other (connected or more distant) domains, simply because they were shown to work well on audio signals or other types or signals. In the same spirit, different models are often combined (in parallel and/or sequentially).

We have thus organized the presentation according to the type of layers used in the networks, with a progressive and "inclusive" approach in terms of complexity: a network within a given category can contain layers from another previously presented category. We thus first present systems based on feedforward neural networks (FFNNs). We then focus on CNNs and recurrent neural networks (RNNs), which generally incorporate some feedforward layers. Next, we review architectures combining CNNs with RNNs,

namely, convolutional recurrent neural networks (CRNNs). Then, we focus on neural networks with residual connections and with attention mechanisms. Finally, we present SSL systems with an encoder-decoder architecture.

A. FFNNs

The FFNN was the first and simplest type of artificial neural network to be designed. In such a network, data move in one direction from the input layer to the output layer, possibly *via* a series of hidden layers (Goodfellow *et al.*, 2016; LeCun *et al.*, 2015). Non-linear activation functions are usually used after each layer (possibly except for the output layer). While this definition of FFNN is very general and may include architectures such as CNNs (discussed in the next subsection), here we mainly focus on architectures made of fully-connected layers known as Perceptron and Multi-Layer Perceptron (MLP) (Goodfellow *et al.*, 2016; LeCun *et al.*, 2015). A Perceptron has no hidden layer, while the notion of MLP is a bit ambiguous: some authors state that an MLP has one hidden layer, while others allow more hidden layers. In this paper, we call an MLP an FFNN with one or more hidden layers.

A few pioneering SSL methods using shallow neural networks (Perceptron or 1-hidden layer MLP) and applied in "unrealistic" setups (e.g., assuming direct-path sound propagation only) have been briefly mentioned in Sec. II A. One of the first uses of an MLP for SSL was proposed by Kim and Ling (2011), who actually considered several MLPs. One network estimates the NoS, after which a distinct network is used for SSL for each considered NoS. The authors evaluated their method on reverberant data even though they assumed an anechoic setting. Tsuzuki *et al.* (2013) proposed using a complex-valued MLP in order to process complex two-microphone-based features, which led to better results than using a real-valued MLP. Youssef *et al.* (2013) also used an MLP to estimate the azimuth of a sound source from a binaural recording made with a robot head. The interaural time difference (ITD) and the interaural level difference (ILD) values (see Sec. V) were separately fed into the input layer and were each processed by a specific set of neurons. A single-hidden-layer MLP was used by Xiao *et al.* (2015), taking GCC-PHAT-based features as inputs and tackling SSL as a classification problem (see Sec. VIII), which showed an improvement over conventional methods on simulated and real data. A similar approach was proposed by Vesperini *et al.* (2016), but the localization was done by regression in the horizontal plane.

Naturally, MLPs with deeper architecture (i.e., more hidden layers) have also been investigated for SSL. Roden *et al.* (2015) compared the performance of an MLP with two hidden layers and different input types, the number of hidden neurons being linked to the type of input features (see Sec. V for more details). Yiwere and Rhee (2017) used an MLP with three hidden layers (tested with different numbers of neurons) to output source azimuth and distance estimates. An MLP with four hidden layers was tested by He *et al.* (2018a) for multi-source localization and speech/non-speech classification, showing similar results as a 4-layer CNN (see Sec. IV B).

Ma *et al.* (2015) proposed using a different MLP for different frequency sub-bands, with each MLP having eight hidden layers. This idea is based on the assumption that, in the presence of multiple sources, each frequency band is mostly dominated by a single source, which enables the training to be done exclusively on single-source data. The output of each sub-band MLP corresponds to a probability distribution on azimuth regions, and the final azimuth estimations are obtained by integrating the probability values over the frequency bands. Another system in the same vein was proposed by Takeda *et al.* in several papers (Takeda and Komatani, 2016a,b, 2017; Takeda *et al.*, 2018). In these works, the eigenvectors of the recorded signal interchannel correlation matrix were separately fed per frequency band into parallel branches of the network, particularly into specific fully-connected layers. Then, several additional fully-connected layers progressively integrated the frequency-dependent outputs (see Fig. 2). The authors showed that this specific architecture outperforms a more conventional 7-layer MLP and the classical MUSIC algorithm on anechoic and reverberant single- and multi-source signals. Opochinsky *et al.* (2019) proposed a small 3-layer MLP to estimate the azimuth of a single source using the relative transfer function (RTF, see Sec. V A 1) of the signal. Their approach is weakly supervised since one part of the loss function is computed without the ground truth DoA labels (see Sec. VIII).

An indirect use of an MLP was explored by Pak and Shin (2019), who used a 3-layer MLP to enhance the interaural phase difference (IPD) (see Sec. V) of the input signal, which was then used for DoA estimation.

B. Convolutional neural networks

CNNs are a popular class of DNNs widely used for pattern recognition due to their property of being translation equivariant (Cohen *et al.*, 2019; Goodfellow *et al.*, 2016). They have been successfully applied to various tasks, such as image classification (e.g., Krizhevsky *et al.*, 2017), natural language processing (NLP) (e.g., Kim, 2014), or automatic speech recognition (e.g., Waibel *et al.*, 1989). CNNs have also been used for SSL, as detailed below.

To our knowledge, Hirvonen (2015) was the first to use a CNN for SSL. He employed this architecture to classify an audio signal containing one speech or musical source into one of eight spatial regions (see Fig. 3). This CNN is composed of four convolutional layers to extract feature maps from multichannel magnitude spectrograms (see Sec. V), followed by four fully-connected layers for classification. Classical pooling is not used because, according to the author, it does not seem relevant for audio representations. Instead, a 4-tap stride with a 2-tap overlap is used to reduce the number of parameters. This approach shows good performance on single-source signals and is capable of adapting

https://doi.org/10.1121/10.0011809

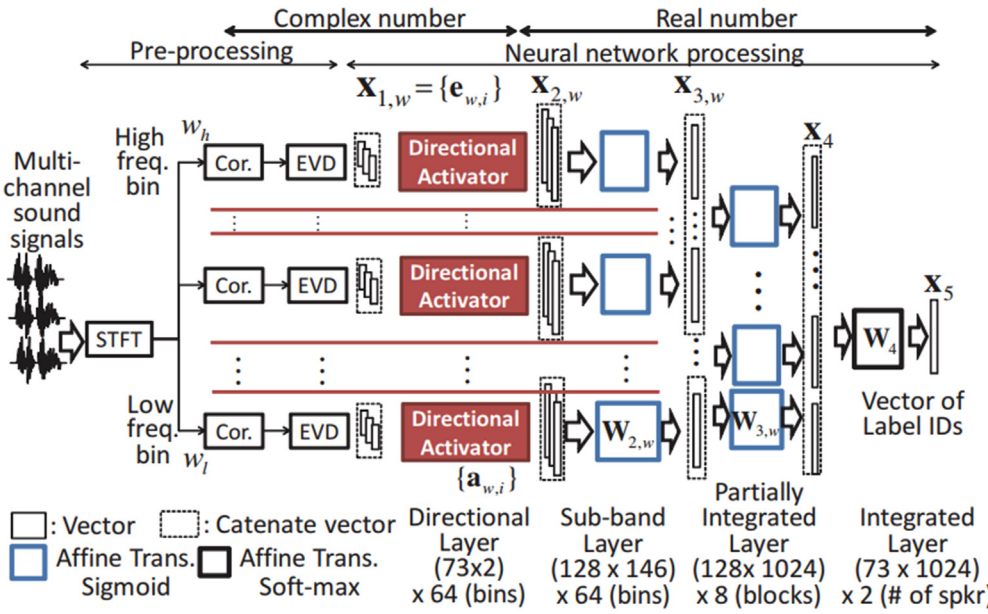

FIG. 2. (Color online) The MLP architecture used by Takeda *et al*. in several papers (Takeda and Komatani, 2016a,b, 2017; Takeda *et al*., 2018). Multiple subband feedforward layers, indexed by w, are trained to extract features from the CPS matrix eigenvectors $\mathbf{e}_{w,i}$, which are used as directional activation functions. The obtained subband vectors $\mathbf{X}_{2,w}$ are integrated across subbands progressively *via* other feedforward layers, giving $\mathbf{X}_{3,w}$ and then $\mathbf{X}_4$. The output layer finally classifies its input in one of the candidate DoAs (the entries of the vector $\mathbf{X}_5$). Note: Reprinted from Takeda and Komatani (2016a). Copyright by IEEE; reprinted with permission.

to different configurations without hand-engineering. However, two topical issues of such a system were pointed out by the author: the robustness of the network with respect to a shift in source location, and the difficulty of interpreting the hidden features.

Chakrabarty and Habets also designed a CNN to predict the azimuth of one (Chakrabarty and Habets, 2017a) or two (Chakrabarty and Habets, 2017b, 2019b) speakers in reverberant environments. The input features are the multichannel STFT phase spectrograms (see Sec. V). In Chakrabarty and Habets (2017a), they proposed using three successive convolutional layers with 64 filters of size 2×2 to consider neighboring frequency bands and microphones. In Chakrabarty and Habets (2017b), they reduced the filter size to 2×1 (1 in the frequency axis) because of the W-disjoint orthogonality (WDO) assumption for speech signals, which assumes that several speakers are not simultaneously active in a same TF bin (Rickard, 2002). In Chakrabarty and Habets (2019b),

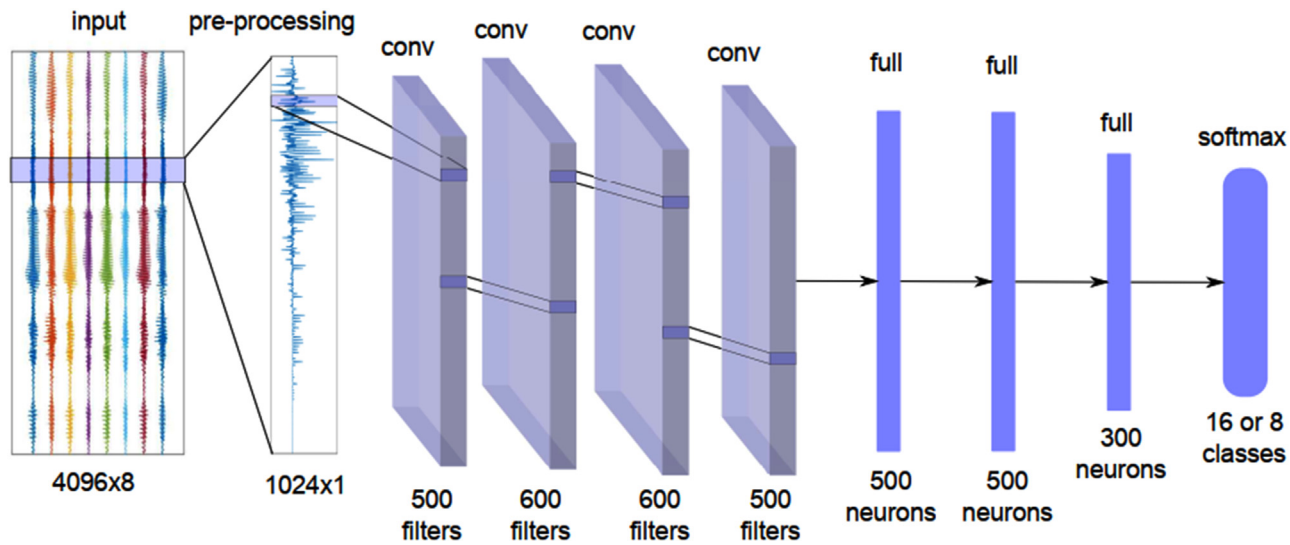

FIG. 3. (Color online) The CNN architecture proposed by Hirvonen (2015) for SSL. The input is an 8-channel signal. For each short-term frame, the 8 magnitude spectra (of 128 frequency bins) are concatenated to form a 1024×1 tensor, which is fed into a series of four convolutional layers with 500 or 600 learnable kernels. The extracted features then pass through several feedforward layers containing 500 or 300 neurons. The output layer contains eight neurons (or 16 if the source type is also considered) and estimates the probability of a source being present in eight candidate DoAs using a softmax activation function. Note: Reprinted from Hirvonen (2015); copyright by the author; reprinted with permission.

they demonstrated that for an M-microphone array, the optimal number of convolutional layers for exploiting phase correlations between the neighboring microphones is $M - 1$.

He *et al.* (2018a) compared a 4-layer MLP and a 4-layer CNN for the multi-speaker detection and localization task. The results showed similar accuracy for both architectures. A deeper architecture was proposed by Yalta *et al.* (2017), with 11 to 20 convolutional layers depending on the experiments. These deeper CNNs showed robustness against noise compared to MUSIC, as well as smaller training time, but this was partly due to the presence of residual blocks (see Sec. IV E). A similar architecture was presented by He *et al.* (2018b), with many convolutional layers and some residual blocks, although with a specific multi-task configuration. The end of the network was split into two convolutional branches, one for azimuth estimation, and the other for speech/non-speech signal classification.

While most localization systems aim to estimate the azimuth or both the azimuth and elevation, Thuillier *et al.* (2018) investigated the estimation of only the elevation angle using a CNN with binaural input features: the ipsilateral and contralateral head-related transfer function (HRTF) magnitude responses (see Sec. V). Vera-Diaz *et al.* (2018) chose to apply a CNN directly on raw multichannel waveforms, assembled side by side as an image, to predict the Cartesian coordinates (x, y, z) of a single static or moving speaker. The successive convolutional layers contain around a hundred filters from size 7×7 for the first layers to 3×3 for the last layer. Ma and Liu (2018) also used a CNN to perform regression, but they used the CPS matrix as an input feature (see Sec. V). To estimate both the azimuth and elevation, Nguyen *et al.* (2018) used a relatively small CNN (two convolutional layers) in regression mode, with binaural input features. A similar approach was considered by Sivasankaran *et al.* (2018) for speaker localization based on a CNN. They showed that injecting a speaker identifier, particularly a mask estimated for the speaker uttering a given keyword, alongside the binaural features at the input layer improved the DoA estimation.

A joint VAD and DoA estimation CNN was developed by Vecchiotti *et al.* (2018). They showed that both problems can be handled jointly in a multi-room environment using the same architecture, although considering separate input features (GCC-PHAT and log-mel-spectrograms) in two separate input branches. These branches are then concatenated in a further layer. Vecchiotti *et al.* (2019b) extended this work by exploring several variant architectures and experimental configurations, and Vecchiotti *et al.* (2019a) developed an end-to-end auditory-inspired system based on a CNN, with Gammatone filter layers included in the neural architecture. A method based on mask estimation was proposed by Zhang *et al.* (2019b), in which a TF mask was estimated and used to either clean or be appended to the input features, facilitating the DoA estimation by a CNN.

Nguyen *et al.* (2020a) presented a multi-task CNN containing ten convolutional layers with average pooling, inferring both the NoS and the sources' DoA. They evaluated their network on signals with up to four sources, showing very good performance in both simulated and real environments. A small 3-layer CNN was employed by Varanasi *et al.* (2020) to infer both azimuth and elevation using signals decomposed with third-order SH (see Sec. V). The authors tried several combinations of input features, including using only the magnitude and/or the phase of the spherical harmonic decomposition.

In the context of hearing aids, a CNN was applied to both VAD and DoA estimation by Varzandeh *et al.* (2020). This system is based on two input features, GCC-PHAT and periodicity degree, both fed separately into two convolutional branches. These two branches are then concatenated in a further layer, which is followed by feed-forward layers. Fahim *et al.* (2020) applied an 8-layer CNN to the so-called modal coherence of first-order Ambisonics input features (see Sec. V) for the localization of multiple sources in a reverberant environment. They proposed a new method to train a multi-source DoA estimation network with only single-source training data, showing an improvement over the system of Chakrabarty and Habets (2019b), especially for signals with three speakers. Hao *et al.* (2020) investigated a real-time implementation of SSL using a CNN with a relatively small architecture (three convolutional layers).

Krause *et al.* (2020a) investigated the use of several types of convolution. They reported that networks using three-dimensional (3D) convolutions (on the time, frequency, and channel axes) achieved better localization accuracy compared to those based on two-dimensional (2D) convolutions, complex convolutions, and depth-wise separable convolutions (all of them on the time and frequency axes), but with a high computational cost. They also showed that the use of depth-wise separable convolutions leads to a good trade-off between accuracy and model complexity (to our knowledge, they were the first to explore this type of convolutions).

Bologni *et al.* (2021) proposed a neural network architecture including a set of 2D convolutional layers for frame-wise feature extraction, followed by several one-dimensional (1D) convolutional layers in the time dimension for temporal aggregation. Diaz-Guerra *et al.* (2021b) applied 3D convolutional layers on SRP-PHAT power maps computed for both azimuth and elevation estimation. They also used a couple of 1D causal convolutional layers at the end of the network to perform single-source tracking. Their whole architecture was designed to function in fully causal mode so that it can be adapted for real-time applications. Wu *et al.* (2021a) proposed using a supervised image mapping approach inspired from computer vision works and referred to as *image translation*. They used a CNN (completed with residual layers, see Sec. IV E) to map an input 2D image [DoA features extracted by conventional beamforming and reshaped as a function of Cartesian coordinates (x, y)] into an output 2D image of the target source position (in which the pixel intensity is decreasing rapidly with the distance to the source), from which the source location is obtained.

https://doi.org/10.1121/10.0011809

As mentioned in the introduction, the DCASE Challenge includes a SELD task, and CNNs have also been used in some of the challenge candidate systems (Politis *et al.*, 2020b). Chytas and Potamianos (2019) used convolutional layers with hundreds of filters of size 4×10 for azimuth and elevation estimation in a regression mode. Kong *et al.* (2019) compared different numbers of convolutional layers for SELD, while an 8-layer CNN was proposed by Noh *et al.* (2019) to improve the results over the baseline.

An indirect use of a CNN was proposed by Salvati *et al.* (2018). They trained the neural network to estimate a weight for each of the narrow-band SRP components fed at the input layer in order to compute a weighted combination of these components. In their experiments, they showed on a few test examples that this allowed for a better fusion of the narrow-band components and reduced the effects of noise and reverberation, leading to better localization accuracy.

In the DoA estimation literature, a few works have explored the use of *dilated convolutions* in DNNs. Dilated convolutions, also known as *atrous* convolutions, are a type of convolutional layer in which the convolution kernel is wider than the classical one but zeros are inserted so that the number of parameters remains the same. Formally, a 1D dilated convolution with a dilation factor l is defined by

$$(x * k)(n) = \sum_i x(n - li)k(i), \tag{9}$$

where x is the input and k the convolution kernel. The conventional linear convolution is obtained with $l = 1$. This definition extends to multidimensional convolution.

Chakrabarty and Habets (2019a) demonstrate that incorporating dilated convolutions with gradually increasing dilation factors reduces the optimal number of convolutional layers of their original CNN architecture (Chakrabarty and Habets, 2019b) (discussed previously in this section). This leads to an architecture with similar SSL performance and lower computational cost.

C. RNNs

RNNs are neural networks designed for modeling temporal sequences of data (Goodfellow *et al.*, 2016; LeCun *et al.*, 2015). Particular types of RNNs include long short-term memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) and gated recurrent units (GRUs) (Cho *et al.*, 2014). These two types of RNNs have become very popular thanks to their capability to circumvent the training difficulties that regular RNNs face, in particular the vanishing and exploding gradient problems (Goodfellow *et al.*, 2016; LeCun *et al.*, 2015).

There are few published works on SSL using only RNNs, as recurrent layers are often combined with convolutional layers (see Sec. IV D). Nguyen *et al.* (2021a) used an RNN to align SED and DoA predictions, which were obtained separately for each possible sound event type. The RNN was ultimately used to determine which SED prediction matched which DoA estimation. A bidirectional LSTM network was used by Wang *et al.* (2019) to estimate a TF mask to enhance the signal, further facilitating DoA estimation by conventional methods such as SRP or subspace methods.

D. CRNNs

CRNNs are neural networks containing one or more convolutional layers and one or more recurrent layers. CRNNs have been regularly exploited for SSL since 2018 because of the respective capabilities of these layers: The convolutional layers have proven to be suitable for extracting relevant features for SSL, and the recurrent layers are well designed for integrating the information over time.

In the series of papers by Adavanne *et al.* (Adavanne *et al.*, 2019a, 2018, 2019b), the authors used a CRNN for SELD, in a multi-task configuration, with first-order Ambisonics (FOA) input features (see Sec. V). In Adavanne *et al.* (2018), their architecture contained a series of successive convolutional layers, each followed by a max-pooling layer and two bidirectional gated recurrent unit (BGRU) layers. Then, a feedforward layer provided an estimation of the spatial pseudo-spectrum (SPS) provided by the MUSIC algorithm (Schmidt, 1986), acting as an intermediary output (see Fig. 4). This SPS was then fed into the second part of the neural network, which was composed of two convolutional layers, a dense layer, two BGRU layers, and a final feedforward layer for azimuth and elevation estimation by classification. The use of an intermediary SPS output has been proposed to help the neural network learn a representation that has proven to be useful for SSL using traditional methods.

In Adavanne *et al.* (2019a) and Adavanne *et al.* (2019b), this intermediary output was no longer used. Instead, the DoA was directly estimated using a block of convolutional layers, a block of BGRU layers, and a feedforward layer. This system is able to localize and detect several sound events even if they overlap in time, provided they are of different types (e.g., speech and car, see the discussion in Sec. II B). This CRNN was the baseline system for Task 3 of the DCASE Challenge in 2019 and 2020. Therefore, it has inspired many other works, and many DCASE Challenge candidate systems were built on the system of Adavanne *et al.* (2019a) with various modifications and improvements.

For example, Lin and Wang (2019) added Gaussian noise to the input spectrograms to train the network to be more robust to noise. Lu (2019) integrated some additional convolutional layers and replaced the BGRU layers with bidirectional LSTM layers. Leung and Ren (2019) used the same architecture with all combinations of cross-channel power spectra, whereas the replacement of input features with group delays was tested by Nustede and Anemüller (2019). GCC-PHAT features were added as input features by Maruri *et al.* (2019). Zhang *et al.* (2019a) used data augmentation during training and averaged the output of the network for a more stable DoA estimation. Xue *et al.* (2019)

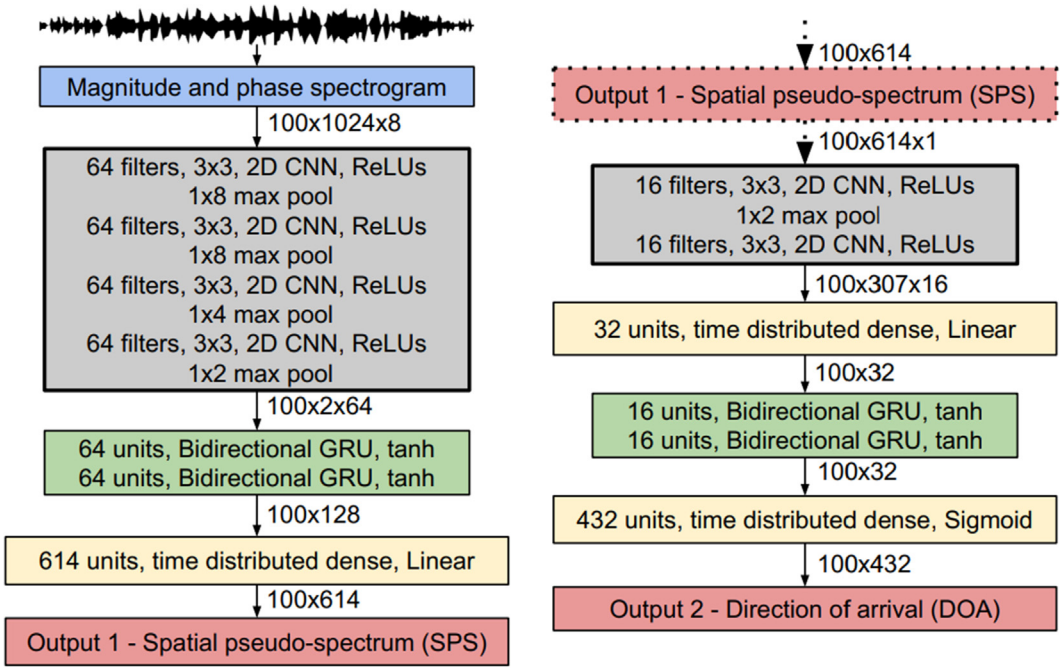

FIG. 4. (Color online) The CRNN architecture of Adavanne *et al.* (2019a), Adavanne *et al.* (2019b), and Adavanne *et al.* (2018), which has inspired numerous SELD systems. The input is the multichannel STFT-domain FOA magnitude and phase spectrogram. First, features are extracted by four successive convolutional layers with sixty-four 3×3 kernels, each followed by a max-pooling layer. Then two BGRU layers with 64 units each and tanh activations are used to capture the temporal evolution of the extracted features. An intermediate SPS output is then computed using a time distributed feedforward layer (i.e., this layer is computed separately on each vector of the temporal axis). Then, two 16-kernel convolution layers followed by a 32-unit time distributed feedfoward layer and two 16-units BGRU layers process the estimated SPS. A final 432-unit time distributed feedforward layer with sigmoid activation function is employed to infer the DoA. Note: Reprinted from Adavanne *et al.* (2018); copyright by IEEE; reprinted with permission.

sent the input features separately into different branches of convolutional layers, log-mel, and constant Q-transform features on the one hand, and phase spectrograms and CPS features on the other hand (see Sec. V). Cao *et al.* (2019b) concatenated the log-mel spectrogram and GCC-PHAT features and fed them into two separate CRNNs for SED and DoA estimation (they also incorporated the intensity vector in Cao *et al.*, 2019a). In contrast to the baseline of Adavanne *et al.* (2019a), more convolutional layers and one single BGRU layer were used. The convolutional part of the DoA network was transferred from the SED CRNN, which was followed by fine-tuning of the DoA branch, labelling this method as *two-stage*. This led to a notable improvement in localization performance over the DCASE Challenge baseline of Adavanne *et al.* (2019a). Small changes to this baseline were also tested by Pratik *et al.* (2019), such as the use of Bark-scale spectrograms as input features, the modification of the activation function or pooling layers, and the use of data augmentation, resulting in noticeable improvements for some experiments.

The same baseline neural architecture of Adavanne *et al.* (2019a) was used by Kapka and Lewandowski (2019), with one separate (but identical, except for the output layer) CRNN instance for each subtask: source counting (up to two sources), DoA estimation of source 1 (if applicable), DoA estimation of source 2 (if applicable), and sound type classification. The authors showed that their method was more efficient than the baseline. Krause and Kowalczyk (2019) explored different manners of splitting the SED and DoA estimation tasks in a CRNN. While some configurations showed an improvement in SED, the localization accuracy was below the baseline for the reported experiments. Park *et al.* (2019b) investigated a combination of a gated linear unit (GLU, a convolutional block with a gated mechanism) and a trellis network (containing convolutional and recurrent layers, see the paper by Bai *et al.* (2019) for details), yielding better results than the baseline. The authors extended this work for the DCASE 2020 Challenge by improving the overall architecture and investigating other loss functions (Park *et al.*, 2020). A non-direct DoA estimation scheme was also derived by Grondin *et al.* (2019), who estimated the TDoA using a CRNN, from which they inferred the DoA.

We also found propositions of CRNN-based systems in the 2020 edition of the DCASE Challenge. Singla *et al.* (2020) used the same CRNN as in the baseline of Adavanne *et al.* (2019a), except that they did not use two separated output branches for SED and DoA estimation. Instead, they concatenated the SED output with the output of the previous layer to estimate the DoA. Song (2020) used separated neural networks similar to the one of Adavanne *et al.* (2019a) to address NoS estimation and DoA estimation in a sequential way. Multiple CRNNs were trained by Tian (2020): one to estimate the NoS (up to two sources), another to estimate the DoA assuming one active source, and another (same as the baseline) to estimate the DoAs of two simultaneously active sources. Cao *et al.* (2020) designed an end-to-end

https://doi.org/10.1121/10.0011809

CRNN architecture to detect and estimate the DoA of possibly two instances of the same sound event. The addition of 1D convolutional filters was investigated by Ronchini *et al.* (2020) to exploit the information along the feature axes. Sampathkumar and Kowerko (2020) augmented the baseline system of Adavanne *et al.* (2019a) by providing the network with more input features (log-mel spectrograms, GCC-PHAT, and intensity vector, see Sec. V).

Independently of the DCASE Challenge, the CRNN of Adavanne *et al.* (2019a) was adapted by Comminiello *et al.* (2019) to receive quaternion FOA input features, which slightly improved the CRNN performance. Perotin *et al.* proposed using a CRNN with bidirectional LSTM layers on the FOA pseudo-intensity vector to localize one (Perotin *et al.*, 2018b) or two (Perotin *et al.*, 2019b) speakers. They showed that this architecture achieves very good performance in simulated and real reverberant environments with static speakers (both types of input features are discussed in Sec. V). This work was extended by Grumiaux *et al.* (2021a), who obtained a substantial improvement in performance over the CRNN of Perotin *et al.* (2019b) by adding more convolutional layers with less max-pooling, to localize up to three simultaneous speakers.

Non-square convolutional filters and a unidirectional LSTM layer were used in the CRNN architecture of Li *et al.* (2018). Xue *et al.* (2020) presented a CRNN with two types of input features: the phase of the CPS and the signal waveforms. The former was first processed by a series of convolutional layers before being concatenated with the latter. Another improvement of the network of Adavanne *et al.* (2019a) was proposed by Komatsu *et al.* (2020), who replaced the classical convolutional blocks with GLUs, based on the hypothesis that GLUs are better suited for extracting relevant features from phase spectrograms. This has led to a notable improvement of localization performance compared to the baseline of Adavanne *et al.* (2019a). Bohlender *et al.* (2021) proposed an extension of the system of Chakrabarty and Habets (2019b), in which LSTMs and temporal convolutional networks (TCNs) replaced the last dense layer of the former architecture. A TCN was made of successive 1D dilated causal convolutional layers with increasing dilated factors (Lea *et al.*, 2017). The authors showed that taking the temporal context into account with such temporal layers actually improves the localization accuracy.

Finally, we can mention the original approach of Nguyen *et al.* (2020c) in which a two-step hybrid approach with two CRNNs is used: In the first step, a first CRNN is used for SED and a single-source histogram-based (conventional) method is used for DoA estimation. In the second step, a second CRNN-based network, referred to as sequence matching network (SMN), is used to match the estimated sequences from the SED and DoA branches. This approach is motivated by the fact that overlapping sounds often have different onsets and offsets, and by matching the outputs of the two branches, an estimated DoA can be associated with the corresponding sound class. This approach was extended to localize moving sources in the framework of the DCASE 2020 Challenge, by adapting the resolution of the azimuth and elevation histograms and by using an ensemble of SMNs (Nguyen *et al.*, 2020b).

E. Residual neural networks

Residual neural networks were originally introduced by He *et al.* (2016), who pointed out that designing very deep networks can lead the gradients to explode or vanish due to the non-linear activation functions, as well as the degradation of the overall performance. Residual connections are designed to enable a feature to bypass a layer block in parallel to the conventional process through this layer block. This allows the gradients to flow directly through the network, usually leading to a better training.

To our knowledge, the first use of a network with residual connections for SSL was proposed by Yalta *et al.* (2017). As illustrated in Fig. 5, this network includes three residual blocks, which are stacks of layers with one of the layers having residual connections with another layer deeper in the stack. Each of these blocks is made of three convolutional layers, the first and last of which are designed with 1×1 filters, with the middle layer designed with 3×3 filters. A residual connection is used between the input and output of each residual block. The same type of residual block was used for SSL by He *et al.* (2018b, 2019a) in parallel to sound classification as speech or non-speech. Suvorov *et al.* (2018) used a series of 1D convolutional layers with several residual connections for single-source localization, directly from the multichannel waveform.

Pujol *et al.* (2019, 2021) integrated residual connections alongside 1D dilated convolutional layers with increasing dilation factors. They used the multichannel waveform as the network input. After the input layer, the architecture was divided into several subnetworks containing the dilated convolutional layers, which functioned as filter banks. Ranjan *et al.* (2019) combined a modified version of the original ResNet architecture (He *et al.*, 2016) with recurrent layers for SELD. This was shown to reduce the DoA error by more than 20° compared to the baseline of Adavanne *et al.* (2019a). Similarly, Bai *et al.* (2021) also used the ResNet model of (He *et al.*, 2016) followed by two GRU layers and two fully-connected layers for SELD. Kujawski *et al.* (2019) also adopted the original ResNet architecture and applied it to the single-source localization problem.

Another interesting architecture containing residual connections was proposed by Naranjo-Alcazar *et al.* (2020) for the DCASE 2020 Challenge. Before the recurrent layers (consisting of two BGRUs), three residual blocks successively processed the input features. These residual blocks contained two residual convolutional layers, followed by a squeeze-excitation module (Hu *et al.*, 2020). These modules aim to improve the modeling of interdependencies between input feature channels compared to classical convolutional layers. Similar squeeze-excitation mechanisms were used by Sundar *et al.* (2020) for multi-source localization. Another

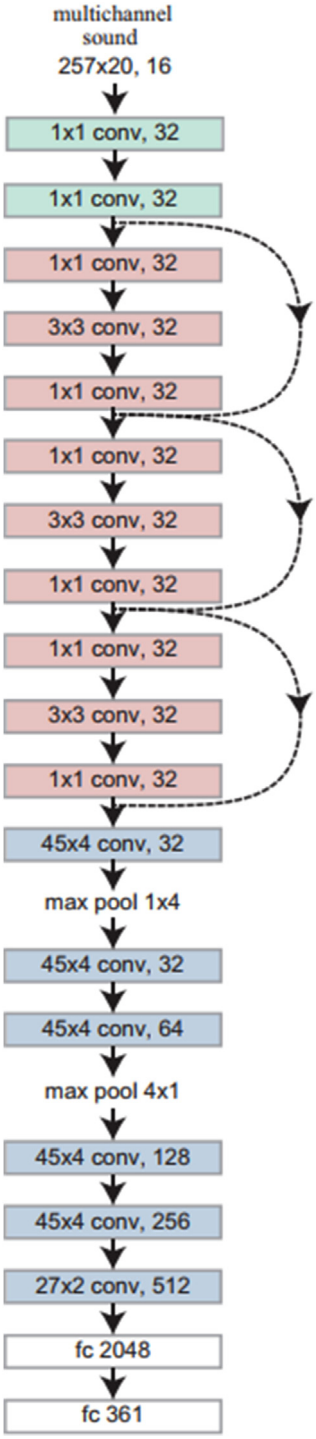

FIG. 5. (Color online) The residual neural network architecture used by Yalta *et al.* (2017). Three residual blocks are employed in this network, which are each composed of two convolutional layers with 32 1×1 filters with another convolutional layer with 32 3×3 filters in-between. For all three residual blocks, the input is added to the output with a residual connection, showed with a dashed arrow in this diagram. The authors show that the use of residual connections not only reduces the learning cost, but also improves the model performance. Note: Reprinted from Yalta *et al.* (2017); under Creative Commons Attribution-NoDerivatives 4.0 International License.

combination of a residual network with squeeze-excitation blocks was reported by Huang and Perez (2021), who implemented it in the framework of a sample-level CNN (i.e., a CNN applied on the time-domain signal samples) (Lee *et al.*, 2017). The resulting blocks are further followed by two Conformer blocks (see the next subsection). The motivation for combining these different models was their observed effectiveness in other audio processing tasks such as SED.

Shimada *et al.* (2020b, 2020a) adapted the MMDenseLSTM architecture, originally proposed by Takahashi *et al.* (2018) for sound source separation, to the SELD problem. This architecture consists of a series of blocks made of convolutions and recurrent layers with residual connections. Their system showed very good performance among the other participants to the DCASE 2020 Challenge. Wang *et al.* (2020) used an ensemble learning approach in which several variants of residual neural networks and recurrent layers were trained to estimate the DoA, achieving the best performance of the DCASE 2020 Challenge.

Guirguis *et al.* (2020) designed a neural network with a TCN in addition to classical 2D convolutions and residual connections. Instead of using recurrent layers as usually considered, the architecture was composed of TCN blocks that were made of several residual blocks, including a 1D dilated convolutional layer with an increasing dilated factor. The authors showed that replacing recurrent layers with TCNs made the hardware implementation of the network more efficient while slightly improving the SELD performance compared to the baseline of Adavanne *et al.* (2019a).

Yasuda *et al.* (2020) exploited a CRNN with residual connections in an indirect way for DoA estimation using an FOA pseudo-intensity vector input (see Sec. V E). A CRNN was first used to remove the reverberant part of the FOA pseudo-intensity vector, after which another CRNN was used to estimate a TF mask, which was applied to attenuate TF bins with a large amount of noise. The source DoA was finally estimated directly from the dereverberated and denoised pseudo-intensity vector.

F. Attention-based neural networks

An *attention mechanism* is a method that allows a neural network to put emphasis on vectors of a temporal sequence that are more relevant for a given task. Originally, attention was proposed by Bahdanau *et al.* (2016) to improve sequence-to-sequence models such as RNNs for machine translation. The general principle is to allocate a different weight to the vectors of the input sequence when using a combination of these vectors for estimating a vector of the output sequence. The model is trained to compute the optimal weights that reflect both the link between vectors of the input sequence (self-attention) and the relevance of the input vectors to explain each output vector (attention at the decoder). This pioneering work has inspired the now popular *Transformer* architecture proposed by Vaswani *et al.* (2017), which greatly improved the machine translation performance. In the Transformer, RNNs are removed, i.e., they are totally replaced by attention models.

https://doi.org/10.1121/10.0011809

Attention models are now used in an increasing number of DL applications, including SSL. Phan *et al.* (2020a,b) submitted an attention-based neural system for the DCASE 2020 Challenge. Their architecture was made of several convolutional layers, followed by a BGRU, after which a self-attention layer was used to infer the activity and the DoA of several distinct sound events at each time step. Schymura *et al.* (2020) added an attention mechanism after the recurrent layers of a CRNN to output an estimation of the sound source activity and its azimuth/elevation. Compared to the baseline of Adavanne *et al.* (2019a), the addition of attention demonstrated a better use of temporal information for SELD. An extension of the system of Chakrabarty and Habets (2019b) based on attention mechanisms has been proposed by Mack *et al.* (2020). Attention is employed to estimate binary masks to focus on frequency bins where the target source is dominant. The first attention stage appears right after the input layer (analogously to Chakrabarty and Habets, 2019b), their network uses phase spectrograms as inputs), while the second attention stage takes place after new features have been extracted using convolutional layers. Adavanne *et al.* (2021) used a self-attention layer after a GRU in order to estimate the association matrix which matches predictions and references. This solves the optimal assignment problem and resulted in large improvements in terms of localization error.

Multi-head self-attention (MHSA), which is the parallel use of several Transformer-type attention models (Vaswani *et al.*, 2017), has also inspired SSL methods. In the DCASE 2021 Challenge, Emmanuel *et al.* (2021) employed a MHSA layer right after several convolution modules tailored to learn varying spectral characteristics. Yalta *et al.* (2021) proposed using the whole encoder part of the Transformer architecture, in addition to several convolutional layers, to extract features from the input data. Wang *et al.* (2021) adapted the Conformer architecture, originally designed by Gulati *et al.* (2020) for automatic speech recognition, to SSL. This architecture is composed of a feature extraction module based on ResNet and a MHSA module that learns local and global context representations. The authors demonstrated the benefit of using a specific data augmentation technique on this model. Zhang *et al.* (2021) also employed this architecture in the DCASE 2021 Challenge. As briefly mentioned in the previous subsection, Conformer blocks were also used in the architecture proposed by Huang and Perez (2021), where they followed a sample-level CNN with residual connections and squeeze-excitation. A Conformer block was also used in the architecture proposed for SELD by Rho *et al.* (2021), after convolutional and fully-connected layers and before BGRU layers. Cao *et al.* (2021) positioned an 8-head attention layer after a series of convolutional layers to track the source location predictions over time for different sources (up to two sources in their experiments). Schymura *et al.* (2021) used three 4-head self-attention encoders along the time axis after a series of convolutional layers before estimating the activity and location of several sound events (see Fig. 6). This neural architecture showed an improvement over the DCASE Challenge baseline of Adavanne *et al.* (2019a). In the same line, Xinghao *et al.* (2021) replaced the conventional convolutional layers of the baseline with a combination of adaptive convolutional layers (using dilated convolutions with different dilation factors) and attention blocks. Another example of MHSA-based Transformer model for SSL can be found in the work of Park *et al.* (2021a). In this work, a pretrained model is fine-tuned with transfer learning. The output sequence corresponding to each 3 s-sequence of input data is averaged to provide one DoA estimation. Sudarsanam *et al.* (2021) enriched the CRNN baseline of Adavanne *et al.* (2019a) with a set of several MHSA blocks followed by fully-connected layers. They provided an analysis of the influence of the number and dimension of the MHSA blocks (the optimal number was found to be 2) and the number of heads (optimal was 8),

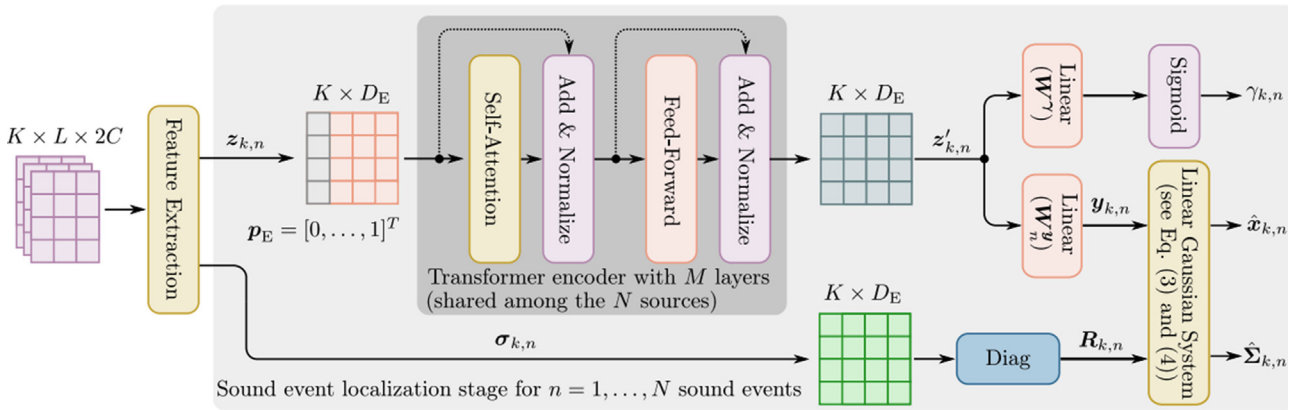

FIG. 6. (Color online) The self-attention-based neural network architecture of Schymura *et al.* (2021). The input is the multi-channel spectrogram shaped as a $K \times L \times 2C$ tensor, with K the number of frequency bins, L the number of frames, and C the number of channels. A feature extraction is first done with convolutional layers (not detailed in the figure) to produce $\mathbf{z}_{k,n}$ to which is attached a positional encoding vector $\mathbf{p}_E$. Then, a Transformer encoder computes a new representation of shape $K \times D_E$, which is used to compute the source activity $\gamma_{k,n}$ and the mean $\hat{\mathbf{x}}_{k,n}$ of the multivariate Gaussian distributions representing the target sources' location (the corresponding covariance matrix $\hat{\mathbf{\Sigma}}_{k,n}$ is computed *via* a parallel (simpler) mechanism.) Note: Reprinted from Schymura *et al.* (2021); copyright by the authors; reprinted with permission.

as well as the effect of positional embedding, normalization layers, and residual connections. Grumiaux *et al.* (2021b) showed that replacing the recurrent layers of a CRNN with self-attention encoders yielded a notable reduction in the computation time. Moreover, the use of MHSA slightly improved localization performance upon the baseline CRNN architecture of Perotin *et al.* (2019b) for the considered multiple speaker localization task.

Finally, we can mention the use of cross-modal attention (CMA) models for SSL by Lee *et al.* (2021b). A CMA model is the generalization of self-attention with two data streams in place of one, which is used in the Transformer decoder (Vaswani *et al.*, 2017). Lee *et al.* (2021b) used two separate SED and DoA estimation CNN blocks to separately produce SED and DoA embeddings (this comes in contrast with most DCASE candidate systems where the first blocks are shared between SED and DoA estimation). Then these embeddings are merged, first with a weighted linear combination and then with a second, more complex, alignment process using two mirrored CMA models. Finally, the SED and DoA outputs of the CMA modules are each sent to three parallel fully-connected networks for final estimation (this is because in the DCASE 2021 Challenge SELD Task, up to three sources can be simultaneously active).

In a general manner, it appears that attention modules, and MHSA in particular, have a tendency to replace the recurrent units in the recent SSL DNNs, following the "Attention is all you need" seminal line of Vaswani *et al.* (2017). This is because compared to RNNs, attention modules can model longer-term dependencies at a lower computational cost and can highly benefit from parallel computations, especially at training time. This tendency is also observed in other application domains, as we will discuss in Sec. IX.

G. Encoder-decoder neural networks

An encoder-decoder network is an architecture made of two building blocks: an *encoder*, which is fed by the input features and outputs a specific representation of the input data, and a *decoder*, which transforms the new data representation from the encoder into the desired output data. Architectures following this principle have been largely explored in the DL literature due to their capacity to provide compact data representations in an unsupervised manner (Goodfellow *et al.*, 2016).

1. Autoencoder (AE)

An AE is an encoder-decoder neural network that is trained to output a copy of its input. Often, the dimension of the encoder's last layer output is small compared to the dimension of the data. This layer is then known as the *bottleneck* layer and it provides a compressed encoding of the input data. Originally, AEs were made of feed-forward layers, but this term is also contemporaneously used to designate AE networks with other types of layers, such as convolutional or recurrent layers. To the best of our knowledge, the first use of an AE for DoA estimation was reported by Zermini *et al.* (2016). They used a simple AE to estimate TF masks for each possible DoA, which were then used for source separation. An interesting AE-based method was presented by Huang *et al.* (2020), in which an ensemble of AEs was trained to reproduce the multichannel input signal at the output, with one AE per candidate source position. Since the common latent information among the different channels is the dry signal, each encoder approximately deconvolves the signal from a given microphone. These dry signal estimates should be similar provided that the source is indeed at the assumed position; hence, the localization is performed by finding the AE with the most consistent latent representation. However, it is not clear whether this model can generalize well to unseen source positions and acoustic conditions.

Le Moing *et al.* (2020) presented an AE with a large number of convolutional layers (and transposed convolutional layers, which are layers of the decoder that process the inverse operation of the corresponding convolutional layer at the encoder), which estimates the potential source activity of each subregion in the (x, y) plane divided in a grid, making it possible to locate multiple sources. They evaluated several types of outputs (binary, Gaussian-based, and binary followed by regression refinement), each of which showed promising results on the simulated and real data. An extension of this work was presented in Le Moing *et al.* (2021), in which they proposed using adversarial training (see Sec. VIII) to improve network performance on real data, as well as on microphone arrays unseen in the training set, in an unsupervised training scheme. To do this, they introduced a novel *explicit transformation* layer that helped the network to be invariant to the microphone array layout. Another encoder-decoder architecture was proposed by He *et al.* (2021b), in which a multichannel waveform was fed into a filter bank with learnable parameters, after which a 1D convolutional encoder-decoder network processed the filter bank output. The output of the last decoder was then fed separately into two branches, one for SED and the other for DoA estimation.

An encoder-decoder structure with one encoder followed by two separate decoders was proposed by Wu *et al.* (2021b). Signals recorded from several microphone arrays were first transformed in the STFT domain (see Sec. V) and then stacked in a 4D-tensor (whose dimensions were time, frequency, microphone array, and microphone). This tensor was then sent to the encoder block, which was made of a series of convolutional layers followed by several residual blocks. The output of the encoder was then fed into two separate decoders, the first of which was trained to output a probability of source presence for each candidate (x, y) region, while the second was trained in the same way but with a range compensation to make the network more robust. The same general encoder-decoder line was adopted in the 2D image mapping approach proposed by Wu *et al.* (2021a). Note that here, the network is composed of convolutional layers at the encoder and transposed convolutional layers at the decoder, which is typical for image mapping applications in computer vision.

https://doi.org/10.1121/10.0011809

Indirect use of an AE was proposed by Vera-Diaz *et al.* (2020), who used convolutional and transposed convolutional layers to estimate the TDoA from GCC-based input features. The main idea was to rely on the encoder-decoder capacity to reduce the dimension of the input data so that the bottleneck representation forced the decoder to output a smoother version of the TDoA. This technique was shown to outperform the classical GCC-PHAT method in the reported experiments. This work was extended in the presence of two sources (Vera-Diaz *et al.*, 2021).

2. Variational autoencoder (VAE)

A VAE is a generative model that was originally proposed by Kingma and Welling (2014) and Rezende *et al.* (2014) and is now very popular in the DL community. A VAE can be seen as a probabilistic version of an AE. Unlike a classical AE, a VAE learns a probability distribution of the data at the output of the decoder and also models the probability distribution of the so-called latent vector at the bottleneck layer, which makes the VAE strongly connected to the concept of unsupervised representation learning (Bengio *et al.*, 2013). New data can thus be obtained with the decoder by sampling these distributions.

To our knowledge, Bianco *et al.* (2020) were the first to apply a VAE for SSL. Their VAE, made of convolutional layers, was trained to generate the phase of inter-microphone RTFs (see Sec. V A 1), jointly with a classifier that estimates the speaker's DoA from the RTF phases. The interest in using a VAE is that this generative model, originally designed for unsupervised training, is here trained in a semi-supervised configuration using a large dataset of unlabeled RTF data together with a limited set of labeled data (RTF values + corresponding DoA labels). In such a limited labeled dataset configuration, this model was shown to outperform an SRP-PHAT-based method as well as a supervised CNN in reverberant scenarios. This semi-supervised (or weakly supervised) approach is further discussed in Sec. IX A. An extension of this work has been further proposed in Bianco *et al.* (2021), with refined network architectures and more realistic acoustic scenarii.

3. U-Net architecture

A U-Net architecture is a particular fully-convolutional neural network originally proposed by Ronneberger *et al.* (2015) for biomedical image segmentation. In U-net, the input features are decomposed into successive feature maps throughout the encoder layers and then recomposed into "symmetrical" feature maps throughout the decoder layers, similarly to CNNs. Having the same dimension for feature maps at the same level in the encoder and decoder enables one to propagate information directly from an encoder level to the corresponding level of the decoder *via* residual connections. This leads to the typical U-shape schematization (see Fig. 7).

Regarding SSL and DoA estimation, several works have been inspired by the original U-Net paper. Chazan *et al.* (2019) employed such an architecture to estimate one TF mask per considered DoA (see Fig. 7), in which each TF bin was associated with a single particular DoA. This spectral mask was finally applied for source separation. This system was extended by Hammer *et al.* (2021) to account for multiple moving speakers. Another joint localization and separation system based on a U-Net architecture was proposed by Jenrungrot *et al.* (2020). In this system, a U-Net was trained based on 1D convolutional layers and GLUs. The input is the multichannel raw waveform accompanied by an angular window that helps the network to perform separation on a particular zone. If the output of the network on the window is empty, no source is detected, otherwise, one or more sources are detected and the process is repeated with a smaller angular window until the angular window reaches 2°. This system shows interesting results on both

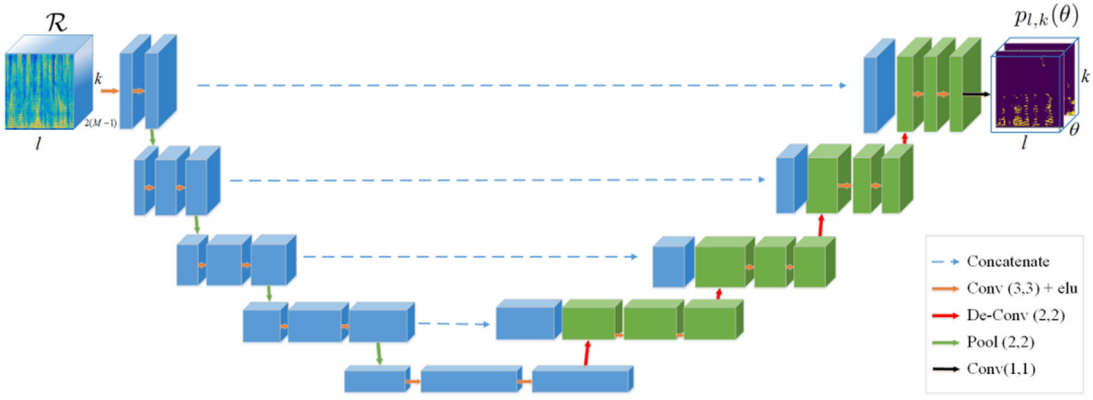

FIG. 7. (Color online) The U-Net network architecture of Chazan *et al.* (2019). The input matrix $\mathcal{R}$ contains angular features extracted from the RTFs (see Sec. V A 1) (l, k, and M denote the time index, the frequency bin, and the number of microphones, respectively). Several stages of encoders (in blue) and decoders (in green) are used. At each encoder (or decoder) stage, two or three convolutional layers with 3×3 kernels are employed to compute a new representation which is used as the input of the next encoder (or decoder, respectively), except for the bottleneck stage from which the output is fed as input into the upper-stage decoder. Residual connections are used to concatenate one encoder output to the input of the same stage decoder, to alleviate the loss information problem. The output of this system consists of one TF mask $p_{l,k}(\theta)$ per considered DoA θ. Note: Reprinted from Chazan *et al.* (2019); copyright by IEEE; reprinted with permission.

synthetic and real reverberant data containing up to eight speakers.

For the DCASE 2020 Challenge, a U-Net with several BGRU layers in-between the convolutional blocks was proposed for SELD by Patel *et al.* (2020). The last transposed convolutional layer of this U-Net outputs a single-channel feature map per sound event, corresponding to its activity and DoA for all frames. This showed an improvement over the baseline of Adavanne *et al.* (2019a) in terms of DoA error. Comanducci *et al.* (2020a) used a U-Net architecture in the second part of their proposed neural network to estimate the source coordinates x and y. The first part, composed of convolutional layers, learns to map GCC-PHAT features to the so-called ray space (where source positions correspond to linear patterns (*cf.* Bianchi *et al.*, 2016), which is an intermediate representation used as the input of the U-Net architecture.

V. INPUT FEATURES

In this section, we provide an overview of the variety of input feature types found in the DL-based SSL literature. Generally, the considered features can be low-level signal representations such as waveforms or spectrograms, handcrafted features such as binaural features, or they can be borrowed from traditional SP methods such as MUSIC or GCC-PHAT.

Overwhelmingly, the input features for the SSL neural networks are based on some representation readily used in signal processing, often emphasizing spatial and/or TF information embedded in the signal. This seems to yield good results, despite the growing trend in other domains to learn the feature representation directly from raw data. One interpretation may be that the network architectures in SSL are usually of a relatively modest size, as compared to end-to-end models used in some other domains, e.g., NLP. A few publications have compared different types of input features for SSL (e.g., Krause *et al.*, 2020b; Roden *et al.*, 2015).

It is also quite common to provide the network with concatenated features of different nature (even if these carry redundant information), which usually has positive impact on performance. This can be attributed to the flexibility of the learning process, which seemingly adapts the network weights such that the pertinent information is efficiently "routed" from such an input to the upper layers of the network, where it is merged into an abstract, optimized feature representation.

We organized this section into the following feature categories: inter-channel, CC-based, spectrogram-based, Ambisonics, intensity-based, and finally the direct use of the multichannel waveforms. Note that, as stated previously, different kinds of features are often combined at the input layer of SSL neural networks.

A. Inter-channel features

1. RTF

The RTF is a very general inter-channel feature that has been widely used for conventional (non-deep) SSL and other spatial audio processing such as source separation and beamforming (Gannot *et al.*, 2017) and acoustic echo cancellation (Valero and Habets, 2017), and is now considered for DL-based SSL as well. The RTF is defined for a given sound source position and for a microphone pair as the ratio $H(f) = A_2(f)/A_1(f)$ of the source-to-microphone ATFs of the two microphones, $A_2(f)$ and $A_1(f)$ (here we are working in the frequency or STFT domain and we recall that an ATF is the discrete Fourier transform of the corresponding RIR). It is thus strongly dependent on the source DoA (for a given recording set-up). In a multichannel set-up with more than two microphones, we can define an RTF for each microphone pair. Often, one microphone is used as a reference microphone, and the ATFs of all other microphones are divided by the ATF of this reference microphone.

As an ATF ratio, an RTF is thus a vector with an entry defined for each frequency bin. If only one directional source is present in the recorded signals and if the (diffuse) background noise is negligible, Eq. (2) shows that an RTF estimate can be obtained for each STFT frame (indexed by n), each frequency bin, and each microphone pair (indexed by i and k) by taking the ratio between the STFT transforms of the recorded waveforms of the two considered channels, $X_i(f,n)$ and $X_k(f,n)$:

$$\hat{H}_{i,k}(f) = \frac{X_k(f,n)}{X_i(f,n)} \approx \frac{A_k(f)S(f,n)}{A_i(f)S(f,n)} = \frac{A_k(f)}{A_i(f)} = H_{i,k}(f), \qquad (10)$$

where $S(f, n)$ is the STFT of the source signal. In the case where a background/sensor noise is present, more sophisticated RTF estimation procedures must be used (e.g., Cohen, 2004; Li *et al.*, 2015; Markovich-Golan and Gannot, 2015). If multiple sources are present, things become more complicated, but using the natural sparsity of speech/audio signals in the TF domain, i.e., only at most one source is assumed to be active in each TF bin (Rickard, 2002), the same principle as for one active source can be applied separately in each TF bin. Therefore, a multiple set of estimated RTFs at different frequencies (and possibly at different time frames if the sources are static or not moving too fast) can be used for multi-source localization. The reader is referred to (Gannot *et al.*, 2017) and references therein for more information on the RTF estimation problem.

An RTF is a complex-valued vector. In practice, an equivalent real-valued pair of vectors is often used. We can use either the real and imaginary parts or the modulus and argument. Often, the log-squared value of the interchannel power ratio is used, i.e., the interchannel power ratio in dB, and the argument of the RTF estimate ideally corresponds to the difference of the ATF phases. Such RTF-based representations have been used in several DNN-based systems for SSL. For example, Chazan *et al.* (2019), Hammer *et al.* (2021), Bianco *et al.* (2021), and Bianco *et al.* (2020) used as input features the arguments of the measured RTFs obtained from all microphone pairs.

https://doi.org/10.1121/10.0011809

2. Binaural features

Binaural features have also been used extensively for SSL, in both conventional and deep systems (Argentieri *et al.*, 2015). These features correspond to a specific two-channel recording set-up, one which attempts to reproduce human hearing in the most realistic way possible. Toward this aim, a dummy head/body with in-ear microphones is used to mimic the source-to-human-ear propagation, and in particular the effects of the head and external ear (pinnae), which are important for source localization by the human perception system. In an anechoic binaural set-up environment, the (two-channel) source-to-microphone impulse response is referred to as the binaural impulse response (BIR). The frequency-domain representation of a BIR is the HRTF. Both BIR and HRTF are functions of the source DoA. To take into account the room acoustics in a real-world SSL application, BIRs are extended to binaural room impulse responses (BRIRs), which combine head/body effects and room effects (in particular reverberation, see further discussion on BRIR simulation in Sec. VII A).

Several binaural features are derived from binaural recordings: The interaural level difference corresponds to the short-term log-power magnitude of the ratio between the two binaural channels in the STFT domain, $X_2(f,n)$ and $X_1(f,n)$,

$$ILD(f,n) = 20\log_{10}\left|\frac{X_2(f,n)}{X_1(f,n)}\right|. \tag{11}$$

The interaural phase difference is the argument of this ratio,

$$IPD(f,n) = \angle\frac{X_2(f,n)}{X_1(f,n)}, \tag{12}$$

and the interaural time difference is the delay that maximizes the CC between the two channels, similarly to the TDoA in Eq. (6). Just like the RTF, these features are actually vectors with frequency-dependent entries. In fact, the ILD and IPD are strongly related (not to say similar) to the log-power and argument of the RTF, as shown by comparing Eqs. (11) and (12) with Eq. (10), the difference relying more on the set-up than on the features themselves. The RTF can be seen as a more general (multichannel) concept, whereas binaural features refer to the specific two-channel binaural setup. As for the RTF, the ILD, IPD, and ITD implicitly encode the position of a source. Again, when several sources are present, the sparsity of speech/audio signals in the TF domain allows ILD/IPD/ITD values to provide information on the position of several simultaneously active sources.

Youssef *et al.* (2013) used ILD and ITD vectors fed separately into specific input branches of an MLP. Ma *et al.* (2015) and Yiwere and Rhee (2017) concatenated the CC of the two binaural channels with the ILD before feeding it into the input layer of their network. Ma *et al.* (2015) justify this choice with two arguments. The first one is to avoid the noise-sensitivity of the peak-picking operation for the computation of the ITD, the second one is because of the systematic changes in the CC function according to the source azimuth. Nguyen *et al.* (2018) used the IPD as the argument of a unitary complex number that was decomposed into real and imaginary parts. These parts were concatenated to the ILD for several frequency bins and several time frames, leading to a 2D tensor that was then fed into a CNN. Pang *et al.* (2019) also used a CNN to process ILD and IPD features in the TF domain, but the ILD and IPD 2D-tensors were directly concatenated at the input of the CNN. A system relying only on the IPD was proposed by Pak and Shin (2019). An MLP was trained to output a clean version of the noisy input IPD in order to better retrieve the DoA using a conventional method. Sivasankaran *et al.* (2018) used as input features the concatenation of the cosine and sine of the IPDs for several frequency bins and time frames. This choice was based on a previous work that showed similar performance for this type of input feature compared to classical phase maps, but with a lower dimension. In an original way, Thuillier *et al.* (2018) employed unusual binaural features. They used the ipsilateral and contralateral spectra. These features were shown to be relevant for elevation estimation using a CNN. We finally found other DNN-based systems that used ILD (e.g., Roden *et al.*, 2015; Zermini *et al.*, 2016), ITD (e.g., Roden *et al.*, 2015), or IPD (e.g., Shimada *et al.*, 2020a; Shimada *et al.*, 2020b; Subramanian *et al.*, 2021b; Zermini *et al.*, 2016) in addition to other types of features.

B. CC-based features

Another manner for extracting and exploiting inter-channel information that depends on source location is to use features based on the CC between the signals of different channels. In particular, as seen in Sec. III, a variant of CC known as GCC-PHAT is a common feature used in classical localization methods (Knapp and Carter, 1976). It is less sensitive to speech signal variations than standard CC, but it may be adversely affected by noise and reverberation (Blandin *et al.*, 2012). Therefore, it has been used within the framework of neural networks, which was revealed to be robust to this type of disturbance/artefact. In several systems, GCC-PHAT has been computed for each microphone pair and several time delays, all concatenated to form a 1D vector used as the input of an MLP (e.g., He *et al.*, 2018a; Vesperini *et al.*, 2016; Xiao *et al.*, 2015). Other architectures include convolutional layers to extract useful information from multi-frame GCC-PHAT features, e.g., (Comanducci *et al.*, 2020a; He *et al.*, 2018a; Li *et al.*, 2018; Lu, 2019; Maruri *et al.*, 2019; Noh *et al.*, 2019; Pratik *et al.*, 2019; Song, 2020; Vecchiotti *et al.*, 2019b; Vecchiotti *et al.*, 2018).

Some SSL systems rely on the CPS, which we already mentioned in Sec. III and which is linked to the CC by a Fourier transform operation (in practice, short-term estimates of the CPS are obtained by multiplying the STFT of one channel with the conjugate STFT of the other channel).

Leung and Ren (2019) and Xue *et al.* (2020) sent the CPS into a CRNN architecture to improve localization performance over the baseline of Adavanne *et al.* (2019a) (see Sec. IV). Grondin *et al.* (2019) also used the cross-spectrum for each microphone pair in the convolutional block of their architecture, whereas GCC-PHAT features were concatenated in a deeper layer. The CPS was also used by Ma and Liu (2018) as an input feature. Acoustic imaging has traditionally shown some interest in the CPS feature to predict localization and sound pressure level of competing sources; coupled with different architectures, from the simple MLP (Castellini *et al.*, 2021) to the complex CNN DenseNet network (Xu *et al.*, 2021a), authors have shown that the use of DNN could outperform traditional deconvolution methods, either in performance or computation time.

Traditional localization methods, such as MUSIC (Schmidt, 1986) or ESPRIT (Roy and Kailath, 1989), have been widely examined in the literature (see Sec. III). These methods are based on the eigen-decomposition of the CC matrix of a multichannel recording. Several DNN-based SSL systems (Takeda and Komatani, 2016a,b, 2017; Takeda *et al.*, 2018) have been inspired by these methods and reuse such features as input for their neural networks. Nguyen *et al.* (2020a) computed the spatial pseudo-spectrum based on the MUSIC algorithm and then used it as input features for a CNN.

Power map methods, which were discussed in Sec. III, have also been used to derive input features for DNN-based SSL systems. Salvati *et al.* (2018) proposed calculating the narrowband normalized steered response power for a set of candidate TDoAs corresponding to an angular grid and feeding it into a convolutional layer. This led to a localization performance improvement compared to the traditional SRP-PHAT method. Such power maps were also used by Diaz-Guerra *et al.* (2021b) as inputs of 3D convolutional layers. In acoustic imaging, a SRP map is also a standard feature where finding the position and the acoustic level is the main goal. Some recent works used a CNN (Gonçalves Pinto *et al.*, 2021) or a U-Net (Lee *et al.*, 2021b) to produce clean deconvolved maps, hence going beyond the intrinsic resolution of the array.

C. Spectrogram-based features

Alternatively to inter-channel features or CC-based features which already encode relative information between channels, another approach is to provide an SSL system directly with "raw" multichannel information, i.e., without any pre-processing in the channel dimension.

This does not prevent some pre-processing in the other dimensions and, from a historical perspective, we notice that many models in this line use spectral or spectro-temporal features instead of raw waveforms (see next subsection) as inputs. In practice, (multichannel) STFT spectrograms are typically used (Vincent *et al.*, 2018). These multichannel spectrograms are generally organized as 3D tensors, with one dimension for time (or frames), one for frequency (bins), and one for channel. The general spirit of DNN-based SSL methods is that the network should be able to "see" by itself and automatically extract and exploit the differences between TF spectrograms along the channel dimension while exploiting the sparsity of TF signal representation.

In several works, the individual spectral vectors from the different STFT frames were provided independently to the neural model, meaning that the network did not take into account their temporal correlation (and a localization result is generally obtained independently for each frame). Thus, in that case, the network input is a matrix of size $M \times K$, with M being the number of microphones, and K being the number of considered STFT frequency bins. Hirvonen (2015) concatenated the log-spectra of eight channels for each individual analysis frame and sent it into a CNN as a 2D matrix. Chakrabarty and Habets (2017a,b, 2019a,b) and Mack *et al.* (2020) used the multichannel phase spectrogram as input features, disregarding the magnitude information. This choice is motivated by the fact that it allows to easily generate a training dataset from white noise signals. As an extension of this work, phase maps were also exploited by Bohlender *et al.* (2021).

When several consecutive frames are considered, the STFT coefficients for multiple timesteps and multiple frequency bins form a 2D matrix for each recording channel. Usually, these spectrograms are stacked together in a third dimension to form the 3D input tensor. Several systems considered only the magnitude spectrograms (e.g., Patel *et al.*, 2020; Pertilä and Cakir, 2017; Wang *et al.*, 2019; Yalta *et al.*, 2017), while others considered only the phase spectrogram (e.g., Subramanian *et al.*, 2021b; Zhang *et al.*, 2019b). When considering both magnitude and phase, they can also be stacked in a third dimension (as well as channels). This representation has been employed in many DNN-based SSL systems (e.g., Guirguis *et al.*, 2020; He *et al.*, 2021a; Kapka and Lewandowski, 2019; Krause and Kowalczyk, 2019; Krause *et al.*, 2020a; Lin and Wang, 2019; Maruri *et al.*, 2019; Schymura *et al.*, 2021; Zhang *et al.*, 2019a). Yang *et al.* (2021a) dedicated different input branches of their CRNN to magnitude and phase features. Other authors have proposed to decompose the complex-valued spectrograms into real and imaginary parts (e.g., Hao *et al.*, 2020; He *et al.*, 2018b; Küçük *et al.*, 2019; Le Moing *et al.*, 2020). Finally, Leung and Ren (2019) tried several combinations of features computed from the complex multi-channel spectrogram, including the magnitude and phase, the real and imaginary parts, and the CPS. They claim that providing this redundant information could help the neural network for better localization.

While basic (STFT) spectrograms consider equally-spaced frequency bins, mel-scale spectrograms and Bark-scale spectrograms are represented with a non-linear sub-bands division, corresponding to a perceptual scale (low-frequency sub-bands have a higher resolution than high-frequency sub-bands) (Peeters, 2004). Mel-spectrograms were preferred to STFT spectrograms in several SSL neural networks (e.g., Cao *et al.*, 2019a; Cao *et al.*, 2019b;

https://doi.org/10.1121/10.0011809

Kong *et al.*, 2019; Ranjan *et al.*, 2019; Vecchiotti *et al.*, 2018). The Bark scale was also explored for spectrograms in the SSL system of Pratik *et al.* (2019).

D. Ambisonic signal representation

In the SSL literature, numerous systems utilize the Ambisonics format, i.e., the SH decomposition coefficients (Jarrett *et al.*, 2017), to represent the input signal. Ambisonics is a multichannel format that is increasingly used due to its capability to represent the spatial properties of a sound field, while being agnostic to the microphone array configuration (Zotter and Frank, 2019).

The SH decomposition is done for the acoustic pressure measured on the surface of a sphere $\mathbb{S}^2$, concentric with the microphone array. For a fixed sound source in far field, the decomposition coefficient of order ℓ and degree $m \in [-\ell, \ell]$, in the STFT domain, is given as follows (Jarrett *et al.*, 2017):

$$B_{\ell,m}(f,n) = \int_{\Omega\in\mathbb{S}^2} X(f,n,\Omega) Y^*_{\ell,m}(\Omega) d\Omega, \tag{13}$$

where $X(f,n,\Omega)$ and $Y_{\ell,m}(\Omega)$ are the acoustic pressure and the SH function, at the direction Ω, respectively. In practice, this integral is approximated by a quadrature rule, since the number of microphones consisting of an array is finite. Such approximation implies that the pressure $X(f,n,\Omega)$ is assumed to be an (almost) "order-limited" function on the sphere (Rafaely, 2019), meaning that $B_{\ell>L,m}(f,n) = 0$, for some maximal order L (that depends on the number of microphones in the array). Hence, for FOA ($L = 1$), the Ambisonics representation (13) counts only 4 coefficients (channels) per TF bin. Alternatively, the Higher-Order Ambisonics (HOA), $L > 1$, signals have more than four channels.

The plane wave, bearing an amplitude $S(f, n)$, and coming from a direction Ω, admits a simple SH representation $B_{\ell,m}(f,n) = S(f,n)Y_{\ell,m}(\Omega)$ (Rafaely, 2019). Therefore, as opposed to other types of microphone arrays, the Ambisonic channels are in phase, since the spatial response of each channel $Y_{\ell,m}(\Omega)$ is TF-independent.[2] Analogous to Eq. (3), the multichannel Ambisonic spectrogram $\mathbf{B}(f,n)$, due to J sources and reverberation, is given by the multivariate expression

$$\mathbf{B}(f,n) = \sum_{j=1}^{J}\sum_{r=0}^{\infty} A_{jr}(f,n) S_j(f,n) \mathbf{Y}(\Omega_{jr}) + \mathbf{N}(f,n), \tag{14}$$

where A_{jr} is the amplitude of the rth reflection of the source S_j (with $r = 0$ corresponding to the direct path), $\mathbf{Y}$ is the vector whose entries are appropriate spherical harmonics $Y_{\ell,m}$ for all considered Ambisonic orders, and $\mathbf{N}$ is the additive noise vector. Note that the complex-valued amplitudes A_{jr} account for the attenuation and phase shift of a corresponding plane wave component.

The FOA spectrograms, decomposed into magnitude and phase components, have been used by Adavanne *et al.* (2019a), Adavanne *et al.* (2018, 2019b); Guirguis *et al.* (2020), Kapka and Lewandowski (2019), and Krause and Kowalczyk (2019). Varanasi *et al.* (2020) and Poschadel *et al.* (2021a,b) used third-order Ambisonics spectrograms. Poschadel *et al.* (2021a) and Poschadel *et al.* (2021b) compared the performance of a CRNN with HOA spectrograms from order 1 to 4, showing that the higher the order, the better the localization accuracy of the network (but still below the performance of the so-called FOA *pseudo-intensity* features, which we will discuss in Sec. V E). They used the phase and magnitude for both elevation and azimuth estimation. Another way of representing the Ambisonics format was proposed by Comminiello *et al.* (2019). Based on the FOA spectrograms, they proposed considering them as quaternion-based input features, which proved to be a suitable representation in previous works (Parcollet *et al.*, 2018). To cope with this type of input feature, a neural network was adapted from the one of Adavanne *et al.* (2019a), showing an improvement over the baseline.

E. Intensity-based features

Sound intensity is an acoustic quantity defined as the product of sound pressure and particle velocity (Jacobsen and Juhl, 2013; Rossing, 2007). In the frequency or TF domain, sound intensity is a complex vector whose real part (known as "active" intensity) is proportional to the gradient of the phase of sound pressure, i.e., it is orthogonal to the wavefront. This is a useful property that has been extensively used for SSL (e.g., Evers *et al.*, 2014; Hickling *et al.*, 1993; Jarrett *et al.*, 2010; Kitić and Guérin, 2018; Nehorai and Paldi, 1994; Pavlidi *et al.*, 2015; Tervo, 2009). The imaginary part ("reactive" intensity) is related to oscillatory local energy transfers, and its physical interpretation is less obvious (Maysenhölder, 1993). Hence, it has been largely ignored by the SSL community, even though it is relevant in room acoustics (Nolan *et al.*, 2019). While the pressure is directly measurable by regular microphones, particle velocity requires specific sensors, such as acoustic vector-sensors (Jacobsen and Juhl, 2013; Nehorai and Paldi, 1994), e.g., the "Microflown" transducer (de Bree, 2003). Otherwise, it has to be approximated using the acoustic pressure measurements. Under certain conditions, particle velocity can be assumed to be proportional to the spatial gradient of sound pressure (Merimaa, 2006; Rossing, 2007), which allows for the estimation by, e.g., the finite difference method (Tervo, 2009) or using the FOA channels discussed in the previous section (Zotter and Frank, 2019). The latter approximation is often called (FOA) complex *pseudo-intensity* vector (Jarrett *et al.*, 2010),

$$\mathbf{I}(f,n) = B_{0,0}(f,n)\mathbf{B}_{\ell=1,m}(f,n)^*, \tag{15}$$

where $\mathbf{B}^*_{\ell=1,m}(f,n)$ is the vector of first-order SH coefficients, excluding the zero-order $B_{0,0}(f,n)$. In free field conditions, assuming the presence of a single source at the TF

bin (f, n), the entries of $\Re(\mathbf{I}(f,n))$ are the Cartesian coordinates of a vector colinear with the DoA of the source (with $\Re$ denoting the real part of a complex value).

The first use of an Ambisonics pseudo-intensity vector for DL-based SSL was reported by Perotin *et al.* (2018b), showing superiority in performance compared to the use of the raw Ambisonics waveforms and traditional Ambisonics-based methods. Interestingly, the authors demonstrated that using both active and reactive intensity improves SSL performance. Moreover, they normalized the intensity vector of each frequency band by its energy, which can be shown to yield features similar to RTFs in the spherical harmonics domain (Daniel and Kitić, 2020; Jarrett *et al.*, 2017). Yasuda *et al.* (2020) proposed using two CRNNs to refine the input FOA pseudo-intensity vector. The first CRNN is trained to estimate denoising and separation masks under the assumption that there are two active sources and that the WDO hypothesis holds. The second CRNN estimates another mask to remove the remaining unwanted components (e.g., reverberation). The two networks, hence, produce an estimate of the "clean" intensity vector for each active source (the NoS is estimated by their system as well). The pseudo-intensity vector has consequently been used in several other recent works, e.g., (Cao *et al.*, 2021; Cao *et al.*, 2019a; Grumiaux *et al.*, 2021a; Grumiaux *et al.*, 2021b; Nguyen *et al.*, 2021a; Park *et al.*, 2020; Perotin *et al.*, 2019a; Perotin *et al.*, 2019b; Song, 2020; Tang *et al.*, 2019).

Sound intensity was also explored by Liu *et al.* (2021) without the Ambisonics representation. The authors computed the instantaneous complex sound intensity using an average of the sound pressure across the four considered channels and two orthogonal particle velocity components using the differences in sound pressure for both microphone pairs. They kept only the real part of the estimated sound intensity (active intensity) and applied a PHAT weighting to improve the robustness against reverberation.

F. Waveforms

Since 2018, several authors have proposed directly providing their neural network models with the raw multichannel recorded signal waveforms. This idea relies on the DNN's capability to find the best representation for SSL without the need of hand-crafted features or pre-processing of any kind. This is in line with the general trend of DL to go toward an *end-to-end* approach that is observed in many other applications, including in speech/audio processing. Of course, this goes together with the always increasing size of networks, datasets, and computational power.

To our knowledge, Suvorov *et al.* (2018) were the first to apply this idea. They trained their neural network directly with the recorded eight-channel waveforms, stacking many 1D convolutional layers to extract high-level features for the final DoA classification. Vera-Diaz *et al.* (2018), Vecchiotti *et al.* (2019a), Chytas and Potamianos (2019), Cao *et al.* (2020), and Pujol *et al.* (2019, 2021) sent the raw multichannel waveforms into 2D convolutional layers. Huang and Perez (2021) sent the raw multichannel waveforms (in microphone format and FOA format) into a 1D CNN with residual connections and squeeze-excitation blocks. Note that this model is used for SELD and the authors motivate the use of raw waveform inputs by the fact that "SED and DOA may have some common features that are better preserved in the raw audio [wave]form." Huang *et al.* (2020) sent the multichannel waveforms into an AE. Jenrungrot *et al.* (2020) shifted the waveforms of each channel to make them temporally aligned according to the TDoA before being injected into the input layer of their network. In the same vein, Huang *et al.* (2018, 2019) proposed time-shifting the multichannel signal by calculating the time delay between the microphone position and the candidate source location, which requires scanning for all candidate locations.

A potential disadvantage of waveform-based features is that the architectures exploiting such data are often more complex, as one part of the network needs to be dedicated to feature extraction. Moreover, some papers have reported that learning the "optimal" feature representations from raw data becomes more difficult when noise is present in the input signals (Wichern *et al.*, 2019) or may even harm generalization, in some cases (Sato *et al.*, 2021). However, it is interesting to mention that the visual inspection of the learned weights of the input layers of some end-to-end (waveform-based) neural networks has revealed that they resemble the filterbanks that are usually applied in the pre-processing stage of SSL (see Sec. V C) and other various classical speech/audio processing tasks (Luo and Mesgarani, 2019; Sainath *et al.*, 2017).

G. Other types of features

Varzandeh *et al.* (2020) have proposed unusual types of features that do not belong to one of the categories described previously. Particularly, they have used a periodicity degree feature together with GCC-PHAT features in a CNN. The periodicity degree is computed for a given frame and period. It is equal to the ratio between the harmonic power signal for the given period and the total power signal. This conveys information about the harmonic content of the source signal to the CNN.

VI. OUTPUT STRATEGIES

In this section, we discuss the different strategies proposed in the literature to obtain a final DoA estimate. We generally divide the strategies into two categories: classification and regression. When the SSL network is designed for the classification task, the source location search space is generally divided into several zones, corresponding to different classes, and the neural network outputs a probability value for each class. As for regression, the goal is to directly estimate (continuous) source position/direction values, which are usually either Cartesian coordinates (x, y, z), or spherical coordinates (θ, ϕ, r) (although the source-

https://doi.org/10.1121/10.0011809

microphone distance r is very rarely considered). However, the latter is an important factor as it can affect the estimation accuracy (for instance, due to the influence of the direct-to-reverberant ratio, DRR; Vincent *et al.*, 2018). Therefore, in order to obtain a robust model, the training dataset needs to be sufficiently diverse such that the network is exposed to sources at different directions, but also at different source-microphone distances. In the last subsection, we report a few non-direct methods in which the neural network does not estimate the location of a source in its output layer. Instead, it either helps another (conventional) algorithm to finally retrieve the desired DoA, or the location estimate is a by-product of some intermediate network layer. A reader particularly interested in the comparison between the classification and regression approaches may consult the papers of Tang *et al.* (2019) and Perotin *et al.* (2019a).

A. DoA estimation *via* classification

Many systems treat DoA estimation as a classification problem, i.e., each class represents a certain zone in the considered search space. In other words, space is divided into several subregions, usually of similar size, and the neural network is trained to produce a probability of active source presence for each subregion. Such a classification problem is often addressed by using a feedforward layer as the last layer in the network, with as many neurons as the number of considered subregions. Two activation functions are generally associated with the final layer neurons: the softmax and sigmoid functions. Softmax ensures that the sum of all neuron outputs is 1, so it is suitable for a single-source localization scenario. With a sigmoid, all neuron outputs are within $[0, 1]$ independently from each other, which is suitable for multi-source localization. The last layer output is often referred to as the *spatial (pseudo)-spectrum*, whose peaks correspond to a high probability of source activity in the corresponding zone.

As already mentioned in Sec. II C, the final DoA estimate(s) is/are generally extracted using a peak picking algorithm: If the number of sources J is known, the selection of the J highest peaks gives the multi-source DoA estimation; if the NoS is unknown, usually the peaks above a certain user-defined threshold are selected, leading to a joint NoS and localization estimations. Some preprocessing, such as spatial spectrum smoothing or angular distance constraints, can be used for better DoA estimation. Hence, such a classification strategy can be readily used for single-source and/or multi-source localization, as the neural network is trained to estimate a probability of source activity in each zone, regardless of the NoS.

1. Spherical coordinates

Regarding the quantization of the source location space, namely, the localization *grid*, different approaches have been proposed. Most early works focused on estimating only the source's *azimuth* θ relative to the microphone array position, dividing the 360° azimuth space into N_θ regions of equal size, leading to a grid quantization step of $360/N_\theta$. Without being exhaustive, we found in the literature many different values for N_θ, e.g., $N_\theta = 7$ (Roden *et al.*, 2015), $N_\theta = 8$ (Hirvonen, 2015), $N_\theta = 20$ (Suvorov *et al.*, 2018), $N_\theta = 37$ (Vecchiotti *et al.*, 2019a), $N_\theta = 72$ (Ma *et al.*, 2015), and $N_\theta = 360$ (Xiao *et al.*, 2015). Some other works did not consider the whole 360° azimuth space. For example, Chazan *et al.* (2019) focused on the region $[0, 180]$ with $N_\theta = 13$.

Estimating the elevation ϕ alone has not been frequently investigated in the literature, probably because of the lack of interesting applications in indoor scenarios. To the best of our knowledge, only one paper focused on estimating the elevation alone (Thuillier *et al.*, 2018). The authors divided the whole elevation range into nine regions of equal size. The majority of recent SSL neural networks are trained to estimate both source azimuth and elevation, whenever the microphone array geometry makes it possible. To do this, several options have been proposed in the literature. One can use two separate output layers, each with the same number of neurons as the number of subregions in the corresponding dimension. For example, the output layer of the neural architecture proposed by Fahim *et al.* (2020) is divided into two branches with fully connected layers, one for azimuth estimation (N_θ neurons), and the other for elevation estimation (N_ϕ neurons). One can also have a single output layer where each neuron corresponds to a zone in the unit sphere, i.e., a unique pair (θ, ϕ) (e.g., Grumiaux *et al.*, 2021b; Perotin *et al.*, 2019b). Finally, one can directly design two separate neural networks, with each estimating the azimuth or the elevation angle, e.g., (Varanasi *et al.*, 2020).

However, most of the neural networks following the classification strategy for joint azimuth and elevation estimation are designed so that the output corresponds to a 2D grid on the unit sphere. For example, Perotin *et al.* (2018b, 2019b) and Grumiaux *et al.* (2021a) used a quasi-uniform spherical grid with 429 classes, each represented by a unique neuron in the output layer of their network. Adavanne *et al.* (2018) sampled the unit sphere in the whole azimuth axis but in the limited elevation range of $[-60°, 60°]$, yielding an output vector corresponding to 432 classes.

Distance estimation has barely been investigated in the SSL literature, highlighting the fact that it is a difficult problem. Roden *et al.* (2015) addressed the distance estimation along with azimuth or elevation prediction by dividing the distance range into five candidate classes. Yiwere and Rhee (2017) quantized the distance range into four classes and estimated it along with three possible azimuth values. In the paper by Takeda and Komatani (2016b), the azimuth axis was classified with $I = 72$ classes along with the distance and height of the source, but these last two quantities were classified into a very small set of possible pairs: (30, 30), (90, 30) and (90, 90) (in centimeters). Bologni *et al.* (2021) trained a CNN to classify a single-source signal into a 2D map representing the azimuth and distance dimensions.

2. Cartesian coordinates

A few works applied the classification paradigm to estimate the Cartesian coordinates. Le Moing *et al.* (2021), Le Moing *et al.* (2020), and Ma and Liu (2018) divided the horizontal (x, y) plane into small regions of the same size, with each being a class in the output layer. However, this representation suffers from a decreasing angular difference between the regions that are far from the microphone array, which is probably why regression is usually preferred for estimating Cartesian coordinates.

B. DoA estimation *via* regression

In regression SSL networks, the source location estimate is directly given by the continuous value provided by one or several output neurons (whether we consider Cartesian or spherical coordinates, and how many source coordinates are of interest). This technique offers the advantage of a potentially more accurate DoA estimation since there is no quantization. Its drawback is twofold. First, the NoS needs to be known or assumed, as there is no way to estimate if a source is active or not based on a localization regression. Second, regression-based SSL usually faces the well-known source permutation problem (Subramanian *et al.*, 2021b), which occurs in the multi-source localization configuration and is common with DL-based source separation methods. Indeed, during the computation of the loss function at the training time, there is an ambiguity in the association between target and actual output—in other words, which estimate should be associated with which target? This issue also arises during the evaluation. One possible solution is to force the SSL network training to be permutation invariant (Subramanian *et al.*, 2021b), in line with what was proposed for audio source separation (Yu *et al.*, 2017).

As for classification, when using regression, there is a variety of possibilities for the type of coordinates to be estimated. The choice among these possibilities is driven more by the context or the application than by design limitations since regression generally requires only a few output neurons.

1. Spherical coordinates

Tsuzuki *et al.* (2013) proposed a complex-valued neural approach for SSL. The output of the network is a complex number of unit amplitude whose argument is an estimate of the azimuth of the source. A direct regression scheme was employed by Nguyen *et al.* (2018) with a two-neuron output layer that predicts the azimuth and elevation values in a single-source environment. The system of Opochinsky *et al.* (2019) performed only azimuth estimation. Regarding the DCASE 2019 Challenge (Politis *et al.*, 2020b), a certain number of candidate systems have used two neurons per event type to estimate the azimuth and elevation of the considered event (e.g., Cao *et al.*, 2019a; Chytas and Potamianos, 2019; Park *et al.*, 2019b), while the event activity was jointly estimated in order to extract (or not) the corresponding coordinates. Sudo *et al.* (2019) proposed representing the output as a quaternion including the cosinus and sinus of the azimuth and elevation angles, from which they retrieve the DoA angle values. This enables to tackle the problem of discontinuity at angle interval boundaries (for instance, at $-180°$ and $180°$).

In the system of Maruri *et al.* (2019), azimuth and elevation estimations were done separately in two network branches, each containing a specific dense layer. Sundar *et al.* (2020) proposed a regression method relying on a preceding classification step: dividing the azimuth space into I equal subregions, with the output of the neural network being made of $3I$ neurons. Assuming there is at most one active source per subregion, three neurons are associated with each of them: one neuron is trained to detect the presence of a source, while the other two neurons estimate the distance and azimuth of that source. The loss function for training is a weighted sum of categorical cross-entropy (for the classification task) and mean square error (for the regression task).

2. Cartesian coordinates

Another way to predict the DoA with regression is to estimate the Cartesian coordinates of the source(s). Vesperini *et al.* (2016) designed their network output layer with only two neurons to estimate the coordinates x and y in the horizontal plane, with an output range normalized within $[0, 1]$, which represents the scaled version of the room size in each dimension. Following the same idea, Vecchiotti *et al.* (2019b) and Vecchiotti *et al.* (2018) also used two neurons to estimate (x, y) but added a third one to estimate the source activity.

The estimation of the three Cartesian coordinates (x, y, z) has been investigated in several systems. Vera-Diaz *et al.* (2018) and Krause *et al.* (2020a) designed the output layer with three neurons to estimate the coordinates of a single source with regression. Adavanne *et al.* (2019a) and Adavanne *et al.* (2019b) chose the same strategy. However, they performed SELD for several types of event, and thus there are three output neurons to provide (x, y, z) estimates for each event type, plus another output neuron to estimate whether or not this event is active. The hyperbolic tangent activation function is used for the localization neurons to keep the output values in the $[-1, 1]$ range, leading to a DoA estimate on the unit sphere. The same strategy was followed in an extension of this work by Comminiello *et al.* (2019).

In Shimada *et al.* (2020a), the authors proposed the activity-coupled cartesian direction of arrival (ACCDOA) representation which encodes the DoA with the source activity in a single vector, separately for each sound class to be localized. More specifically, the ACCDOA vector encodes the Cartesian coordinates (x, y, z), is then normalized and then multiplied by the source activity ($\in [0, 1]$). Using a threshold, the active sources can be detected using this vector norm, and their respective DoAs can be retrieved from

https://doi.org/10.1121/10.0011809

the normalized Cartesian coordinates. This ACCDOA output representation has then been used in other works (e.g., Emmanuel *et al.*, 2021; Naranjo-Alcazar *et al.*, 2021; Nguyen *et al.*, 2021b; Shimada *et al.*, 2021; Shimada *et al.*, 2020b; Sudarsanam *et al.*, 2021).

C. Non-direct DoA estimation

Neural networks have also been used in the regression mode to estimate intermediate quantities, which are then used by a non-neural algorithm to predict the final DoA.

Pertilä and Cakir (2017) proposed using a CNN in the regression mode to estimate a TF mask. This mask was then applied to the noisy multichannel spectrogram to obtain an estimate of the clean multichannel spectrogram, and a classical SRP-PHAT method was next applied to retrieve the final DoA. Another TF mask estimation was done by Wang *et al.* (2019) using a bidirectional LSTM network to improve traditional DoA estimation methods, such as GCC-PHAT or MUSIC. Pak and Shin (2019) trained an MLP to remove unwanted artefacts of the IPD input features. The cleaned feature was then used to estimate the DoA with a non-neural method. Yasuda *et al.* (2020) proposed a method to filter out reverberation and other non-desired effects from the intensity vector by TF mask estimation. The filtered intensity vector led to a better DoA estimation than an intensity-based conventional method. Yang *et al.* (2021a) used a two-stage neural network system to estimate the direct-path RTF (DP-RTF), that is, the part of the RTF that corresponds to the direct source-to-microphone propagation (Li *et al.*, 2016b). In Yang *et al.* (2021a), the source DoA is the direction parameter of a DP-RTF taken from a dictionary of pre-computed DP-RTFs, corresponding to the closest match with the network estimate.

Huang *et al.* (2018, 2019) employed neural networks on multichannel waveforms, shifted in time with a delay corresponding to a certain candidate source location, to estimate the original dry signal. Doing this for a set of candidate locations, they then calculated the sum of CC coefficients between the estimated dry source signals for all candidate source locations. The final estimated location was obtained as the one leading to the maximum sum.

A joint localization and separation scheme was proposed by Jenrungrot *et al.* (2020). The neural network was trained to estimate the signal coming from a certain direction within a certain angular window, whose parameters were injected as an input to each layer. Thus, the network acted like a radar and scanned through all directions, then progressively reduced the angular window up to a desired angular resolution.

Several works proposed employing neural networks for a better prediction of the TDoA, which is then used to determine the DoA as often done in traditional methods. Grondin *et al.* (2019) estimated the TDoA in the regression mode using a hyperbolic tangent activation function at the output layer. Vera-Diaz *et al.* (2020) used an AE to estimate a function from GCC-based features (similar to TDoA) that exhibited a clear peak corresponding to the estimated DoA. Their work was extended in the presence of two sources (Vera-Diaz *et al.*, 2021). In Comanducci *et al.* (2020b), the authors employed a U-Net in a regression manner to clean GCC-based features from noise and reverberation.

Subramanian *et al.* (2021a) proposed a neural system based on a stacked localization network, parametric beamformers and a speech recognition network. Since each of these modules is differentiable, the system is trained in the end-to-end mode, using an ASR-specific cost function. Despite being optimized for the ASR, the trained system also exhibits very good performance in terms of source separation and localization, whose predictions are the intermediate results, retrievable at the output of the corresponding processing modules.

VII. DATA

In this section, we detail the different approaches taken to deal with data during model training or testing. Because we are dealing with indoor domestic/office environments, noise and reverberation are common in real-world signals. We successively inspect the use of synthetic and recorded datasets in DNN-based SSL.

A. Synthetic data

A well-known limitation of supervised learning (see Sec. VIII) for SSL is the lack of labeled training data. In a general manner, it is difficult to produce datasets of recorded signals with corresponding source position metadata in diverse spatial configurations (and possibly with diverse spectral content) that would be sufficiently large for efficient SSL neural model training. Therefore, one often has to *simulate* a large amount of data to obtain an efficient SSL system.

To generate realistic data, taking into account reverberation, one needs to simulate the room acoustics. This is usually done by synthesizing the RIR that models the sound propagation for a "virtual" source-microphone pair. This is done for all microphones of the array (and for a large number of source positions and microphone array positions, see next). Then, a "dry" (i.e., clean reverberation-free monophonic) source signal is convolved with this RIR to obtain the simulated microphone signal (this is done for every channel of the microphone array). As already stated in Sec. I B, the foundation of SSL relies on the fact that the relative location of a source with respect to the microphone array position is implicitly encoded in the (multichannel) RIR, and an SSL DNN learns to extract and exploit this information from examples. Therefore, such data generation has to be done with many different dry signals and for a large number of simulated RIRs with different source and microphone array positions. The latter must be representative of the configurations in which the SSL system will be used in practice. Moreover, other parameters, such as room dimensions and reverberation time, may have to be varied to take into account other factors of variations in SSL.

One advantage of this approach is that many dry signal datasets exist, in particular for speech signals (e.g., Garofolo *et al.*, 1993a; Garofolo *et al.*, 1993b; Lamel *et al.*, 1991). Therefore, many SSL methods are trained with dry speech signals convolved with simulated RIRs. Chakrabarty and Habets (2017a,b) used white noise as the dry signal for training and speech signals for testing. This approach is reminiscent of the work of Deleforge *et al.* (2013) and Deleforge *et al.* (2015) based on a GMR and already mentioned in Sec. III. Using white noise as the dry signal enables the acquisition of training data that are "dense" in the TF domain. However, Vargas *et al.* (2021) showed that training on speech or music signals leads to better results than noise-based training, even when the signals are simulated with a generative adversarial network (GAN). Furthermore, the results of Krause *et al.* (2021) indicate that using speech, noise, and sound events data altogether leads to better localization performance, even compared to matched training and test signals.

As for RIR simulation, there exist several methods (and variants thereof) and acoustic simulation software. Detailing these methods and software implementations is out of the scope of this article, but an interested reader may consult appropriate references (e.g., Rindel, 2000; Siltanen *et al.*, 2010; Svensson and Kristiansen, 2002). Let us only mention that the simulators based on the image source method (ISM) (Allen and Berkley, 1979) have been widely used in the SSL community, probably due to the fact that they offer a relatively good trade-off between the simulation fidelity, in particular regarding the "head" of an RIR, i.e., the direct propagation and early reflections (Rindel, 2000), and computational complexity. Among publicly available libraries, the RIR generator of Habets (2006), the related signal generator (Habets, 2022), the Roomsim toolbox of Campbell *et al.* (2005), and its extension to mobile sources called Roomsimove (Vincent and Campbell, 2008), the Spherical Microphone Impulse Response (SMIR) generator of Jarrett *et al.* (2012), the Pyroomacoustics toolbox of Scheibler *et al.* (2018), and the Multichannel Room Acoustics Simulator (MCRoomSim) of Wabnitz *et al.* (2010), are very popular. Such libraries have been used by, e.g., Bianco *et al.* (2020); Chakrabarty and Habets (2019b); Grumiaux *et al.* (2021a); Li *et al.* (2018); Nguyen *et al.* (2020a); Perotin *et al.* (2019b); Salvati *et al.* (2018); Varanasi *et al.* (2020). An efficient open-source implementation of the ISM method, relying on graphic processing unit (GPU) acceleration, has been recently presented by Diaz-Guerra *et al.* (2021a) and used in Diaz-Guerra *et al.* (2021b) to simulate moving sources.

Other improved models based on the ISM have also been used to simulate impulse responses, such as the one presented by Hirvonen (2015). This model relies on that of Lehmann and Johansson (2010), which adds a diffuse reverberation model to the original ISM method. Hübner *et al.* (2021) proposed a low-complexity model-based training data generation method that includes a deterministic model for the direct path and a statistical model for late reverberation. It has been demonstrated that the SSL neural network, trained using the data generated by this method, achieves comparable localization performance as the same architecture trained on a dataset generated by the usual ISM. However, the proposed simulation method is computationally more efficient. An investigation of several simulation methods was done by Gelderblom *et al.* (2021), with extensions of ISM, namely, ISM with directional sources, and ISM with a diffuse field due to scattering. Gelderblom *et al.* (2021) compared the simulation algorithms *via* the training of an MLP (in both regression and classification modes) and showed that ISM with scattering effects and directional sources leads to the best SSL performance. More sophisticated software, such as ICARE® (Bouatouch *et al.*, 2006), often combine ISM with efficient ray-tracing and statistical methods, permitting simulation of more complicated room geometries and acoustic effects. Note, however, that none of the methods based on approximating the sound propagation by geometrical acoustics is capable of precisely simulating certain wave phenomena, such as diffraction (Kuttruff, 2016).

Training and testing binaural SSL systems require either directly using signals recorded in a binaural setup (see next subsection) or using a dataset of two-channel BIRs and convolving these BIRs with (speech/audio) dry signals, just like for simulations in conventional set-up. Most of the time, the BIRs are recorded ones (see next subsection; there exist a few BIR simulators, but we will not detail this quite specific aspect here). To take into account the room acoustics in a real-world SSL application, BIR effects are often combined with RIR effects. This is not obtained by trivially cascading the BIR and RIR filters, since the BIR depends on the source DoA, meaning that one would have to integrate it with RIR components from many incoming directions (Bernschütz, 2016). However, such a process is included in several RIR simulators, which are able to produce the corresponding combined response, called the binaural room impulse response (BRIR), e.g., (Campbell *et al.*, 2005). Recall that BIRs are often manipulated in the frequency domain (referred as HRTFs), where they are a function of both frequency and source DoA.

B. Real data

Collecting real labeled data is crucial to assessing the robustness of an SSL neural network in a real-world environment. However, it is a cumbersome task. As of today, only a few datasets of such recordings exist. Among them, several impulse response datasets are publicly available and have been used to generate training and/or testing data.

The distant-speech interaction for robust home applications (DIRHA) simulated corpus presented by Cristoforetti *et al.* (2014) has been used to simulate microphone speech signals based on real RIRs, recorded in a multi-room environment (Vecchiotti *et al.*, 2018; Vesperini *et al.*, 2016). Another database consisting of recorded RIRs from three rooms with different acoustic characteristics is publicly

available (Hadad *et al.*, 2014), using three microphone array configurations to capture signals from several source azimuth positions in the range $[-90°, 90°]$. The RIR dataset published by Fernandez-Grande *et al.* (2021) is intended to be used for DoA estimation and contains measurements from a three-channel array. Other RIR datasets have been published by, e.g., Szöke *et al.* (2019), Eaton *et al.* (2015), Hahmann *et al.* (2021a), Koyama *et al.* (2021), Kristoffersen *et al.* (2021), and Riezu and Grande (2021). The last four were initially designed for sound field analysis and synthesis, and they contain measurements from single-channel microphones (i.e., not microphone arrays). However, the acquired RIRs correspond to multiple positions within a room and could be potentially used to emulate microphone arrays.

As for BIR dataset recordings, a physical head-and-torso simulator (HATS) (aka "dummy head") is used, with ear microphones plugged into the dummy head ears. To isolate head and torso effects from other environmental effects such as reverberation, binaural recordings are generally made in an anechoic room. For example, the dataset published by Thiemann and Van De Par (2015) was collected using four different dummy heads and used for SSL by Roden *et al.* (2015).

The Surrey Binaural Room Impulse Responses database was published by Francombe (2017) and has been used for SSL by, e.g., Ma *et al.* (2015) to synthesize signals for evaluating the proposed method. This database has been recorded using a HATS in four room configurations, with sound coming from loudspeakers. It thus combines binaural effects with room effects.

Several challenges have also been organized for some years, and evaluation datasets with real recordings have been constituted to assess the candidate systems. Datasets were created for the SELD task of the DCASE Challenge, in 2019 (Adavanne *et al.*, 2019c), 2020 (Politis *et al.*, 2020a), and 2021 (Politis *et al.*, 2021). These datasets contain sound events in reverberant and noisy environments, synthesized from recordings of real RIRs. These data come in two four-microphone spatial audio formats: tetrahedral microphone array and FOA. The dataset comprises 12 sound event types, including, e.g., barking dog, female/male speech, or ringing, with up to three simultaneous events overlapping. In the 2019 dataset, the sources are static, whereas they are both static and moving in the 2020 and 2021 datasets, with more diverse acoustic conditions. Finally, in the 2021 edition of the DCASE dataset, additional sound events have been added to the recordings to play the role of (directional) interferers (that are not bound to be classified). These datasets have been used in many SSL systems (e.g., Cao *et al.*, 2019a; Cao *et al.*, 2020; Grondin *et al.*, 2019; Mazzon *et al.*, 2019; Naranjo-Alcazar *et al.*, 2020; Park *et al.*, 2019b; Shimada *et al.*, 2020b; Wang *et al.*, 2020). Very recently, another SELD challenge focused on 3D sound has been announced (Guizzo *et al.*, 2021), where a *pair* of FOA microphones was used to capture a large number of RIRs in an office room, from which the audio data were generated.

The acoustic source Localization and Tracking (LOCATA) challenge (Evers *et al.*, 2020) has been one of the most comprehensive challenges targeting the localization of speech sources. The challenge tasks include single and multiple SSL, each of which are a setting where the sources and/or microphones are static or mobile. The recordings have been made using several types of microphone arrays, namely, the planar array from Brutti *et al.* (2010), the *em32 Eigenmike* spherical array, a hearing aid, and a set of microphones mounted on a robot head. The ground truth data include position information obtained through an optical tracking system, hand-labeled VAD metadata, and dry (or close-talking) source signals. This dataset has been used in a number of works to validate the effectiveness of a proposed method on "real-life" recordings (e.g., Diaz-Guerra *et al.*, 2021b; Grumiaux *et al.*, 2021a; Pak and Shin, 2019; Sundar *et al.*, 2020; Tang *et al.*, 2019; Varanasi *et al.*, 2020; Yang *et al.*, 2021b).

A few audio-visual datasets have also been developed and are publicly available, in which the audio data are enriched with video information. This type of dataset is dedicated to the development and testing of audio-visual localization and tracking techniques, which are out of the scope of this survey paper. Among these corpora, the AV16.3 corpus (Lathoud *et al.*, 2004) and the CHIL database (Stiefelhagen *et al.*, 2007) have provided an evaluative basis for several (purely audio) SSL systems (Vera-Diaz *et al.*, 2018, 2020, 2021) by considering only the audio part of the audiovisual dataset.

Finally, we also found a series of papers in which neural networks were tested using real data specifically recorded for the presented work in the researchers' own laboratories, (e.g., Chazan *et al.*, 2019; Grumiaux *et al.*, 2021a; Grumiaux *et al.*, 2021b; He *et al.*, 2018a; He *et al.*, 2021a; Le Moing *et al.*, 2020; Nguyen *et al.*, 2020a; Perotin *et al.*, 2019a; Perotin *et al.*, 2018b; Varanasi *et al.*, 2020).

C. Data augmentation techniques

To limit the massive use of simulated data, which can limit the robustness of the network on real-world data, and to overcome the limitation in the amount of real data, several authors have proposed resorting to data augmentation techniques. Without producing more recordings, data augmentation allows for the creation of additional training examples, often leading to improved network performance.

For the DCASE Challenge, many submitted systems were trained using data augmentation techniques on the train dataset. Mazzon *et al.* (2019) proposed and evaluated three techniques to augment the training data, taking advantage of the FOA representation used by their SSL neural network: swap or inversion of FOA channels, label-oriented rotation (the rotation is applied to result in the desired label), or channel-oriented rotation (the rotation is directly applied with the desired matrix). Interestingly, the channel-oriented rotation method gave the worst results in their experiments, while the other two methods showed an improvement in

neural network performance. Zhang *et al.* (2019a) applied the SpecAugment method of Park *et al.* (2019a), which led to new data examples by masking certain time frames or frequencies of a spectrogram, or both at the same time. This method was also employed by, e.g., Bai *et al.* (2021); Krause *et al.* (2021); Shimada *et al.* (2021); Yalta *et al.* (2021). In the work of Pratik *et al.* (2019), new training material was created with the *Mixup* method of Zhang *et al.* (2018), which relies on convex combinations of an existing training data pair. Noh *et al.* (2019) used pitch shifting and block mixing data augmentation (Salamon and Bello, 2017). The techniques of Mazzon *et al.* (2019) and Zhang *et al.* (2019a) were employed by Shimada *et al.* (2021) and Shimada *et al.* (2020b) to create new mixtures, along with another data augmentation method proposed by Takahashi *et al.* (2016), which is based on random mixing of two training signals.

Wang *et al.* (2021) applied four new data augmentation techniques to the DCASE dataset (Politis *et al.*, 2021). The first one applies the benefit of the FOA format to changing the location of the sources by swapping audio channels. The second method is based on the extraction of spatial and spectral information on the sources, which are then modified and recombined to create new training examples. The third one relies on mixing multiple examples, resulting in new multi-source labelled mixtures. The fourth technique is based on random TF masking. The authors evaluated the benefits of these data augmentation methods both when used separately and when applied sequentially.

VIII. LEARNING STRATEGIES

In a general manner, when training a neural network to accomplish a certain task, one needs to choose a training paradigm that often depends on the type and amount of available data. In the DNN-based SSL literature, most of the systems rely on supervised learning, although several examples of semi-supervised and weakly supervised learning can also be found.

A. Supervised learning

When training a neural network with supervised learning, the training dataset must contain the output target (also known as the label, especially in the classification mode) for each corresponding input data. A cost function (or loss function) is used to quantify the error between the output target and the actual output of the neural network for a given input data, and training consists of minimizing the average loss function over the training dataset. We have seen in Sec. VI that in a single-source SSL scenario with the classification paradigm, a softmax output function is generally used. In that case, the cost function is generally the categorical cross-entropy (e.g., Chakrabarty and Habets, 2017a; Perotin *et al.*, 2018b; Yalta *et al.*, 2017). When dealing with multiple sources, still with the classification paradigm, sigmoid activation functions and a binary cross-entropy loss function are used (e.g., Chakrabarty and Habets, 2017b; Grumiaux *et al.*, 2021a; Perotin *et al.*, 2019b). With a regression scheme, the choice for the cost function is the mean square error in most systems, e.g., (Adavanne *et al.*, 2019a; He *et al.*, 2021a; Krause *et al.*, 2020a; Nguyen *et al.*, 2018; Pertilä and Cakir, 2017; Salvati *et al.*, 2018; Shimada *et al.*, 2020a). We also sometimes witness the use of other cost functions, such as the angular error (Perotin *et al.*, 2019a) and the ℓ_1-norm (Jenrungrot *et al.*, 2020).

The limitation of supervised training is that the training relies on a great amount of labeled training data, whereas only a few real-world datasets with limited size have been collected for SSL. These datasets are not sufficient for robust training with DL models. To cope with these issues, one can opt for a data simulation method, as seen in Sec. VII A, or data augmentation techniques, as seen in Sec. VII C. Otherwise, alternative training strategies can be employed, such as semi-supervised and weakly supervised learning, as presented hereafter.

B. Semi-supervised and weakly supervised learning

Unsupervised learning refers to model training with a dataset that does not contain labels. In the present SSL framework, this means that we would have a dataset of recorded acoustic signals without the knowledge of source position/direction, and hence unsupervised learning alone is not applicable to SSL in practice. Semi-supervised learning refers to when part of the learning is done in a supervised manner, and another part is done in an unsupervised manner. Usually, the network is pre-trained with labeled data training and refined (or fine-tuned) using unsupervised learning, i.e., without resorting to labels. In the SSL literature, semi-supervised learning has been proposed to improve the performance of the neural network on conditions unseen during supervised training or on real data, compared to its performance when trained only in the supervised manner. It can be seen as an alternative manner to enrich a labeled training dataset of too limited size or conditions (see Sec. VII).

For example, Takeda and Komatani (2017) and Takeda *et al.* (2018) adapted a pre-trained neural network to unseen conditions in a unsupervised way. For the cost function, the cross-entropy was modified to be computed only with the estimated output, so that the overall entropy was minimized. They also applied a parameter selection method dedicated to avoid overfitting, as well as early stopping. Bianco *et al.* (2020) combined supervised and unsupervised learning using a VAE-based system. A generative network was trained to infer the phase of RTFs, which were used as input features in a classifier network. The cost function directly encompasses a supervised term and an unsupervised term and, during the training, the examples can come with or without labels.

Le Moing *et al.* (2021) proposed a semi-supervised approach to adapt the network to real-world data after it was trained with a simulated dataset. This strategy was implemented with *adversarial training* (Goodfellow *et al.*, 2014). In the present SSL context, a discriminator network was

https://doi.org/10.1121/10.0011809

trained to label incoming data as synthetic or real, and the generator network learned to fool the discriminator. This enabled the adaptation of the DoA estimation network to infer from real data.

A different kind of training, named weakly supervised, was used by He *et al.* (2019a) and He *et al.* (2021a). The authors fine-tuned a pre-trained neural network by adapting the cost function to account for weak labels, which is the NoS, presumably known. This helped to improve the network performance by reducing the amount of incoherent predictions. Weak supervision was also used by Opochinsky *et al.* (2019). Under the assumption that only a few training data come with labels, a triplet loss function is computed. For each training step, three examples are drawn: a *query* sample, acting as a usual example, a *positive* sample close to the query sample, and a *negative* sample from a more remote source position. The triplet loss (named so because of these three components) is then derived so that the network learns to infer the position of the positive sample closer to the query sample than the negative sample.

IX. CONCLUSIONS AND PERSPECTIVES

In this paper, we have presented a comprehensive overview of the literature on SSL techniques based on DL methods from 2011 to 2021. We attempted to categorize the many publications in this domain according to different characteristics of the methods in terms of source (mixture) configuration, neural network architecture, input data type, output strategy, training, and test datasets, and learning strategy. Tables II–V summarize our survey: They gather the references of the reviewed DL-based SSL papers with the main characteristics of the proposed methods (the ones that were used in our taxonomy of the different methods) being reported into different columns. We believe these tables can be very useful for a quick search of methods with a given set of characteristics.

To conclude this survey paper, we can comment on some current trends and draw a series of perspectives on the future directions that would be interesting to investigate to improve the performance of SSL systems and gain a better understanding of their behavior. Note that some of these perspectives appeal to general methodological issues in deep learning that are common to many applications, and some others are more specific to SSL. Note also that this list of research directions is not meant to be exhaustive.

A. Adaptation to (limited sets of) real-world data

In a general manner, we observe a drop in performance when DNNs trained on simulated data are tested on real-world signals. This effect is well-known in the DL research in general, it is a particular case of the poor generalization capability of DNNs in the case of significant train-test data mismatch (Goodfellow *et al.*, 2016; LeCun *et al.*, 2015). We recall that this problem remains particularly crucial in SSL due to the difficulty of developing massive labeled datasets (i.e., with reliable annotations of ground-truth source location) and the use of simulated training data. This is valid for training datasets generated using the usual "shoebox" acoustic simulations. Such geometry is rarely encountered in real-world environments. Moreover, the placement of the simulated microphone array is often unrealistic (e.g., it is floating in the air, whereas a practical recording device is often positioned on a table, leading to strong reflections).

A first approach to tackle this problem is to consider more sophisticated room acoustics simulators, capable of taking into account more complex room geometries and acoustic phenomena, such as scattering or diffraction, see the related discussion in Sec. VII A. However, this presents the limitation of a heavier computation cost, which should be balanced with the amount of data to be generated. Another line of research is to progressively train the network with more and more realistic signals, e.g., first with signals generated with simulated SRIRs, then fine-tuning the network with signals generated with real SRIRs, then further fine-tuning it with recorded data. This is in line with the general methodology of domain adaptation (DA) (Kouw and Loog, 2019) and transfer learning (Bengio, 2012; Zhuang *et al.*, 2020) used in many applications of DL, which aims at improving the performance of a network on a particular domain (in our case, real-world data) after it has been trained on another domain (here, simulated data). For SSL, the idea is to "optimize" the model to the target acoustic environment and/or sound sources. DA is a promising research field on its own, and it has only recently attracted the attention of the SSL community. To our best knowledge, the adversarial approach of Le Moing *et al.* (2021) and the entropy-based adaptation of Takeda and Komatani (2017) are the only representatives of DA for SSL.

Another line of research would be to inspire from weakly-supervised SSL methods based on *manifold learning* (Laufer-Goldshtein *et al.*, 2020). The general principle is that the high-dimensional multichannel observed data live in a low-dimensional acoustic space, controlled by a limited number of latent variables (mainly, room dimensions, source and microphone positions, and reflection coefficients). This low-dimensional space, or manifold, can be identified using a large set of unlabeled data and unsupervised data dimension reduction techniques. Then a limited set of labeled data can be used to identify the relationship between observed data and source positions "in the manifold," and thus estimate the source positions from new observed data (using, e.g., interpolation techniques). This principle was largely developed by Laufer-Goldshtein *et al.* (2020), who proposed several non-deep manifold identification techniques and corresponding SSL algorithms. The same principle can be applied with a DL approach, in particular with deep latent-variable generative models such as the VAE, in the line with the semi-supervised VAE-SSL model of Bianco *et al.* (2021) and Bianco *et al.* (2020) already mentioned in Sec. IV G 2 (see also an example of weakly supervised VAE-based source-filter decomposition of speech signals by Sadok *et al.*, 2022). To our knowledge, SSL based on "deep manifold learning" is still a largely

TABLE II. Summary of DL-based SSL systems published from 2011 to 2018, organized in chronological then alphabetical order. **Type**: R, regression, C, classification. **Learning**: S, supervised, SS, semi-supervised, WS, weakly supervised. **Sources**: **NoS**, considered number of sources; **Kno.** indicates if the NoS is known or not before estimating the DoA (✓, yes, ✗, no), **Mov.** specifies if moving sources are considered. **Data**: SA, synthetic anechoic; RA, real anechoic; SR, synthetic reverberant; RR, real reverberant.

							Sources			Data							
										Train				Test			
Author	Year	Architecture	Type	Learning	Input features	Output	NoS	Kno.	Mov.	SA	RA	SR	RR	SA	RA	SR	RR
Kim and Ling (2011)	2011	MLP	R	S	Power of multiple beams	θ	1-5	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓
Tsuzuki *et al.* (2013)	2013	MLP	R	S	Time delay, phase delay, sound pressure diff.	θ	1	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗
Youssef *et al.* (2013)	2013	MLP	R	S	ILD, ITD	θ	1	✗	✗	✓	✗	✓	✗	✓	✗	✓	✓
Hirvonen (2015)	2015	CNN	C	S	Magnitude spectrograms	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Ma *et al.* (2015)	2015	MLP	C	S	Binaural cross correlation + ILD	θ	1-3	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓
Roden *et al.* (2015)	2015	MLP	C	S	ILD, ITD, binaural magnitude + phase spectrogr., binaural real + imaginary spectrograms	θ / ϕ / r	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Xiao *et al.* (2015)	2015	MLP	C	S	GCC-PHAT	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Takeda and Komatani (2016b)	2016	MLP	C	S	Complex eigenvectors from correlation matrix	θ, z, r	0-1	✓	✗	✗	✓	✗	✓	✗	✓	✗	✓
Takeda and Komatani (2016a)	2016	MLP	C	S	Complex eigenvectors from correlation matrix	θ	0-2	✗	✗	✗	✓	✗	✓	✗	✓	✗	✓
Vesperini *et al.* (2016)	2016	MLP	R	S	GCC-PHAT	x, y	1	✓	✗	✗	✗	✓	✓	✗	✗	✓	✓
Zermini *et al.* (2016)	2016	AE	C	S	Mixing vector + ILD + IPD	θ		✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Chakrabarty and Habets (2017a)	2017	CNN	C	S	Phase map	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Chakrabarty and Habets (2017b)	2017	CNN	C	S	Phase map	θ	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Pertilä and Cakir (2017)	2017	CNN	R	S	Magnitude spectrograms	TF Mask	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Takeda and Komatani (2017)	2017	MLP	C	SS	Complex eigenvectors from correlation matrix	θ, ϕ	1	✓	✗	✗	✓	✗	✓	✗	✓	✗	✓
Yalta *et al.* (2017)	2017	Res. CNN	C	S	Magnitude spectrograms	θ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Yiwere and Rhee (2017)	2017	MLP	C	S	Binaural cross correlation + ILD	θ, d	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Adavanne *et al.* (2018)	2018	CRNN	C	S	Magnitude + phase spectrograms	SPS, θ, ϕ	∞	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
He *et al.* (2018a)	2018	MLP, CNN	C	S	GCC-PHAT	θ	0-2	✗/✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
He *et al.* (2018b)	2018	Res. CNN	C	S	Real + imaginary spectrograms	θ	∞	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓
Huang *et al.* (2018)	2018	DNN	R	S	Waveforms	dry signal	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Li *et al.* (2018)	2018	CRNN	C	S	GCC-PHAT	θ	1	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗
Ma and Liu (2018)	2018	CNN	C	S	CPS	x, y	3	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗
Nguyen *et al.* (2018)	2018	CNN	R	S	ILD + IPD	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Perotin *et al.* (2018b)	2018	CRNN	C	S	Intensity	θ, ϕ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Salvati *et al.* (2018)	2018	CNN	C/R	S	Narrowband SRP components	SRP weights	1	✓	✗	?				✗	✗	✓	✓
Sivasankaran *et al.* (2018)	2018	CNN	C	S	IPD	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Suvorov *et al.* (2018)	2018	Res. CNN	C	S	Waveforms	θ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Takeda *et al.* (2018)	2018	MLP	C	SS	Complex eigenvectors from correlation matrix	θ	1	✓	✗	✗	✓	✗	✓	✗	✓	✗	✓
Thuillier *et al.* (2018)	2018	CNN	C	S	Ipsilateral + contralateral ear input signal	ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Vecchiotti *et al.* (2018)	2018	CNN	R	S	GCC-PHAT + mel spectrograms	x, y	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Vera-Diaz *et al.* (2018)	2018	CNN	R	S	Waveforms	x, y, z	1	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓

18 December 2024 13:32:40
https://doi.org/10.1121/10.0011809
JASA

JASA
https://doi.org/10.1121/10.0011809

TABLE III. Summary of DL-based SSL systems published in 2019, organized in alphabetical order. See Table I's caption for acronyms specification.

							Sources			Data							
										Train				Test			
Author	Year	Architecture	Type	Learning	Input features	Output	NoS	Kno.	Mov.	SA	RA	SR	RR	SA	RA	SR	RR
Adavanne *et al.* (2019a)	2019	CRNN	R	S	FOA magnitude + phase spectrograms	x, y, z	1	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓
Adavanne *et al.* (2019b)	2019	CRNN	R	S	FOA magnitude + phase spectrograms	x, y, z	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cao *et al.* (2019a)	2019	CRNN	R	S	Log-Mel spectrogr. + GCC-PHAT + intensity	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Cao *et al.* (2019b)	2019	CRNN	R	S	Log-Mel spectrogr. + GCC-PHAT	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Chakrabarty and Habets (2019a)	2019	CNN	C	S	Phase map	θ	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Chakrabarty and Habets (2019b)	2019	CNN	C	S	Phase map	θ	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Chazan *et al.* (2019)	2019	U-net	C	S	Phase map of the RTF between each mic pair	θ	∞	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
Chytas and Potamianos (2019)	2019	CNN	R	S	Waveforms	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Comminiello *et al.* (2019)	2019	CRNN	R	S	Quaternion FOA	x, y, z	1	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗
Grondin *et al.* (2019)	2019	CRNN	R	S	CPS + GCC-PHAT	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
He *et al.* (2019a)	2019	Res. CNN	C	WS	Real + imaginary spectrograms	θ	1-2	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
Huang *et al.* (2019)	2019	CNN	R	S	Waveforms	dry signal	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Kapka and Lewandowski (2019)	2019	CRNN	R	S	Magnitude + phase spectrograms	x, y, z	1-2	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓
Kong *et al.* (2019)	2019	CNN	R	S	Log-Mel magnitude FOA spectrograms	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Krause and Kowalczyk (2019)	2019	CRNN	R	S	Magnitude / phase spectrograms	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Küçük *et al.* (2019)	2019	CNN	C	S	Real + imaginary spectrograms	θ	1	✓	✗	✗	✗	✓	✓	✗	✗	✓	✓
Kujawski *et al.* (2019)	2019	Res. CNN	R	S	Beamforming map	x, y	1	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗
Leung and Ren (2019)	2019	CRNN	R	S	CPS + real/imag. spectro + mag./phase spectro	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Lin and Wang (2019)	2019	CRNN	C	S	Magnitude and phase spectrograms	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Lu (2019)	2019	CRNN	R	S	GCC-PHAT	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Maruri *et al.* (2019)	2019	CRNN	R	S	GCC-PHAT + magnitude + phase spectrograms	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Mazzon *et al.* (2019)	2019	CRNN	R	S	Mel-spectrograms + GCC-PHAT/intensity	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Noh *et al.* (2019)	2019	CNN	C	S	GCC-PHAT	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Nustede and Anemüller (2019)	2019	CRNN	R	S	Group delays	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Opochinsky *et al.* (2019)	2019	MLP	R	WS	RTFs	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Pak and Shin (2019)	2019	MLP	R	S	IPD	(clean) IPD		✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Pang *et al.* (2019)	2019	CNN	R	S	ILD + IPD	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Park *et al.* (2019b)	2019	CRNN	R	S	Log-Mel spectrograms + intensity	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Perotin *et al.* (2019b)	2019	CRNN	C	S	FOA pseudo-intensity	θ, ϕ	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Perotin *et al.* (2019a)	2019	CRNN	C/R	S	FOA pseudo-intensity	θ, ϕ / x, y, z	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Pratik *et al.* (2019)	2019	CRNN	R	S	GCC-PHAT + Mel/Bark spectrograms	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Pujol *et al.* (2019)	2019	Res. CNN	R	S	Waveforms	x, y	1	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
Ranjan *et al.* (2019)	2019	Res. CRNN	C	S	Log-Mel spectrograms	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Sudo *et al.* (2019)	2019	CRNN	R	S	cos(IPD), sin(IPD)	$\cos(\theta)$, $\sin(\theta)$, $\cos(\phi)$, $\sin(\phi)$	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Tang *et al.* (2019)	2019	CRNN	C/R	S	FOA pseudo-intensity	θ, ϕ / x, y, z	1	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
Vecchiotti *et al.* (2019b)	2019	CNN	R	S	GCC-PHAT + Mel-spectrograms	x, y	1	✓	✗	✗	✗	✓	✓	✗	✗	✓	✓
Vecchiotti *et al.* (2019a)	2019	CNN	C	S	Waveforms	θ	1	✓	✗	✗	✓	✗	✓	✓	✓	✗	✓

TABLE III. (*Continued.*)

Author	Year	Architecture	Type	Learning	Input features	Output	Sources			Data								
							NoS	Kno.	Mov.	Train				Test				
										SA	RA	SR	RR	SA	RA	SR	RR
Wang *et al.* (2019)	2019	RNN	R	S	Magnitude spectrograms	TF Mask	1	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
Xue *et al.* (2019)	2019	CRNN	R	S	Log-Mel spectr. + CQT + phase spectrogr. + CPS	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Zhang *et al.* (2019a)	2019	CRNN	R	S	Magnitude and phase spectrograms	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Zhang *et al.* (2019b)	2019	CNN	C	S	Phase spectrograms	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗

under-considered and open topic in the literature, yet it offers a promising direction to deal with limited annotated datasets.

B. Flexibility of the trained models

As opposed to conventional SP techniques, which can be parameterized to adapt to the changes in the system setup, DL methods for SSL generally assume identical setups for the training and the inference phase. Particularly, the number, geometrical arrangement, and the directivity of the microphones composing an array are usually assumed to be fixed. This is a serious disadvantage since the network needs to be retrained for different microphone arrays, although the task (SSL) remains the same. A partial remedy is to use array-agnostic inputs, such as Ambisonics (e.g., Adavanne *et al.*, 2018; Grumiaux *et al.*, 2021a; Perotin *et al.*, 2019b), CPS eigenvectors (e.g., Takeda and Komatani, 2016b), or spatial pseudo-spectra (e.g., Nguyen *et al.*, 2020a; Wu *et al.*, 2021a). Another possibility is to adopt array-invariant techniques from end-to-end multichannel speech enhancement (Luo *et al.*, 2020).

Moreover, one could apply transfer learning and DA techniques, discussed previously, that could enable the models trained for a particular microphone array to adapt to another. Such techniques could not only be beneficial for the changes in microphone array setups but also the changes in the input signal parameters, such as the sampling rate, frame, and overlap length, as well as the type of the STFT window function. A radical approach would be to make the method inherently independent of parameterization, by treating the input signal as a point cloud, as recently suggested by Subramani and Smaragdis (2021).

C. Multi-task learning

Multi-task training is a general methodology to improve the performance of a DNN-based system on a given task by training the model to jointly and simultaneously tackle several other tasks (Ruder, 2017; Zhang and Yang, 2021). It has been observed in practice that this often leads to better performance on the first target task. This principle is most often implemented in the following manner: An early part of the model (e.g., a common feature extraction module composed of several layers or several layer blocks) is common for the different tasks, then the model splits into different branches, each one specialized in one of the different tasks. The common part is assumed to allow the discovery of an efficient signal representation, and the fact that this representation is used for several downstream tasks somehow reinforces the efficiency of the representation extraction.

This principle can be applied to SSL. In fact, it has already been extensively illustrated in this survey with the SELD Task of the DCASE Challenge (Politis *et al.*, 2020b) and the many candidates that have been proposed to this challenge (and that we have reported in this survey). The vast majority of the candidate DNNs follow the previously established architecture, with a common feature extraction

JASA
https://doi.org/10.1121/10.0011809
18 December 2024 13:32:40

TABLE IV. Summary of DL-based SSL systems published in 2020, organized in alphabetical order. See the Table I caption for acronyms specification.

							Sources			Data							
										Train				Test			
Author	Year	Architecture	Type	Learn.	Input features	Output	NoS	Kno.	Mov.	SA	RA	SR	RR	SA	RA	SR	RR
Bianco *et al.* (2020)	2020	VAE	C	SS	RTFs	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Cao *et al.* (2020)	2020	CRNN	R	S	FOA waveforms	θ, ϕ	0-2	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓
Comanducci *et al.* (2020a)	2020	CNN/U-Net	C	S	GCC-PHAT	x, y	1	✓	✗	✓	✗	✓	✗	✗	✗	✓	✓
Comanducci *et al.* (2020b)	2020	U-Net	R	S	GCC	Clean GCC	1	✓	✗	✓	✗	✓	✗	✗	✗	✓	✓
Fahim *et al.* (2020)	2020	CNN	C	S	FOA modal coherence	θ, ϕ	1-7	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Hao *et al.* (2020)	2020	CNN	C	S	Real + imaginary spectrograms + spectral flux	θ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Huang *et al.* (2020)	2020	AE	R	S	Waveforms	θ	1	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗
Hübner *et al.* (2021)	2020	CNN	C	S	Phase map	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
Jenrungrot *et al.* (2020)	2020	U-Net	R	S	Waveforms	θ	0-8	✗	✓	✗	✗	✓	✗	✗	✗	✓	✓
Mack *et al.* (2020)	2020	CNN + attention	C	S	Phase map	θ	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Le Moing *et al.* (2020)	2020	AE	C,R	S	Real + imaginary spectrograms	x, y	1-3	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓
Le Moing *et al.* (2021)	2020	AE	C	SS	Real + imaginary spectrograms	x, y	1-3	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓
Naranjo-Alcazar *et al.* (2020)	2020	Res. CRNN	R	S	Log-Mel magnitude spectrograms + GCC-PHAT	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Nguyen *et al.* (2020a)	2020	CNN	C	S	Spatial pseudo-spectrum	θ	0-4	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓
Nguyen *et al.* (2020b)	2020	CRNN	R	S	DoAs from histogram-based method	θ, ϕ	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Nguyen *et al.* (2020c)	2020	CRNN	R	S	DoAs from histogram-based method	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Park *et al.* (2020)	2020	CRNN	R	S	Log-Mel energy + intensity	θ, ϕ	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Patel *et al.* (2020)	2020	U-Net	R	S	Mel-spectrograms	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Phan *et al.* (2020a)	2020	CRNN + SA	R	S	FOA log-Mel spectrograms + active/reactive intensity, or GCC-PHAT	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Phan *et al.* (2020b)	2020	CRNN + SA	R	S	FOA log-Mel spectrograms + active/reactive intensity, or GCC-PHAT	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Ronchini *et al.* (2020)	2020	CRNN	R	S	FOA log-Mel spectrograms + log-Mel intensity	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Sampathkumar and Kowerko (2020)	2020	CRNN	R	S	MIC + FOA Mel spectrograms + active intensity + GCC-PHAT	θ, ϕ	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Shimada *et al.* (2020a)	2020	Res. CRNN	R	S	FOA magnitude spectrograms + IPD	ACCDOA	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Shimada *et al.* (2020b)	2020	Res. CRNN	R	S	FOA magnitude spectrograms + IPD	ACCDOA	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Singla *et al.* (2020)	2020	CRNN	R	S	FOA log-Mel spectrograms + log-Mel intensity	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Song (2020)	2020	CRNN	R	S	GCC-PHAT + FOA active intensity	x, y, z	1	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓
Sundar *et al.* (2020)	2020	Res. CNN	C/R	S	Waveforms	d, θ	1-3	✗	✓	✓	✗	✓	✗	✓	✗	✓	✓
Tian (2020)	2020	CRNN	?	S	Ambisonics	?	?	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓
Varanasi *et al.* (2020)	2020	CNN	C	S	3rd spherical harmonics (phase or phase+magnitude)	θ, ϕ	1	✓	✓	✗	✗	✓	✗	✗	✗	✓	✓
Varzandeh *et al.* (2020)	2020	CNN	C	S	GCC-PHAT + periodicity degree	θ	0-1	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓
Vera-Diaz *et al.* (2020)	2020	AE	R	S	GCC-PHAT	time-delay	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Wang *et al.* (2020)	2020	Res. CRNN	R	S	FOA pseudo-intensity + FOA log-Mel spectrograms + GCC-PHAT	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Xue *et al.* (2020)	2020	CRNN	C	S	CPS + waveforms + beamforming output	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Yasuda *et al.* (2020)	2020	Res. CRNN	R	S	FOA log-Mel spectrograms + intensity	denoised IV	2	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓

TABLE V. Summary of DL-based SSL systems published in 2021, organized in alphabetical order. See the Table I caption for acronyms specification.

							Sources			Data							
										Train				Test			
Author	Year	Architecture	Type	Learn.	Input features	Output	NoS	Kno.	Mov.	SA	RA	SR	RR	SA	RA	SR	RR
Adavanne *et al.* (2021)	2021	CRNN + SA	R	S	FOA Mel spectrograms + intensity + GCC-PHAT	x,y,z	2	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Bai *et al.* (2021)	2021	Res. CRNN	R	S	Log-Mel spectrograms + intensity	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Bianco *et al.* (2021)	2021	VAE	C	SS	RTF	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Bohlender *et al.* (2021)	2021	CNN/CRNN	C	S	Phase map	θ	1-3	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
Bologni *et al.* (2021)	2021	CNN	C	S	Waveforms	θ, d	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Cao *et al.* (2021)	2021	SA	R	S	Log-Mel spectrograms + intensity	x, y, z	0-2	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓
Castellini *et al.* (2021)	2021	MLP	R	S	real + imaginary CPS	x, y	1-3	✓	✗	✓	✗	✗	✗	✗	✗	✗	✓
Diaz-Guerra *et al.* (2021b)	2021	CNN	R	S	SRP-PHAT power map	x, y, z	1	✓	✓	✗	✗	✓	✗	✗	✗	✓	✓
Emmanuel *et al.* (2021)	2021	CNN + SA	R	S	Log-spectrograms + intensity	ACCDOA	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Gelderblom *et al.* (2021)	2021	MLP	C/R	S	GCC-PHAT	θ	2	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
Gonçalves Pinto *et al.* (2021)	2021	CNN	R	S	Magnitude CPS	x, y	1-10	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
Grumiaux *et al.* (2021a)	2021	CRNN	C	S	Intensity	θ, ϕ	1-3	✓	✓	✗	✗	✓	✗	✗	✗	✓	✓
Grumiaux *et al.* (2021b)	2021	CNN + SA	C	S	Intensity	θ, ϕ	1-3	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Guirguis *et al.* (2020)	2021	TCN	R	S	Magnitude + phase spectrograms	x, y, z	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Hammer *et al.* (2021)	2021	U-net	C	S	Phase map of the RTF between each mic pair	θ	∞	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓
He *et al.* (2021a)	2021	Res. CNN	C	WS	Magnitude + phase spectrograms	θ	1-4	✓/✗	✗	✗	✗	✓	✓	✗	✗	✗	✓
He *et al.* (2021b)	2021	CNN	R	S	Waveforms	x, y, z	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Huang and Perez (2021)	2021	Res. CNN + SA	R	S	Waveforms	ACCDOA	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Komatsu *et al.* (2020)	2021	CRNN	R	S	FOA magnitude + phase spectrograms	θ, ϕ	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Krause *et al.* (2020a)	2021	CNN	R	S	Magnitude + phase spectrograms	x, y, z	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Krause *et al.* (2020b)	2021	CRNN	R	S	Misc.	θ, ϕ	1	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
Lee *et al.* (2021a)	2021	U-Net	R	S	SRP power map	x,y	1-3	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗
Lee *et al.* (2021b)	2021	CNN + attention	C	S	Log-Mel spectrograms + intensity	θ	1	✓	✓	✗	✗	✓	✗	✗	✗	✓	✓
Liu *et al.* (2021)	2021	CNN	C	S	Intensity	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Naranjo-Alcazar *et al.* (2021)	2021	Res. CRNN	R	S	Log-Mel spectrograms + GCC-PHAT	ACCDOA	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Nguyen *et al.* (2021a)	2021	CRNN	C	S	Intensity/GCC-PHAT	θ, ϕ	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Nguyen *et al.* (2021b)	2021	CNN + RNN/SA	R	S	Log-spectrograms + DRR + SCM eigenvectors	ACCDOA	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Park *et al.* (2021a)	2021	SA	R	S	log-Mel spectrograms + intensity	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Poschadel *et al.* (2021a)	2021	CRNN	C	S	HOA magnitude + phase spectrograms	θ, ϕ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Poschadel *et al.* (2021b)	2021	CRNN	C	S	HOA magnitude + phase spectrograms	θ, ϕ	2-3	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
Pujol *et al.* (2021)	2021	Res. CNN	R	S	Waveforms	θ, ϕ	1	✓	✗	✗	✗	✓	✗	✗	✓	✓	✓
Rho *et al.* (2021)	2021	CRNN + SA	R	S	Log-Mel spectrograms + intensity	θ, ϕ	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Schymura *et al.* (2021)	2021	CNN + SA	R	S	Magnitude + phase spectrograms	θ, ϕ	1	✓	✗	✓	✗	✓	✓	✓	✗	✓	✓
Schymura *et al.* (2020)	2021	CNN + AE + attent.	R	S	FOA magnitude + phase spectrograms	θ, ϕ	1	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓
Shimada *et al.* (2021)	2021	Res. CRNN + SA	R	S	IPD	ACCDOA	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Subramanian *et al.* (2021a)	2021	CRNN	C/R	S	Phase spectrogram	θ	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Subramanian *et al.* (2021b)	2021	CRNN	C		Phase spectrograms, IPD	θ	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Sudarsanam *et al.* (2021)	2021	SA	R	S	Log-Mel spectrograms + intensity	ACCDOA	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓

https://doi.org/10.1121/10.0011809 18 December 2024 13:32:40

https://doi.org/10.1121/10.0011809

TABLE V. (*Continued.*)

Author	Year	Architecture	Type	Learn.	Input features	Output	Sources: NoS	Sources: Kno.	Sources: Mov.	Data: Train: SA	Data: Train: RA	Data: Train: SR	Data: Train: RR	Data: Test: SA	Data: Test: RA	Data: Test: SR	Data: Test: RR
Vargas *et al.* (2021)	2021	CNN	C	S	Phase map	θ	1	✓	✗	✗	✗	✓	✓	✗	✗	✓	✓
Vera-Diaz *et al.* (2021)	2021	AE	R	S	GCC-PHAT	time-delay	2	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Wang *et al.* (2021)	2021	SA	R	S	Mel-spectr. + intensity/Mel-spectr. + GCC-PHAT	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Wu *et al.* (2021b)	2021	AE	R	S	Likelihood surface	x, y	1	✓	✗	✗	✗	✓	✓	✗	✗	✓	✓
Wu *et al.* (2021a)	2021	CNN AE	R	S	Beamforming heatmap image	x, y	1	✓	✗	✗	✗	✓	✓	✗	✗	✓	✓
Xinghao *et al.* (2021)	2021	CNN + SA	R	S	Log-Mel spectrograms + intensity	ACCDOA	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Xu *et al.* (2021a)	2021	DenseNet	R	S	Real CPS	x, y	6-25	✗/✓	✗	✓	✗	✗	✗	✗	✗	✓	✗
Yalta *et al.* (2021)	2021	SA	R	S	Log-Mel spectrograms + intensity	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
Yang *et al.* (2021a)	2021	CRNN	C	S	Log-magnitude and phase spectrograms	θ	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Yang *et al.* (2021b)	2021	CRNN	C	S	Log-magnitude and phase spectrograms	θ	1	✓	✓	✗	✗	✓	✗	✗	✗	✓	✓
Zhang *et al.* (2021)	2021	CNN + SA	R	S	Log-spectrograms + intensity + GCC-PHAT	x, y, z	1	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓

module followed by two SED and SSL branches. In 2021, the ACCDOA representation was adopted by many researchers, see Sec. VI B 2, and allowed for a joint SED and SSL process up to the very last model layer. We believe that combining the SSL task with other tasks (alternately to SED or in addition to it) such as source separation or ASR could lead to further advances. For example, jointly proceeding to source counting in addition to SSL in the work of Grumiaux *et al.* (2021a) was shown to improve the SSL performance (note that here source counting is explicit and high-resolution, i.e., it consists in estimating the number of active sources at the short-term frame level, whereas it is most often implicit and generally made on a much larger timescale in the SELD task of the DCASE Challenge). Other examples of multi-task learning for SSL can be found in the works of Wu *et al.* (2021a,b). Combining SSL with source separation in a DL framework is further discussed later.

Somewhat different from multi-task approaches, the end-to-end *task-oriented* learning of the entire processing chain (stacked localization, DoA-parameterized beamforming, and ASR blocks) of Subramanian *et al.* (2021a) represents a refreshing idea to address the lack of DoA-annotated data. For instance, by using pre-trained ASR blocks, and by "freezing" all but the localization part of the system during training, one could use the abundant labeled speech corpora as proxy information for the localization task. Such systems could incorporate both neural network modules and processing blocks based on conventional SP, as discussed in the next subsection.

D. Combination of DL and conventional SP techniques

In this survey paper, we have seen how the DL-based data-driven approach to the SSL problem has somehow replaced the conventional SP approach over the last decade. Yet, conventional methods are able to "explicitly" exploit strong prior knowledge of the physical underlying processes *via* signal and propagation models, whereas the exploitation of the spatial information contained in the mixture signal is done mostly "implicitly" by DNNs. Therefore, a major perspective for SSL is to get the best of both worlds, i.e., the combination of DL with conventional multichannel SP techniques.

This can be inspired by what has been done in, e.g., speech enhancement and speech/audio source separation. In the single-channel configuration, DL-based speech enhancement and separation are mostly based on the masking approach in the TF domain. Binary masks or soft masks (reminiscent of the well-known single-channel Wiener filter) are estimated with DNNs from the noisy signal and applied to it to obtain a cleaned version, see the review by Wang and Chen (2018). For multichannel speech enhancement and separation, a straightforward approach is to input the multichannel signal in the mask estimation network. However, more clever strategies can be elaborated. For example, Erdogan *et al.* (2016), Heymann *et al.* (2016), and Higuchi

et al. (2017) proposed combining the DNN-based single-channel masking with beamforming techniques (Van Veen and Buckley, 1988). In these works, the TF-domain masks estimated by a DNN are used to select speech-dominant against noise-dominant TF points, which are then used to estimate speech and noise spatial covariance matrices, respectively, which are finally used to build beamforming filters. These papers report better ASR scores than with direct TF masking or basic beamforming applied separately. This approach was extended by Perotin *et al.* (2018a) with an additional first stage of beamforming in the HOA domain to improve the mask estimation. Joint end-to-end optimization of the mask estimator, the beamformer, and possibly an ASR acoustic model, was considered in the TF domain by Meng *et al.* (2017) and Heymann *et al.* (2017), and in the time domain by Li *et al.* (2016d). Closer to source separation than to beamforming, Nugraha *et al.* (2016) combined a DNN trained to estimate a clean speech spectrogram from a noisy speech spectrogram with the source separation technique based on the spatial covariance matrix (SCM) model and Wiener filtering of Duong *et al.* (2010). Leglaive *et al.* (2019) proposed an unsupervised multichannel speech enhancement system combining a VAE for modeling the (single-channel) clean speech signal and the SCM model for modeling the spatial characteristics of the multi-channel signal.

Although we can find many examples of a combination of DL-based and SP-based approaches for beamforming and source separation, to our knowledge and as shown by our survey, this principle has been poorly applied to SSL so far. Yet, powerful deep models, and in particular deep generative models such as GANs (Goodfellow *et al.*, 2014), VAEs (Kingma and Welling, 2014), and dynamical VAEs (Girin *et al.*, 2021) are now available to model the temporal and/or spectral characteristics of sounds and can be combined with SP-based models. Morevover, as already mentioned earlier in this survey, the connection between audio source separation, diarization, and SSL is strong, reciprocal (each task can help to solve the other ones), and is already exploited in many conventional systems (Gannot *et al.*, 2017; Vincent *et al.*, 2018). Future works may thus consider jointly sound source localization, diarization, and separation/enhancement in a hybrid approach combining powerful DL models and conventional SP techniques. General frameworks for the joint optimization of DNN parameters and "conventional" parameters are now established and can be exploited (Engel *et al.*, 2020; Shlezinger *et al.*, 2020).

E. Moving sources and deep tracking

In this survey, we poorly considered the case of moving sound sources and the necessity to rely in this case on tracking algorithms. These algorithms take as input the results of SSL obtained individually on each time frame and connect them through time. This is generally based on the use of a model of the source dynamics. In the multi-source case, dynamical models are often combined with source appearance models (which would model the sound texture or the different speakers' voices in the case of audio signals), resulting in the formation of source tracks with a consistent source "identity" for each of these tracks. Tracking algorithms also estimate the tracks "birth" and "death," i.e., the time at which the corresponding sources are activated or inactivated. Such multi-object tracking (MOT) algorithms have a long history and their detailed description is beyond the scope of this paper; for a good overview of this domain, see the review papers of Vo *et al.* (2015) and Luo *et al.* (2021).

More recently, deep approaches to the MOT problem have emerged, an evolution mostly driven by the computer vision community (Ciaparrone *et al.*, 2020). For example, RNNs have been used in place of the traditional Kalman filter to model object dynamics for MOT in videos (e.g., Babaee *et al.*, 2018; Liang and Zhou, 2018; Sadeghian *et al.*, 2017; Saleh *et al.*, 2021; Xiang *et al.*, 2019). The current trend is to replace RNNs with Transformer-like models, as discussed at the end of Sec. IV F, (e.g., Meinhardt *et al.*, 2021; Sun *et al.*, 2020; Xu *et al.*, 2021b). The combination of a deep appearance model (automatic speaker recognition, SED) with a deep dynamical model in a sound source tracking system is a largely open problem and certainly a key ingredient for future developments in robust multi-source acoustic scene analysis in adverse acoustic environments and complex scenarios. Given the problem of annotated data scarcity in SSL, DL-based sound source localization and tracking may take inspiration from the unsupervised deep approaches to the MOT problem recently proposed by several researchers (e.g., Crawford and Pineau, 2020; He *et al.*, 2019b; Karthik *et al.*, 2020; Lin *et al.*, 2022; Luiten *et al.*, 2020).

ACKNOWLEDGMENTS

This work was funded by the French Association for Technological Research (ANRT CIFRE contract 2019/0533) and partially funded by the Multidisciplinary Institute in Artificial Intelligence MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

[1]Note that this problem is common to DL-based multi-source SSL methods and conventional methods for which a source activity profile is estimated and peak-picking algorithms are typically used to select the active sources.

[2]In practice, the spatial response of Ambisonic microphones is approximately frequency-independent only within certain bandwidth (dictated by the HOA order), due to spatial aliasing in the high frequency range, and noise amplification at lower frequencies (Zotter and Frank, 2019).

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (**2019a**). "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," IEEE J. Sel. Top. Signal Process. **13**(1), 34–48.

Adavanne, S., Politis, A., and Virtanen, T. (**2018**). "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, September 3–7, Rome, Italy.

Adavanne, S., Politis, A., and Virtanen, T. (**2019b**). "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," in *Proceedings of the Detection and*

Classification of Acoustic Scenes and Events Workshop (DCASE Workshop), New York, NY.
Adavanne, S., Politis, A., and Virtanen, T. (**2019c**). "A multi-room reverberant dataset for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, New York, NY.
Adavanne, S., Politis, A., and Virtanen, T. (**2021**). "Differentiable tracking-based training of deep learning sound source localizers," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 17–20, New Paltz, NY, pp. 211–215.
Ahmad, M., Muaz, M., and Adeel, M. (**2021**). "A survey of deep neural network in acoustic direction finding," in *Proceedings of the IEEE International Conference on Digital Futures and Transformative Technology (ICoDT2)*, May 24–26, Islamabad, Pakistan.
Allen, J. B., and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**(4), 943–950.
Amengual Garí, S. V., Lachenmayr, W., and Mommertz, E. (**2017**). "Spatial analysis and auralization of room acoustics using a tetrahedral microphone," J. Acoust. Soc. Am. **141**(4), EL369–EL374.
Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (**2012**). "Speaker diarization: A review of recent research," IEEE Trans. Audio. Speech. Lang. Process. **20**(2), 356–370.
Arberet, S., Gribonval, R., and Bimbot, F. (**2009**). "A robust method to count and locate audio sources in a multichannel underdetermined mixture," IEEE Trans. Signal Process. **58**(1), 121–133.
Argentieri, S., Danes, P., and Souères, P. (**2015**). "A survey on sound source localization in robotics: From binaural to array processing methods," Comput. Speech Lang. **34**(1), 87–112.
Babaee, M., Li, Z., and Rigoll, G. (**2018**). "Occlusion handling in tracking multiple people using RNN," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, October 7–10, Athens, Greece, pp. 2715–2719.
Bahdanau, D., Cho, K., and Bengio, Y. (**2016**). "Neural machine translation by jointly learning to align and translate," arXiv:1409.0473.
Bai, S., Kolter, J. Z., and Koltun, V. (**2019**). "Trellis networks for sequence modeling," arXiv:1810.06682.
Bai, J., Pu, Z., and Chen, J. (**2021**). "DCASE 2021 Task 3: SELD system based on Resnet and random segment augmentation," Technical Report, DCASE 2021 Challenge.
Ban, Y., Li, X., Alameda-Pineda, X., Girin, L., and Horaud, R. (**2018**). "Accounting for room acoustics in audio-visual multi-speaker tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 15–20, Alberta, Canada, pp. 6553–6557.
Basten, T., de Bree, H., and Sadasivan, S. (**2008**). "Acoustic eyes: A novel sound source localization and monitoring technique with 3D sound probes," in *Proceedings of the International Conference on Noise Vibration Engineering (ISMA)*, September 7–9, Leuven, Belgium.
Benesty, J., Chen, J., and Huang, Y. (**2008**). *Microphone Array Signal Processing* (Springer Science & Business Media, New York).
Bengio, Y. (**2012**). "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of the ICML Workshop Unsupervised & Transfer Learning*, July 2, Bellevue, WA, pp. 17–36.
Bengio, Y., Courville, A., and Vincent, P. (**2013**). "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828.
Bernschütz, B. (**2016**). "Microphone arrays and sound field decomposition for dynamic binaural recording," Ph.D. thesis, Technische Universitaet Berlin, Berlin, Germany.
Bialer, O., Garnett, N., and Tirer, T. (**2019**). "Performance advantages of deep neural networks for angle of arrival estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 12–17, Brighton, UK, pp. 3907–3911.
Bianchi, L., Antonacci, F., Sarti, A., and Tubaro, S. (**2016**). "The ray space transform: A new framework for wave field processing," IEEE Trans. Signal Process. **64**(21), 5696–5706.
Bianco, M. J., Gannot, S., Fernandez-Grande, E., and Gerstoft, P. (**2021**). "Semi-supervised source localization in reverberant environments with deep generative modeling," IEEE Access **9**, 84956–84970.
Bianco, M. J., Gannot, S., and Gerstoft, P. (**2020**). "Semi-supervised source localization with deep generative modeling," in *Proceedings of the MLSP*, September 21–24, Eespo, Finland (virtual conference).
Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., and Deledalle, C.-A. (**2019**). "Machine learning in acoustics: Theory and applications," J. Acoust. Soc. Am. **146**(5), 3590–3628.
Blandin, C., Ozerov, A., and Vincent, E. (**2012**). "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," Signal Process. **92**(8), 1950–1960.
Bohlender, A., Spriet, A., Tirry, W., and Madhu, N. (**2021**). "Exploiting temporal context in CNN based multisource DoA estimation," IEEE/ACM Trans. Audio. Speech. Lang. Process. **29**, 1594–1608.
Bologni, G., Heusdens, R., and Martinez, J. (**2021**). "Acoustic reflectors localization from stereo recordings using neural networks," in *Proceedings of the ICASSP*, June 6–11, Toronto, Canada (virtual conference), pp. 1–5.
Bouatouch, K., Deille, O., Maillard, J., Martin, J., and Noé, N. (**2006**). "Real time acoustic rendering of complex environments including diffraction and curved surfaces," in *Proceedings of the Audio Engineering Society (AES) Convention*, May 20–23, Paris, France.
Brandstein, M., and Ward, D. (**2001**). *Microphone Arrays: Signal Processing Techniques and Applications* (Springer Science & Business Media, New York).
Brutti, A., Cristoforetti, L., Kellermann, W., Marquardt, L., and Omologo, M. (**2010**). "WOZ acoustic data collection for interactive TV," Lang. Resour. Eval. **44**(3), 205–219.
Bush, D., and Xiang, N. (**2018**). "A model-based Bayesian framework for sound source enumeration and direction of arrival estimation using a coprime microphone array," J. Acoust. Soc. Am. **143**(6), 3934–3945.
Campbell, D., Palomaki, K., and Brown, G. (**2005**). "A Matlab simulation of shoebox room acoustics for use in research and teaching," Comput. Inform. Syst. **9**(3), 48.
Candes, E. J., Romberg, J. K., and Tao, T. (**2006**). "Stable signal recovery from incomplete and inaccurate measurements," Comm. Pure Appl. Math. **59**(8), 1207–1223.
Cao, Y., Iqbal, T., Kong, Q., An, F., Wang, W., and Plumbley, M. D. (**2021**). "An improved event-independent network for polyphonic sound event localization and detection," in *Proceedings of the ICASSP*, June 6–11, Toronto, Canada (virtual conference).
Cao, Y., Iqbal, T., Kong, Q., Galindo, M. B., Wang, W., and Plumbley, M. D. (**2019a**). "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," Technical Report, DCASE 2019 Challenge.
Cao, Y., Iqbal, T., Kong, Q., Zhong, Y., Wang, W., and Plumbley, M. D. (**2020**). "Event-independent network for polyphonic sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, November 2–4, Tokyo, Japan.
Cao, Y., Kong, Q., Iqbal, T., An, F., Wang, W., and Plumbley, M. (**2019b**). "Polyphonic sound event detection and localization using a two-stage strategy," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, October 25–26, New York, NY.
Castellini, P., Giulietti, N., Falcionelli, N., Dragoni, A. F., and Chiariotti, P. (**2021**). "A neural network based microphone array approach to grid-less noise source localization," Appl. Acoust. **177**, 107947.
Chakrabarty, S., and Habets, E. A. P. (**2017a**). "Broadband DoA estimation using convolutional neural networks trained with noise signals," in *Proceedings of the WASPAA*, October 19, New Paltz, NY, pp. 136–140.
Chakrabarty, S., and Habets, E. A. P. (**2017b**). "Multi-speaker localization using convolutional neural network trained with noise," arXiv:1712.04276.
Chakrabarty, S., and Habets, E. A. P. (**2019a**). "Multi-scale aggregation of phase information for reducing computational cost of CNN based DoA estimation," in *Proceedings of EUSIPCO*, September 2–6, A Coruña, Spain.
Chakrabarty, S., and Habets, E. A. P. (**2019b**). "Multi-speaker DoA estimation using deep convolutional networks trained with noise signals," IEEE J. Sel. Top. Signal Process. **13**(1), 8–21.
Chang, S.-Y., Li, B., Simko, G., Sainath, T. N., Tripathi, A., van den Oord, A., and Vinyals, O. (**2018**). "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proceedings of the ICASSP*, April 15–20, Calgary, Canada, pp. 5549–5553.
Chardon, G., and Daudet, L. (**2012**). "Narrowband source localization in an unknown reverberant environment using wavefield sparse decomposition," in *Proceedings of the ICASSP*, March 27–29, Kyoto, Japan, pp. 9–12.

Chazan, S. E., Hammer, H., Hazan, G., Goldberger, J., and Gannot, S. (**2019**). "Multi-microphone speaker separation based on deep DoA estimation," in *Proceedings of EUSIPCO*, September 2–6, A Coruña, Spain.
Chiariotti, P., Martarelli, M., and Castellini, P. (**2019**). "Acoustic beamforming for noise source localization – Reviews, methodology and applications," Mech. Syst. Signal Process. **120**, 422–448.
Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (**2014**). "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv:1406.1078.
Choi, J., and Chang, J.-H. (**2020**). "Convolutional neural network-based DoA estimation using stereo microphones for drone," in *Proceedings of the ICEIC*, January 19–22, Barcelona, Spain, pp. 1–5.
Chollet, F. (**2017**). *Deep Learning with Python* (Simon and Schuster, New York).
Chytas, S. P., and Potamianos, G. (**2019**). "Hierarchical detection of sound events and their localization using convolutional neural networks with adaptive thresholds," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, October 25–26, New York, NY.
Ciaparrone, G., Luque Sánchez, F., Tabik, S., Troiano, L., Tagliaferri, R., and Herrera, F. (**2020**). "Deep learning in video multi-object tracking: A survey," Neurocomputing **381**, 61–88.
Cobos, M., Antonacci, F., Alexandridis, A., Mouchtaris, A., and Lee, B. (**2017**). "A survey of sound source localization methods in wireless acoustic sensor networks," Wireless Comm. Mobile Comput. **2017**, 1–24.
Cohen, I. (**2004**). "Relative transfer function identification using speech signals," IEEE Trans. Speech Audio Process. **12**(5), 451–459.
Cohen, T., Weiler, M., Kicanaoglu, B., and Welling, M. (**2019**). "Gauge equivariant convolutional networks and the icosahedral CNN," in *Proceedings of the International Conference on Machine Learning*, June 9–15, Long Beach, CA, pp. 1321–1330.
Comanducci, L., Borra, F., Bestagini, P., Antonacci, F., Tubaro, S., and Sarti, A. (**2020a**). "Source localization using distributed microphones in reverberant environments based on deep learning and ray space transform," IEEE/ACM Trans. Audio. Speech. Lang. Process. **28**, 2238–2251.
Comanducci, L., Cobos, M., Antonacci, F., and Sarti, A. (**2020b**). "Time difference of arrival estimation from frequency-sliding generalized cross-correlations using convolutional neural networks," in *Proceedings of the ICASSP*, May 4–8, Barcelona, Spain (virtual conference), pp. 4945–4949.
Comminiello, D., Lella, M., Scardapane, S., and Uncini, A. (**2019**). "Quaternion convolutional neural networks for detection and localization of 3D sound events," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 12–17, Brighton, UK.
Crawford, E., and Pineau, J. (**2020**). "Exploiting spatial invariance for scalable unsupervised object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, February 7–12, New York, NY.
Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., and Abad, A. (**2014**). "The DIRHA simulated corpus," in *Proceedings of the LREC*, May 26–31, Reykjavik, Iceland, pp. 2629–2634.
Daniel, J., and Kitić, S. (**2020**). "Time-domain velocity vector for retracing the multipath propagation," in *Proceedings of the ICASSP*, May 4–8, Barcelona, Spain (virtual conference), pp. 421–425.
Datum, M. S., Palmieri, F., and Moiseff, A. (**1996**). "An artificial neural network for sound localization using binaural cues," J. Acoust. Soc. Am. **100**(1), 372–383.
de Bree, H.-E. (**2003**). "An overview of microflown technologies," Acta Acust. united Ac. **89**(1), 163–172.
DCASE Community (**2022**). "The DCASE Workshop and Challenge website," available at https://dcase.community/ (Last viewed June 27, 2022).
Deleforge, A., Forbes, F., and Horaud, R. (**2013**). "Variational EM for binaural sound-source separation and localization," in *Proceedings of the ICASSP*, May 26–31, Vancouver, Canada.
Deleforge, A., and Horaud, R. (**2012**). "2D sound-source localization on the binaural manifold," in *Proceedings of the MLSP*, September 23–26, Santander, Spain, pp. 1–6.
Deleforge, A., Horaud, R., Schechner, Y. Y., and Girin, L. (**2015**). "Co-localization of audio sources in images using binaural features and locally-linear regression," IEEE/ACM Trans. Audio. Speech. Lang. Process. **23**(4), 718–731.
Diaz-Guerra, D., Miguel, A., and Beltran, J. R. (**2021a**). "gpuRIR: A python library for room impulse response simulation with GPU acceleration," Multimed. Tools Appl. **80**(4), 5653–5671.
Diaz-Guerra, D., Miguel, A., and Beltran, J. R. (**2021b**). "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," IEEE/ACM Trans. Audio. Speech. Lang. Process. **29**, 300–311.
DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. (**2001**). "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, edited by M. Brandstein and D. Ward (Springer, Berlin), pp. 157–180.
Dmochowski, J. P., Benesty, J., and Affes, S. (**2007**). "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *Proceedings of WASPAA*, October 21–24, New Paltz, NY, pp. 18–21.
Dorfan, Y., and Gannot, S. (**2015**). "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," IEEE/ACM Trans. Audio. Speech. Lang. Process. **23**(10), 1692–1703.
Duong, N. Q., Vincent, E., and Gribonval, R. (**2010**). "Under-determined reverberant audio source separation using a full-rank spatial covariance model," IEEE Trans. Audio. Speech. Lang. Process. **18**(7), 1830–1840.
Dvorkind, T. G., and Gannot, S. (**2005**). "Time difference of arrival estimation of speech source in a noisy and reverberant environment," Signal Process **85**(1), 177–204.
Eaton, J., Gaubitch, N. D., Moore, A. H., and Naylor, P. A. (**2015**). "The ACE challenge "Corpus description and performance evaluation," in *Proceedings of WASPAA*, October 18–21, New Paltz, NY, pp. 1–5.
Elbir, A. M. (**2020**). "DeepMUSIC: Multiple signal classification via deep learning," IEEE Sens. Lett. **4**(4), 1–4.
El Zooghby, A., Christodoulou, C., and Georgiopoulos, M. (**2000**). "A neural network-based smart antenna for multiple source tracking," IEEE Trans. Antennas Propagat. **48**(5), 768–776.
Emmanuel, P., Parrish, N., and Horton, M. (**2021**). "Multi-scale network for sound event localization and detection," Technical Report, DCASE 2021 Challenge.
Engel, J., Hantrakul, L., Gu, C., and Roberts, A. (**2020**). "DDSP: Differentiable digital signal processing," arXiv:2001.04643.
Erdogan, H., Hershey, J. R., Watanabe, S., Mandel, M., and Le Roux, J. (**2016**). "Improved MVDR beamforming using single-channel mask prediction networks," in *Proceedings of Interspeech*, September 8–12, San Francisco, CA.
Escolano, J., Xiang, N., Perez-Lorenzo, J. M., Cobos, M., and Lopez, J. J. (**2014**). "A Bayesian direction-of-arrival model for an undetermined number of sources using a two-microphone array," J. Acoust. Soc. Am. **135**(2), 742–753.
Evers, C., Löllmann, H. W., Mellmann, H., Schmidt, A., Barfuss, H., Naylor, P. A., and Kellermann, W. (**2020**). "The LOCATA challenge: Acoustic source localization and tracking," IEEE/ACM Trans. Audio. Speech. Lang. Process. **28**, 1620–1643.
Evers, C., Moore, A. H., and Naylor, P. A. (**2014**). "Multiple source localisation in the spherical harmonic domain," in *Proceedings of IWAENC*, September 8–11, Antibes, France, pp. 258–262.
Fahim, A., Samarasinghe, P. N., and Abhayapala, T. D. (**2020**). "Multi-source DoA estimation through pattern recognition of the modal coherence of a reverberant soundfield," IEEE/ACM Trans. Audio. Speech. Lang. Process. **28**, 605–618.
Falong, L., Hongbing, J., and Xiaopeng, Z. (**1993**). "The ML bearing estimation by using neural networks," J. Electron. (China) **10**(1), 1–8.
Fernandez-Grande, E., Bianco, M. J., Gannot, S., and Gerstoft, P. (**2021**). "DTU three-channel room impulse response dataset for direction of arrival estimation 2020," available at https://ieee-dataport.org/open-access/dtu-three-channel-room-impulse-response-dataset-direction-arrival-estimation-2020 (Last viewed June 27, 2022).
Fortunati, S., Grasso, R., Gini, F., Greco, M. S., and LePage, K. (**2014**). "Single-snapshot DOA estimation by using compressed sensing," EURASIP J. Adv. Signal Process. **2014**(1), 1–17.
Foucart, S., and Rauhut, H. (**2013**). "An invitation to compressive sensing," in *A Mathematical Introduction to Compressive Sensing* (Springer, New York), pp. 1–39.
Francombe, J. (**2017**). *IoSR Listening Room Multichannel BRIR Dataset* (University of Surrey, Surrey, UK).
Gannot, S., Haardt, M., Kellermann, W., and Willett, P. (**2019**). "Introduction to the issue on acoustic source localization and tracking in dynamic real-life scenes," IEEE J. Sel. Top. Signal Process. **13**(1), 3–7.

Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (**2017**). "A consolidated perspective on multimicrophone speech enhancement and source separation," IEEE/ACM Trans. Audio. Speech. Lang. Process. **25**(4), 692–730.
Garofolo, J., Graff, D., Paul, D., and Pallett, D. (**1993a**). "CSR-I (WSJ0) Sennheiser LDC93S6B," Linguistic Data Consortium, Philadelphia, PA, https://catalog.ldc.upenn.edu/LDC93S6B (Last viewed June 27, 2022).
Garofolo, J. S., Lamel, L., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (**1993b**). "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, PA, https://catalog.ldc.upenn.edu/LDC93s1 (Last viewed June 27, 2022).
Gelderblom, F. B., Liu, Y., Kvam, J., and Myrvoll, T. A. (**2021**). "Synthetic data for DNN-based DoA estimation of indoor speech," in *Proceedings of the ICASSP*, June 6–11, Toronto, Canada (virtual conference).
Gerstoft, P., Mecklenbräuker, C. F., Seong, W., and Bianco, M. (**2018**). "Introduction to compressive sensing in acoustics," J. Acoust. Soc. Am. **143**(6), 3731–3736.
Gerstoft, P., Mecklenbräuker, C. F., Xenaki, A., and Nannuru, S. (**2016**). "Multisnapshot sparse Bayesian learning for DOA," IEEE Signal Process. Lett. **23**(10), 1469–1473.
Gerzon, M. A. (**1992**). "General metatheory of auditory localisation," in *Proceedings of the Audio Engineering Society (AES) Convention*, March 24–27, Vienna, Austria.
Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., and Alameda-Pineda, X. (**2021**). "Dynamical variational autoencoders: A comprehensive review," FNT Mach. Learn. **15**(1–2), 1–175.
Gonçalves Pinto, W., Bauerheim, M., and Parisot-Dupuis, H. (**2021**). "Deconvoluting acoustic beamforming maps with a deep neural network," in *Proceedings of the Inter-Noise Conference*, August 1–4, Washington, DC, pp. 5397–5408.
Goodfellow, I., Bengio, Y., and Courville, A. (**2016**). *Deep Learning* (MIT Press, Cambridge, MA).
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (**2014**). "Generative Adversarial Nets," in *Proceedings of NIPS*, December 8–13, Montréal, Canada.
Goryn, D., and Kaveh, M. (**1988**). "Neural networks for narrowband and wideband direction finding," in *Proceedings of ICASSP*, April 11–14, New-York, NY, pp. 2164–2167.
Grondin, F., Sobieraj, I., Plumbley, M., and Glass, J. (**2019**). "Sound event localization and detection using CRNN on pairs of microphones," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, October 25–26, New York, pp. 84–88.
Grumiaux, P.-A., Kitic, S., Girin, L., and Guerin, A. (**2020**). "High-resolution speaker counting in reverberant rooms using CRNN with Ambisonics features," in *Proceedings of the EUSIPCO*, January 18–22, Amsterdam, The Netherlands (virtual conference).
Grumiaux, P.-A., Kitic, S., Girin, L., and Guérin, A. (**2021a**). "Improved feature extraction for CRNN-based multiple sound source localization," in *Proceedings of the EUSIPCO*, August 23–27, Dublin, Ireland (virtual conference).
Grumiaux, P.-A., Kitic, S., Srivastava, P., Girin, L., and Guérin, A. (**2021b**). "SALADnet: Self-attentive multisource localization in the Ambisonics domain," in *Proceedings of WASPAA*, October 17–20, New Paltz, NY (virtual conference).
Guirguis, K., Schorn, C., Guntoro, A., Abdulatif, S., and Yang, B. (**2020**). "SELD-TCN: Sound event localization & detection via temporal convolutional networks," in *Proceedings of the EUSIPCO*, January 18–22, Amsterdam, The Netherlands (virtual conference).
Guizzo, E., Gramaccioni, R. F., Jamili, S., Marinoni, C., Massaro, E., Medaglia, C., Nachira, G., Nucciarelli, L., Paglialunga, L., Pennese, M., Pepe, S., Rocchi, E., Uncini, A., and Comminiello, D., "L3DAS21 Challenge: Machine learning for 3D audio signal processing," in IEEE International Workshop on Machine Learning for Signal Processing (MLSP), October 2021, Gold Coast, Queensland, Australia, pp. 1–6.
Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (**2020**). "Conformer: Convolution-augmented Transformer for speech recognition," in *Proceedings of Interspeech*, October 25–29, Shanghai, China (virtual conference), pp. 5036–5040.
Habets, E. A. P. (**2006**). "Room impulse response generator," Technical Report.
Habets, E. A. P. (**2022**). "Signal generator" https://github.com/ehabets/Signal-Generator/ (Last viewed March 31, 2022).
Hadad, E., Heese, F., Vary, P., and Gannot, S. (**2014**). "Multichannel audio database in various acoustic environments," in *Proceedings of IWAENC*, September 8–11, Antibes, France, pp. 313–317.
Hahmann, M., Verburg, S., and Fernandez-Grande, E. (**2021a**). "Acoustic frequency responses of an empty cuboid room," https://data.dtu.dk/articles/dataset/Acoustic_frequency_responses_of_an_empty_cuboid_room/13315289 (Last viewed June 27, 2022).
Hahmann, M., Verburg, S. A., and Fernandez-Grande, E. (**2021b**). "Spatial reconstruction of sound fields using local and data-driven functions," J. Acoust. Soc. Am. **150**(6), 4417–4428.
Hammer, H., Chazan, S. E., Goldberger, J., and Gannot, S. (**2021**). "Dynamically localizing multiple speakers based on the time-frequency domain," EURASIP J. Audio Speech Music Process. **2021**(1), 1–10.
Hao, Y., Küçük, A., Ganguly, A., and Panahi, I. M. S. (**2020**). "Spectral flux-based convolutional neural network architecture for speech source localization and its real-time implementation," IEEE Access **8**, 197047–197058.
Hübner, F., Mack, W., and Habets, E. A. P. (**2021**). "Efficient training data generation for phase-based DoA estimation," in *Proceedings of the ICASSP*, June 6–11, Toronto, Canada (virtual conference).
He, W., Motlicek, P., and Odobez, J.-M. (**2018a**). "Deep neural networks for multiple speaker detection and localization," in *Proceedings of the IEEE ICRA*, May 21–25, Brisbane, Australia, pp. 74–79.
He, W., Motlicek, P., and Odobez, J.-M. (**2018b**). "Joint localization and classification of multiple sound sources using a multi-task neural network," in *Proceedings of Interspeech*, September 2–6, Hyderabad, India, pp. 312–316.
He, W., Motlicek, P., and Odobez, J.-M. (**2019a**). "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *Proceedings of the ICASSP*, May 12–17, Brighton, UK, pp. 770–774.
He, W., Motlicek, P., and Odobez, J.-M. (**2021a**). "Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation," IEEE/ACM Trans. Audio. Speech. Lang. Process. **29**, 1303–1317.
He, Y., Trigoni, N., and Markham, A. (**2021b**). "SoundDet: Polyphonic moving sound event detection and localization from raw waveform," in *Proceedings of the ICML*, July 18–24, Virtual Conference.
He, Z., Li, J., Liu, D., He, H., and Barber, D. (**2019b**). "Tracking by animation: Unsupervised learning of multi-object attentive trackers," in *Proceedings of the IEEE CVPR*, June 16–20, Long Beach, CA, pp. 1318–1327.
He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition," in *Proceedings of CVPR*, June 26–July 1, Las Vegas, NV, pp. 770–778.
Heymann, J., Drude, L., Boeddeker, C., Hanebrink, P., and Haeb-Umbach, R. (**2017**). "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *Proceedings of the ICASSP*, March 5–9, New Orleans, LA.
Heymann, J., Drude, L., and Haeb-Umbach, R. (**2016**). "Neural network based spectral mask estimation for acoustic beamforming," in *Proceedings of the ICASSP*, March 20–25, Shanghai, China.
Hickling, R., Wei, W., and Raspet, R. (**1993**). "Finding the direction of a sound source using a vector sound-intensity probe," J. Acoust. Soc. Am. **94**(4), 2408–2412.
Higuchi, T., Kinoshita, K., Delcroix, M., Zmolkova, K., and Nakatani, T. (**2017**). "Deep clustering-based beamforming for separation with unknown number of sources," in *Proceedings of Interspeech*, August 20–24, Stockholm, Sweden.
Hirvonen, T. (**2015**). "Classification of spatial audio location and content using convolutional neural networks," in *Proceedings of the AES Convention*, May 7–10, Warsaw, Poland.
Hochreiter, S., and Schmidhuber, J. (**1997**). "Long short-term memory," Neural Comput. **9**(8), 1735–1780.
Hogg, A. O., Neo, V. W., Weiss, S., Evers, C., and Naylor, P. A. (**2021**). "A polynomial eigenvalue decomposition MUSIC approach for broadband sound source localization," in *Proceedings of WASPAA*, October 21–24, New Paltz, NY.
Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (**2020**). "Squeeze-and-excitation networks," IEEE Trans. Pattern Anal. Mach. Intell. **42**(8), 2011–2023.

Huang, D., and Perez, R. (**2021**). "SSELDNET: A fully end-to-end sample-level framework for sound event localization and detection," Technical Report, DCASE 2021 Challenge.
Huang, Y., Wu, X., and Qu, T. (**2018**). "DNN-based sound source localization method with microphone array," in *Proceedings of the IECE*, October 28–29, Beijing, China.
Huang, Y., Wu, X., and Qu, T. (**2019**). "A time-domain end-to-end method for sound source localization using multi-task learning," in *Proceedings of the ICSP*, September 28–30, Weihai, China, pp. 52–56.
Huang, Y., Wu, X., and Qu, T. (**2020**). "A time-domain unsupervised learning based sound source localization method," in *Proceedings of the ICSP*, September 12–15, Shanghai, China, pp. 26–32.
Jacobsen, F., and Juhl, P. M. (**2013**). *Fundamentals of General Linear Acoustics* (John Wiley & Sons, New York).
Jarrett, D. P., Habets, E. A., and Naylor, P. A. (**2010**). "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proceedings of EUSIPCO*, August 23–27, Aalborg, Denmark, pp. 442–446.
Jarrett, D. P., Habets, E. A., and Naylor, P. A. (**2017**). *Theory and Applications of Spherical Microphone Array Processing* (Springer, New York).
Jarrett, D., Habets, E., Thomas, M., and Naylor, P. (**2012**). "Rigid sphere room impulse response simulation: Algorithm and applications," J. Acoust. Soc. Am. **132**(3), 1462–1472.
Jenrungrot, T., Jayaram, V., Seitz, S., and Kemelmacher-Shlizerman, I. (**2020**). "The cone of silence: Speech separation by localization," arXiv:2010.06007.
Jha, S., Chapman, R., and Durrani, T. (**1988**). "Bearing estimation using neural networks," in *Proceedings of ICASSP*, April 11–14, New York, NY, pp. 2156–2159.
Jha, S., and Durrani, T. (**1989**). "Bearing estimation using neural optimisation methods," in *Proceedings of IEE International Conference on Artificial Neural Networks*, October 16–18, London, UK, pp. 129–133.
Jha, S., and Durrani, T. (**1991**). "Direction of arrival estimation using artificial neural networks," IEEE Trans. Syst, Man, Cybernet. **21**(5), 1192–1201.
Kapka, S., and Lewandowski, M. (**2019**). "Sound source detection, localization and classification using consecutive ensemble of CRNN models," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, October 25–26, New York, NY.
Karthik, S., Prabhu, A., and Gandhi, V. (**2020**). "Simple unsupervised multi-object tracking," arXiv:2006.02609.
Kim, J., and Hahn, M. (**2018**). "Voice activity detection using an adaptive context attention model," IEEE Signal Process. Lett. **25**(8), 1181–1185.
Kim, Y. (**2014**). "Convolutional neural networks for sentence classification," arXiv:1408.5882.
Kim, Y., and Ling, H. (**2011**). "Direction of arrival estimation of humans with a small sensor array using an artificial neural network," PIER. B **27**, 127–149.
Kingma, D. P., and Welling, M. (**2014**). "Auto-encoding variational Bayes," in *Proceedings of the ICLR*, April 14–16, Banff, Canada.
Kitić, S., Bertin, N., and Gribonval, R. (**2014**). "Hearing behind walls: Localizing sources in the room next door with cosparsity," in *Proceedings of the ICASSP*, May 4–9, Florence, Italy, pp. 3087–3091.
Kitić, S., and Guérin, A. (**2018**). "TRAMP: Tracking by a Real-time AMbisonic-based Particle filter," in *IEEE-AASP Challenge on Acoustic Source Localization and Tracking (LOCATA)*, July 8–11, Sheffield, UK.
Knapp, C., and Carter, G. (**1976**). "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust, Speech, Signal Process. **24**(4), 320–327.
Komatsu, T., Togami, M., and Takahashi, T. (**2020**). "Sound event localization and detection using convolutional recurrent neural networks and gated linear units," in *Proceedings of the EUSIPCO*, August 24–28, Amsterdam, The Netherlands (virtual conference), pp. 41–45.
Kong, Q., Cao, Y., Iqbal, T., Wang, W., and Plumbley, M. D. (**2019**). "Cross-task learning for audio tagging, sound event detection and spatial localization," Technical Report, DCASE 2019 Challenge.
Kounades-Bastian, D., Girin, L., Alameda-Pineda, X., Gannot, S., and Horaud, R. (**2017**). "An EM algorithm for joint source separation and diarisation of multichannel convolutive speech mixtures," in *Proceedings of the ICASSP*, March 5–9, New Orleans, LA, pp. 16–20.
Kouw, W. M., and Loog, M. (**2019**). "A review of domain adaptation without target labels," IEEE Trans. Pattern Anal. Mach. Intell. **43**(3), 766–785.
Koyama, S., Nishida, T., Kimura, K., Abe, T., Ueno, N., and Brunnström, J. (**2021**). "MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods," in *Proceedings of WASPAA*, October 21–24, New Paltz, NY (virtual conference), pp. 1–5.
Krause, D., and Kowalczyk, K. (**2019**). "Arborescent neural network architectures for sound event detection and localization," Technical Report, DCASE 2019 Challenge.
Krause, D., Politis, A., and Kowalczyk, K. (**2020a**). "Comparison of convolution types in CNN-based feature extraction for sound source localization," in *Proceedings of the EUSIPCO*, August 24–28, Amsterdam, The Netherlands (virtual conference), pp. 820–824.
Krause, D., Politis, A., and Kowalczyk, K. (**2020b**). "Feature overview for joint modeling of sound event detection and localization using a microphone array," in *Proceedings of the EUSIPCO*, August 24–28, Amsterdam, The Netherlands (virtual conference), pp. 31–35.
Krause, D., Politis, A., and Kowalczyk, K. (**2021**). "Data diversity for improving DNN-based localization of concurrent sound events," in *Proceedings of the EUSIPCO*, August 23–27, Dublin, Ireland (virtual conference), pp. 236–240.
Kristoffersen, M. S., Møller, M. B., Martínez-Nuevo, P., and Østergaard, J. (**2021**). "Deep sound field reconstruction in real rooms: Introducing the ISOBEL sound field dataset," arXiv:2102.06455.
Krizhevsky, A., Sutskever, I., and Hinton, G. E. (**2017**). "ImageNet classification with deep convolutional neural networks," Commun. ACM **60**(6), 84–90.
Küçük, A., Ganguly, A., Hao, Y., and Panahi, I. M. S. (**2019**). "Real-time convolutional neural network-based speech source localization on smartphone," IEEE Access **7**, 169969–169978.
Kujawski, A., Herold, G., and Sarradj, E. (**2019**). "A deep learning method for grid-free localization and quantification of sound sources," J. Acoust. Soc. Am. **146**(3), EL225–EL231.
Kuttruff, H. (**2016**). *Room Acoustics* (CRC Press, Boca Raton, FL).
Lamel, L., Gauvain, J.-L., and Eskenazi, M. (**1991**). "BREF, a large vocabulary spoken corpus for French," in *Proceedings of Eurospeech*, September 24–26, Genova, Italy, pp. 4–7.
Landschoot, C. R., and Xiang, N. (**2019**). "Model-based Bayesian direction of arrival analysis for sound sources using a spherical microphone array," J. Acoust. Soc. Am **146**(6), 4936–4946.
Lathoud, G., Odobez, J.-M., and Gatica-Perez, D. (**2004**). "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proceedings of the International MLMI*, June 21–23, Martigny, Switzerland, pp. 182–195.
Laufer-Goldshtein, B., Talmon, R., and Gannot, S. (**2020**). "Data-driven multi-microphone speaker localization on manifolds," FNT Signal Process. **14**(1–2), 1–161.
Lea, C., Flynn, M. D., Vidal, R., Reiter, A., and Hager, G. D. (**2017**). "Temporal convolutional networks for action segmentation and detection," in *Proceedings of the CVPR*, July 21–26, Honolulu, HI, pp. 1003–1012.
LeCun, Y., Bengio, Y., and Hinton, G. (**2015**). "Deep learning," Nature **521**(7553), 436–444.
Lee, H., Cho, J., Kim, M., and Park, H. (**2016**). "DNN-based feature enhancement using DoA-constrained ICA for robust speech recognition," IEEE Signal Process. Lett. **23**(8), 1091–1095.
Lee, J., Park, J., Kim, K. L., and Nam, J. (**2017**). "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," arXiv:1703.01789.
Lee, S.-H., Hwang, J.-W., Seo, S.-B., and Park, H.-M. (**2021b**). "Sound event localization and detection using cross-modal attention and parameter sharing for DCASE2021 challenge," Technical Report, DCASE 2021 Challenge.
Lee, S. Y., Chang, J., and Lee, S. (**2021a**). "Deep learning-based method for multiple sound source localization with high resolution and accuracy," Mech. Syst. Signal Process. **161**, 107959.
Leglaive, S., Girin, L., and Horaud, R. (**2019**). "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proceedings of the ICASSP*, May 12–17, Brighton, UK, pp. 101–105.

Lehmann, E. A., and Johansson, A. M. (**2010**). "Diffuse reverberation model for efficient image-source simulation of room impulse responses," IEEE Trans. Audio. Speech. Lang. Process. **18**(6), 1429–1439.
Le Moing, G., Vinayavekhin, P., Agravante, D. J., Inoue, T., Vongkulbhisal, J., Munawar, A., and Tachibana, R. (**2021**). "Data-efficient framework for real-world multiple sound source 2D localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toronto, Canada (virtual conference).
Le Moing, G., Vinayavekhin, P., Inoue, T., Vongkulbhisal, J., Munawar, A., Tachibana, R., and Agravante, D. J. (**2020**). "Learning multiple sound source 2D localization," in *Proceedings of the IEEE MMSP*, September 21–24, Tampere, Finland (virtual conference).
Leung, S., and Ren, Y. (**2019**). "Spectrum combination and convolutional recurrent neural networks for joint localization and detection of sound events," Technical Report, DCASE 2019 Challenge.
Li, B., Sainath, T. N., Weiss, R. J., Wilson, K. W., and Bacchiani, M. (**2016d**). "Neural network adaptive beamforming for robust multichannel speech recognition," in *Proceedings of Interspeech*, September 8–12, San Francisco, CA.
Li, Q., Zhang, X., and Li, H. (**2018**). "Online direction of arrival estimation based on deep learning," in *Proceedings of the ICASSP*, April 15–20, Calgary, Canada, pp. 2616–2620.
Li, X., Girin, L., Badeig, F., and Horaud, R. (**2016a**). "Reverberant sound localization with a robot head based on direct-path relative transfer function," in *Proceedings of the IROS*, October 9–14, Daejeon, Korea, pp. 2819–2826.
Li, X., Girin, L., Horaud, R., and Gannot, S. (**2015**). "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *Proceedings of the ICASSP*, April 19–24, Brisbane, Australia, pp. 320–324.
Li, X., Girin, L., Horaud, R., and Gannot, S. (**2016b**). "Estimation of the direct-path relative transfer function for supervised sound source localization," IEEE/ACM Trans. Audio. Speech. Lang. Process. **24**(11), 2171–2186.
Li, X., Girin, L., Horaud, R., and Gannot, S. (**2017**). "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," IEEE/ACM Trans. Audio. Speech. Lang. Process. **25**(10), 1197–2012.
Li, X., Horaud, R., Girin, L., and Gannot, S. (**2016c**). "Voice activity detection based on statistical likelihood ratio with adaptive thresholding," in *Proceedings of the IEEE IWAENC*, September 13–16, Xi'an, China, pp. 1–5.
Liang, Y., and Zhou, Y. (**2018**). "LSTM multiple object tracker combining multiple cues," in *Proceedings of the IEEE ICIP*, February 7–14, Athens, Greece, pp. 2351–2355.
Lin, X., Girin, L., and Alameda-Pineda, X. (**2022**). "Unsupervised multiple-object tracking with a dynamical variational autoencoder," arXiv:2202.09315.
Lin, Y., and Wang, Z. (**2019**). "A report on sound event localization and detection," Technical Report, DCASE 2019 Challenge.
Liu, N., Chen, H., Songgong, K., and Li, Y. (**2021**). "Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays," J. Acoust. Soc. Am. **149**(2), 1069–1084.
Liu, Z.-M., Huang, Z.-T., and Zhou, Y.-Y. (**2012**). "An efficient maximum likelihood method for direction-of-arrival estimation via sparse Bayesian learning," IEEE Trans. Wireless Commun. **11**(10), 1–11.
Liu, Z.-M., Zhang, C., and Yu, P. S. (**2018**). "Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections," IEEE Trans. Antennas Propagat. **66**(12), 7315–7327.
Lu, Z. (**2019**). "Sound event detection and localization based on CNN and LSTM," Technical Report, DCASE 2019 Challenge.
Luiten, J., Zulfikar, I. E., and Leibe, B. (**2020**). "UnOVOST: Unsupervised offline video object segmentation and tracking," in *Proceedings of the IEEE WACV*, March 1–5, Snowmass Village, CO, pp. 1989–1998.
Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., and Kim, T.-K. (**2021**). "Multiple object tracking: A literature review," Artif. Intell. **293**, 103448.
Luo, Y., Chen, Z., Mesgarani, N., and Yoshioka, T. (**2020**). "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proceedings of the ICASSP 2020*, May 4–8, Barcelona, Spain, pp. 6394–6398.
Luo, Y., and Mesgarani, N. (**2019**). "Conv-TASnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM Trans. Audio. Speech. Lang. Process. **27**(8), 1256–1266.
Ma, N., Brown, G., and May, T. (**2015**). "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proceedings of Interspeech*, September 6–10, Dresden, Germany, pp. 160–164.
Ma, W., and Liu, X. (**2018**). "Phased microphone array for sound source localization with deep learning," Aerosp. Syst. **2**(2), 71–81.
Mabande, E., Sun, H., Kowalczyk, K., and Kellermann, W. (**2011**). "Comparison of subspace-based and steered beamformer-based reflection localization methods," in *Proceedings of the EUSIPCO*, August 29–September 2, Barcelona, Spain, pp. 146–150.
Mack, W., Bharadwaj, U., Chakrabarty, S., and Habets, E. A. P. (**2020**). "Signal-aware broadband DoA estimation using attention mechanisms," in *Proceedings of the ICASSP 2020*, May 4–8, Barcelona, Spain, pp. 4930–4934.
Mandel, M. I., Weiss, R. J., and Ellis, D. P. (**2009**). "Model-based expectation-maximization source separation and localization," IEEE Trans. Audio. Speech. Lang. Process. **18**(2), 382–394.
Markovich-Golan, S., and Gannot, S. (**2015**). "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proceedings of the ICASSP*, April 19–24, Brisbane, Australia.
Maruri, H. A. C., Meyer, P. L., Huang, J., Ontiveros, JAdH., and Lu, H. (**2019**). "GCC-PHAT cross-correlation audio features for simultaneous sound event localization and detection (SELD) in multiple rooms," Technical Report, DCASE 2019 Challenge.
Masuyama, Y., Bando, Y., Yatabe, K., Sasaki, Y., Onishi, M., and Oikawa, Y. (**2020**). "Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling," in *Proceedings of the IEEE/RSJ IROS*, October 25–29, Las Vegas, NV, pp. 4848–4854.
May, T., Van De Par, S., and Kohlrausch, A. (**2011**). "A probabilistic model for robust localization based on a binaural auditory front-end," IEEE Trans. Audio. Speech. Lang. Process. **19**(1), 1–13.
Maysenhölder, W. (**1993**). "The reactive intensity of general time-harmonic structure-borne sound fields," in *Proceedings of the International Congress on Intensity Techniques*, August 31–September 2, Leuven, Belgium, pp. 63–70.
Mazzon, L., Koizumi, Y., Yasuda, M., and Harada, N. (**2019**). "First order Ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, October 25–26, New York.
Meinhardt, T., Kirillov, A., Leal-Taixe, L., and Feichtenhofer, C. (**2021**). "Trackformer: Multi-object tracking with transformers," arXiv:2101.02702.
Meng, Z., Watanabe, S., Hershey, J. R., and Erdogan, H. (**2017**). "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proceedings of the ICASSP*, March 5–9, New Orleans, LA.
Merimaa, J. (**2006**). "Analysis, synthesis, and perception of spatial sound: Binaural localization modeling and multichannel loudspeaker reproduction," Ph.D. thesis, Helsinki University of Technology, Helsinki, Finland.
Nam, S., Davies, M. E., Elad, M., and Gribonval, R. (**2013**). "The cosparse analysis model and algorithms," Appl. Comput. Harmonic Anal. **34**(1), 30–56.
Nannuru, S., Koochakzadeh, A., Gemba, K. L., Pal, P., and Gerstoft, P. (**2018**). "Sparse Bayesian learning for beamforming using sparse linear arrays," J. Acoust. Soc. Am. **144**(5), 2719–2729.
Naranjo-Alcazar, J., Perez-Castanos, S., Cobos, M., Ferri, F. J., and Zuccarello, P. (**2021**). "Sound event localisation and detection using squeeze-excitation residual CNNs," Technical Report, DCASE 2021 Challenge.
Naranjo-Alcazar, J., Perez-Castanos, S., Ferrandis, J., Zuccarello, P., and Cobos, M. (**2020**). "Sound event localization and detection using squeeze-excitation residual CNNs," Technical Report, DCASE 2020 Challenge.
Nehorai, A., and Paldi, E. (**1994**). "Acoustic vector-sensor array processing," IEEE Trans. Signal Process. **42**(9), 2481–2491.
Nguyen, Q., Girin, L., Bailly, G., Elisei, F., and Nguyen, D.-C. (**2018**). "Autonomous sensorimotor learning for sound source localization by a humanoid robot," in *IEEE/RSJ IROS*, October 1–5, Madrid, Spain.
Nguyen, T. N. T., Gan, W.-S., Ranjan, R., and Jones, D. L. (**2020a**). "Robust source counting and DoA estimation using spatial pseudo-

spectrum and convolutional neural network," IEEE/ACM Trans. Audio. Speech. Lang. Process. **28**, 2626–2637.
Nguyen, T. N. T., Jones, D. L., and Gan, W. S. (**2020b**). "Ensemble of sequence matching networks for dynamic sound event localization, detection, and tracking," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, November 2–4, Tokyo, Japan.
Nguyen, T. N. T., Jones, D. L., and Gan, W.-S. (**2020c**). "A sequence matching network for polyphonic sound event localization and detection," in *Proceedings of the ICASSP 2020*, May 4–8, Barcelona, Spain (virtual conference), pp. 71–75.
Nguyen, T. N. T., Nguyen, N. K., Phan, H., Pham, L., Ooi, K., Jones, D. L., and Gan, W.-S. (**2021a**). "A general network architecture for sound event localization and detection using transfer learning and recurrent neural network," in *Proceedings of the ICASSP*, June 6–11, Toronto, Canada (virtual conference), pp. 935–939.
Nguyen, T. N. T., Watcharasupat, K., Nguyen, N. K., Jones, D. L., and Gan, W. S. (**2021b**). "DCASE 2021 Task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection," Technical Report, DCASE 2021 Challenge.
Noh, K., Choi, J.-H., Jeon, D., and Chang, J.-H. (**2019**). "Three-stage approach for sound event localization and detection," Technical Report, DCASE 2019 Challenge.
Nolan, M., Verburg, S. A., Brunskog, J., and Fernandez-Grande, E. (**2019**). "Experimental characterization of the sound field in a reverberation room," J. Acoust. Soc. Am. **145**(4), 2237–2246.
Noohi, T., Epain, N., and Jin, C. T. (**2013**). "Direction of arrival estimation for spherical microphone arrays by combination of independent component analysis and sparse recovery," in *Proceedings of the ICASSP*, May 26–31, Vancouver, Canada, pp. 346–349.
Nugraha, A. A., Liutkus, A., and Vincent, E. (**2016**). "Multichannel audio source separation with deep neural networks," IEEE/ACM Trans. Audio. Speech. Lang. Process. **24**(9), 1652–1664.
Nustede, E. J., and Anemüller, J. (**2019**). "Group delay features for sound event detection and localization," Technical Report, DCASE 2019 Challenge.
Opochinsky, R., Chechik, G., and Gannot, S. (**2021**). "Deep ranking-based DoA tracking algorithm," in *Proceedings of the EUSIPCO*, August 23–27, Dublin, Ireland (virtual conference), pp. 1020–1024.
Opochinsky, R., Laufer-Goldshtein, B., Gannot, S., and Chechik, G. (**2019**). "Deep ranking-based sound source localization," in *Proceedings of WASPAA*, October 21–24, New Paltz, NY, pp. 283–287.
Pak, J., and Shin, J. W. (**2019**). "Sound localization based on phase difference enhancement using deep neural networks," IEEE/ACM Trans. Audio. Speech. Lang. Process. **27**(8), 1335–1345.
Pal, P., and Vaidyanathan, P. P. (**2010**). "Nested arrays: A novel approach to array processing with enhanced degrees of freedom," IEEE Trans. Signal Process. **58**(8), 4167–4181.
Pang, C., Liu, H., and Li, X. (**2019**). "Multitask learning of time-frequency CNN for sound source localization," IEEE Access **7**, 40725–40737.
Parcollet, T., Zhang, Y., Morchid, M., Trabelsi, C., Linarès, G., De Mori, R., and Bengio, Y. (**2018**). "Quaternion Convolutional Neural Networks for End-to-End Automatic Speech Recognition," arXiv:1806.07789.
Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (**2019a**). "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of Interspeech*, September 15–19, Graz, Austria, pp. 2613–2617.
Park, S., Jeong, Y., and Lee, T. (**2021a**). "Many-to-many audio spectrogram transformer: Transformer for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, November 15–19, Barcelona, Spain, pp. 105–109.
Park, S., Lim, W., Suh, S., and Jeong, Y. (**2019b**). "TrellisNet-based architecture for sound event localization and detection with reassembly learning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, October 25–26, New York, NY.
Park, S., Suh, S., and Jeong, Y. (**2020**). "Sound event localization and detection with various loss functions," Technical Report, DCASE 2020 Challenge.
Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (**2021b**). "A review of speaker diarization: Recent advances with deep learning," arXiv:2101.09624.
Patel, S. J., Zawodniok, M., and Benesty, J. (**2020**). "A single stage fully convolutional neural network for sound source localization and detection," Technical Report, DCASE 2020 Challenge.
Pavlidi, D., Delikaris-Manias, S., Pulkki, V., and Mouchtaris, A. (**2015**). "3D localization of multiple sound sources with intensity vector estimates in single source zones," in *Proceedings of the EUSIPCO*, August 31–September 4, Nice, France, pp. 1556–1560.
Peeters, G. (**2004**). "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," CUIDADO Project Report 54.0.
Perotin, L., Défossez, A., Vincent, E., Serizel, R., and Guérin, A. (**2019a**). "Regression versus classification for neural network based audio source localization," in *Proceedings of WASPAA*, October 21–24, New Paltz, NY.
Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (**2018a**). "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *Proceedings of the ICASSP*, April 15–20, Calgary, Canada.
Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (**2018b**). "CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector," in *Proceedings of the IWAENC*, September 17–20, Tokyo, Japan, pp. 241–245.
Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (**2019b**). "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," IEEE J. Sel. Top. Signal Process. **13**(1), 22–33.
Pertilä, P., and Cakir, E. (**2017**). "Robust direction estimation with convolutional neural networks based steered response power," in *Proceedings of the ICASSP*, March 5–9, New Orleans, LA, pp. 6125–6129.
Phan, H., Pham, L., Koch, P., Duong, N. Q. K., McLoughlin, I., and Mertins, A. (**2020a**). "Audio event detection and localization with multitask regression network," Technical Report, DCASE 2020 Challenge.
Phan, H., Pham, L., Koch, P., Duong, N. Q. K., McLoughlin, I., and Mertins, A. (**2020b**). "On multitask loss function for audio event detection and localization," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, November 2–4, Tokyo, Japan, pp. 160–164.
Ping, G., Fernandez-Grande, E., Gerstoft, P., and Chu, Z. (**2020**). "Three-dimensional source localization using sparse Bayesian learning on a spherical microphone array," J. Acoust. Soc. Am. **147**(6), 3895–3904.
Politis, A., Adavanne, S., Krause, D., Deleforge, A., Srivastava, P., and Virtanen, T. (**2021**). "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, November 15–19, Barcelona, Spain.
Politis, A., Adavanne, S., and Virtanen, T. (**2020a**). "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, November 2–4, Tokyo, pp. 165–169.
Politis, A., Mesaros, A., Adavanne, S., Heittola, T., and Virtanen, T. (**2020b**). "Overview and evaluation of sound event localization and detection in DCASE 2019," IEEE/ACM Trans. Audio. Speech. Lang. Process. **29**, 684–698.
Poschadel, N., Hupke, R., Preihs, S., and Peissig, J. (**2021a**). "Direction of arrival estimation of noisy speech using convolutional recurrent neural networks with higher-order Ambisonics signals," in *Proceedings of the EUSIPCO*, August 23–27, Dublin, Ireland (virtual conference).
Poschadel, N., Preihs, S., and Peissig, J. (**2021b**). "Multi-source direction of arrival estimation of noisy speech using convolutional recurrent neural networks with higher-order ambisonics signals," in *Proceedings of the EUSIPCO*, August 23–27, Dublin, Ireland (virtual conference), pp. 1015–1019.
Pratik, P., Jee, W. J., Nagisetty, S., Mars, R., and Lim, C. (**2019**). "Sound event localization and detection using CRNN architecture with Mixup for model generalization," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, October 25–26, New York.
Pujol, H., Bavu, E., and Garcia, A. (**2019**). "Source localization in reverberant rooms using deep learning and microphone arrays," in *Proceedings of the ICA Conference*, September 9–13, Aachen, Germany.

Pujol, H., Bavu, E., and Garcia, A. (**2021**). "BeamLearning: An end-to-end deep learning approach for the angular localization of sound sources using raw multichannel acoustic pressure data," J. Acoust. Soc. Am. **149**(6), 4248–4263.
Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (**2019**). "Deep learning for audio signal processing," IEEE J. Sel. Top. Signal Process. **13**(2), 206–219.
Raangs, R., and Druyvesteyn, E. (**2002**). "Sound source localization using sound intensity measured by a three dimensional PU-probe," in *Proceedings of AES Convention*, May 10–13, Munich, Germany.
Rafaely, B. (**2019**). *Fundamentals of Spherical Array Processing* (Springer, New York).
Ranjan, R., Jayabalan, S., Nguyen, T. N. T., and Lim, W.-S. (**2019**). "Sound events detection and direction of arrival estimation using residual net and recurrent neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, October 25–26, New York.
Rastogi, R., Gupta, P., and Kumaresan, R. (**1987**). "Array signal processing with interconnected neuron-like elements," in *Proceedings of the ICASSP*, April 6–9, Dallas, TX, pp. 2328–2331.
Rezende, D. J., Mohamed, S., and Wierstra, D. (**2014**). "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the ICML*, June 21–26, Beijing, China.
Rho, D., Lee, S., Park, J., Kim, T., Chang, J., and Ko, J. (**2021**). "A combination of various neural networks for sound event localization and detection," Technical Report, DCASE 2021 Challenge.
Rickard, S. (**2002**). "On the approximate W-disjoint orthogonality of speech," in *Proceedings of the ICASSP*, May 13–17, Orlando, FL, pp. 529–532.
Riezu, S. A. V., and Grande, E. F. (**2021**). "Room impulse response dataset—ACT, DTU Elektro (011, IEC; plane, sphere)," https://data.dtu.dk/articles/dataset/Room_Impulse_Response_Dataset_-_ACT_DTU_Elektro_011_IEC_plane_sphere_/14320166 (Last viewed June 27, 2022).
Rindel, J. H. (**2000**). "The use of computer modeling in room acoustics," J. Vibroeng. **3**(4), 219–224.
Roden, R., Moritz, N., Gerlach, S., Weinzierl, S., and Goetze, S. (**2015**). "On sound source localization of speech signals using deep neural networks," in *Proceedings of Deutsche Jahrestagung Akustik (DAGA)*, March 16–19, Nuremberg, Germany.
Roman, N., and Wang, D. (**2008**). "Binaural tracking of multiple moving sources," IEEE Trans. Audio, Speech, Lang. Process. **16**(4), 728–739.
Ronchini, F., Arteaga, D., and Pérez-López, A. (**2020**). "Sound event localization and detection based on CRNN using rectangular filters and channel rotation data augmentation," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, November 2–4, Tokyo, Japan.
Ronneberger, O., Fischer, P., and Brox, T. (**2015**). "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the MICCAI*, October 5–9, Munich, Germany, pp. 234–241.
Rossing, T. D. (**2007**). *Springer Handbook of Acoustics* (Springer, New York).
Roy, R., and Kailath, T. (**1989**). "ESPRIT: Estimation of signal parameters via rotational invariance techniques," IEEE Trans. Acoust, Speech, Signal Process. **37**(7), 984–995.
Ruder, S. (**2017**). "An overview of multi-task learning in deep neural networks," arXiv preprint arXiv:1706.05098.
Sadeghian, A., Alahi, A., and Savarese, S. (**2017**). "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *Proceedings of the ICCV*, October 22–29, Venice, Italy, pp. 300–311.
Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., and Séguier, R. (**2022**). "Learning and controlling the source-filter representation of speech with a variational autoencoder," arXiv:2204.07075.
Sainath, T. N., Weiss, R. J., Wilson, K. W., Li, B., Narayanan, A., Variani, E., Bacchiani, M., Shafran, I., Senior, A., Chin, K., Misra, A., and Kim, C., (**2017**). "Multichannel signal processing with deep neural networks for automatic speech recognition," IEEE/ACM Trans. Audio. Speech. Lang. Process. **25**(5), 965–979.
Salamon, J., and Bello, J. P. (**2017**). "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Process. Lett. **24**(3), 279–283.
Saleh, F., Aliakbarian, S., Rezatofighi, H., Salzmann, M., and Gould, S. (**2021**). "Probabilistic tracklet scoring and inpainting for multiple object tracking," in *Proceedings of the CVPR*, June 19–25, Nashville, TN (virtual conference), pp. 14329–14339.
Salvati, D., Drioli, C., and Foresti, G. L. (**2018**). "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions," IEEE Trans. Emerg. Top. Comput. Intell. **2**(2), 103–116.
Sampathkumar, A., and Kowerko, D. (**2020**). "Sound event detection and localization using CRNN models," Technical Report, DCASE 2020 Challenge.
Sato, I., Liu, G., Ishikawa, K., Suzuki, T., and Tanaka, M. (**2021**). "Does end-to-end trained deep model always perform better than non-end-to-end counterpart?," Electron. Imag. **2021**(10), 240–241.
Sawada, H., Mukai, R., and Makino, S. (**2003**). "Direction of arrival estimation for multiple source signals using independent component analysis," in *IEEE International Symposium on Signal Processing in Applications*, July 1–4, Paris, France, pp. 411–414.
Scheibler, R., Bezzam, E., and Dokmani, I. (**2018**). "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proceedings of the ICASSP*, April 15–20, Calgary, Canada, pp. 351–355.
Schmidt, R. (**1986**). "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propagat. **34**(3), 276–280.
Schwartz, O., and Gannot, S. (**2013**). "Speaker tracking using recursive EM algorithms," IEEE/ACM Trans. Audio. Speech. Lang. Process. **22**(2), 392–402.
Schymura, C., Bönninghoff, B., Ochiai, T., Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., and Kolossa, D. (**2021**). "PILOT: Introducing Transformers for probabilistic sound event localization," in *Proceedings of Interspeech*, August 30–September 3, Brno, Czech Republic.
Schymura, C., Ochiai, T., Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., and Kolossa, D. (**2020**). "Exploiting attention-based sequence-to-sequence architectures for sound event localization," in *Proceedings of the EUSIPCO*, August 24–28, Amsterdam, The Netherlands.
Sehgal, A., and Kehtarnavaz, N. (**2018**). "A convolutional neural network smartphone app for real-time voice activity detection," IEEE Access **6**, 9017–9026.
Shimada, K., Koyama, Y., Takahashi, N., Takahashi, S., and Mitsufuji, Y. (**2020a**). "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *Proceedings of the ICASSP*, May 4–8, Barcelona, Spain (virtual conference).
Shimada, K., Takahashi, N., Koyama, Y., Takahashi, S., Tsunoo, E., Takahashi, M., and Mitsufuji, Y. (**2021**). "Ensemble of ACCDOA- and EINV2-based systems with d3nets and impulse response simulation for sound event localization and detection," Technical Report, DCASE 2021 Challenge.
Shimada, K., Takahashi, N., Takahashi, S., and Mitsufuji, Y. (**2020b**). "Sound event localization and detection using activity-coupled cartesian DoA vector and RD3net," Technical Report, DCASE 2020 Challenge.
Shlezinger, N., Whang, J., Eldar, Y. C., and Dimakis, A. G. (**2020**). "Model-based deep learning," arXiv:2012.08405.
Siltanen, S., Lokki, T., and Savioja, L. (**2010**). "Rays or waves? understanding the strengths and weaknesses of computational room acoustics modeling techniques," in *Proceedings of the ISRA*, August 29–31, Melbourne, Australia.
Singla, R., Tiwari, S., and Sharma, R. (**2020**). "A sequential system for sound event detection and localization using CRNN," Technical Report, DCASE 2020 Challenge.
Sivasankaran, S., Vincent, E., and Fohr, D. (**2018**). "Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment," in *Proceedings of Interspeech*, September 2–6, Hyderabad, India.
Song, J-m. (**2020**). "Localization and detection for moving sound sources using consecutive ensembles of 2D-CRNN," Technical Report, DCASE 2020 Challenge.
Southall, H., Simmers, J., and O'Donnell, T. (**1995**). "Direction finding in phased arrays with a neural network beamformer," IEEE Trans. Antennas Propagat. **43**(12), 1369–1374.
Stiefelhagen, R., Bernardin, K., Bowers, R., Rose, R. T., Michel, M., and Garofolo, J. (**2007**). "The CLEAR 2007 evaluation," in *Proceedings of the Multimodal Technologies for Perception of Humans*, May 8–11, Baltimore, MD, pp. 3–34.
Subramani, K., and Smaragdis, P. (**2021**). "Point cloud audio processing," in *Proceedings of WASPAA*, October 21–24, New Paltz, NY (virtual conference), pp. 31–35.

Subramanian, A. S., Weng, C., Watanabe, S., Yu, M., Xu, Y., Zhang, S.-X., and Yu, D. (**2021a**). "Directional ASR: A new paradigm for E2E multi-speaker speech recognition with source localization," in *Proceedings of the ICASSP*, June 6–11, Toronto, Canada (virtual conference), pp. 8433–8437.

Subramanian, A. S., Weng, C., Watanabe, S., Yu, M., and Yu, D. (**2021b**). "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," arXiv:2102.07955.

Sudarsanam, P. A., Politis, A., and Drossos, K. (**2021**). "Assessment of self-attention on learned features for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, November 15–19, Barcelona, Spain, pp. 100–104.

Sudo, Y., Itoyama, K., Nishida, K., and Nakadai, K. (**2019**). "Improvement of DOA estimation by using quaternion output in sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, October 25–26, New York, NY.

Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., and Luo, P. (**2020**). "Transtrack: Multiple object tracking with transformer," arXiv:2012.15460.

Sundar, H., Wang, W., Sun, M., and Wang, C. (**2020**). "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *Proceedings of the ICASSP*, May 4–8, Barcelona, Spain (virtual conference), pp. 4642–4646.

Suvorov, D., Dong, G., and Zhukov, R. (**2018**). "Deep residual network for sound source localization in the time domain," arXiv:1808.06429.

Svensson, P., and Kristiansen, U. R. (**2002**). "Computational modelling and simulation of acoustic spaces," in *Proceedings of the AES Conference*, June 15–17, Espoo, Finland.

Szöke, I., Skácel, M., Moner, L., Paliesek, J., and Černocký, J. (**2019**). "Building and evaluation of a real room impulse response dataset," IEEE J. Sel. Top. Signal Process. **13**(4), 863–876.

Takahashi, N., Goswami, N., and Mitsufuji, Y. (**2018**). "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proceedings of the IWAENC*, September 17–20, Tokyo, Japan.

Takahashi, N., Gygli, M., Pfister, B., and Gool, L. V. (**2016**). "Deep convolutional neural networks and data augmentation for acoustic event recognition," in *Proceedings of Interspeech*, September 8–12, San Francisco, CA, pp. 2982–2986.

Takeda, R., and Komatani, K. (**2016a**). "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *Proceedings of the IEEE Spoken Language Technology Workshop*, December 13–16, San Juan, Portugal, pp. 603–609.

Takeda, R., and Komatani, K. (**2016b**). "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proceedings of the ICASSP*, March 20–25, Shanghai, China, pp. 405–409.

Takeda, R., and Komatani, K. (**2017**). "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," in *Proceedings of the ICASSP*, March 5–9, New Orleans, LA, pp. 2217–2221.

Takeda, R., Kudo, Y., Takashima, K., Kitamura, Y., and Komatani, K. (**2018**). "Unsupervised adaptation of neural networks for discriminative sound source localization with eliminative constraint," in *Proceedings of the ICASSP*, April 15–20, Calgary, Canada, pp. 3514–3518.

Tang, Z., Kanu, J. D., Hogan, K., and Manocha, D. (**2019**). "Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks," in *Proceedings of Interspeech*, September 15–19, Graz, Austria, pp. 654–658.

Tervo, S. (**2009**). "Direction estimation based on sound intensity vectors," in *Proceedings of the EUSIPCO*, August 24–28, Glasgow, Scotland, pp. 700–704.

Thiemann, J., and Van De Par, S. (**2015**). "Multiple model high-spatial resolution HRTF measurements," in *Proceedings of Deutsche Jahrestagung Akustik (DAGA)*, March 16–19, Nuremberg, Germany.

Thuillier, E., Gamper, H., and Tashev, I. J. (**2018**). "Spatial audio feature discovery with convolutional neural networks," in *Proceedings of the ICASSP*, April 15–20, Calgary, Canada, pp. 6797–6801.

Tian, C. (**2020**). "Multiple CRNN for SELD," Technical Report, DCASE 2020 Challenge.

Tranter, S. E., and Reynolds, D. A. (**2006**). "An overview of automatic speaker diarization systems," IEEE Trans. Audio. Speech Lang. Process. **14**(5), 1557–1565.

Tsuzuki, H., Kugler, M., Kuroyanagi, S., and Iwata, A. (**2013**). "An approach for sound source localization by complex-valued neural network," IEICE Trans. Inform. Syst. **96**(10), 2257–2265.

Ünleren, M. F., and Yaldiz, E. (**2016**). "Direction of arrival estimation by using artificial neural networks," in *Proceedings of the European Modelling Symposium*, November 28–30, Pisa, Italy, pp. 242–245.

Vaidyanathan, P. P., and Pal, P. (**2010**). "Sparse sensing with co-prime samplers and arrays," IEEE Trans. Signal Process. **59**(2), 573–586.

Valero, M. L., and Habets, E. A. (**2017**). "Multi-microphone acoustic echo cancellation using relative echo transfer functions," in *Proceedings of WASPAA*, October 21–24, New Paltz, NY, pp. 229–233.

Van Veen, B. D., and Buckley, K. M. (**1988**). "Beamforming: A versatile approach to spatial filtering," IEEE Acoust. Speech Signal Process. Mag. **5**(2), 4–24.

Varanasi, V., Gupta, H., and Hegde, R. M. (**2020**). "A deep learning framework for robust DoA estimation using spherical harmonic decomposition," IEEE/ACM Trans. Audio. Speech. Lang. Process. **28**, 1248–1259.

Vargas, E., Hopgood, J. R., Brown, K., and Subr, K. (**2021**). "On improved training of CNN for acoustic source localisation," IEEE/ACM Trans. Audio. Speech. Lang. Process. **29**, 720–732.

Varzandeh, R., Adilolu, K., Doclo, S., and Hohmann, V. (**2020**). "Exploiting periodicity features for joint detection and DoA estimation of speech sources using convolutional neural networks," in *Proceedings of the ICASSP*, May 4–8, Barcelona, Spain (virtual conference), pp. 566–570.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (**2017**). "Attention is all you need," arXiv:1706.03762.

Vecchiotti, P., Ma, N., Squartini, S., and Brown, G. J. (**2019a**). "End-to-end binaural sound localisation from the raw waveform," in *Proceedings of the ICASSP*, May 12–17, Brighton, UK, pp. 451–455.

Vecchiotti, P., Pepe, G., Principi, E., and Squartini, S. (**2019b**). "Detection of activity and position of speakers by using deep neural networks and acoustic data augmentation," Expert Syst. with Appl. **134**, 53–65.

Vecchiotti, P., Principi, E., Squartini, S., and Piazza, F. (**2018**). "Deep neural networks for joint voice activity detection and speaker localization," in *Proceedings of the EUSIPCO*, September 3–7, Roma, Italy, pp. 1567–1571.

Vera-Diaz, J. M., Pizarro, D., and Macias-Guarasa, J. (**2018**). "Towards end-to-end acoustic localization using deep learning: From audio signal to source position coordinates," Sensors **18**(10), 3418.

Vera-Diaz, J. M., Pizarro, D., and Macias-Guarasa, J. (**2020**). "Towards domain independence in CNN-based acoustic localization using deep cross correlations," in *Proceedings of the EUSIPCO*, August 24–28, Amsterdam, The Netherlands (virtual conference), pp. 226–230.

Vera-Diaz, J. M., Pizarro, D., and Macias-Guarasa, J. (**2021**). "Acoustic source localization with deep generalized cross correlations," Signal Process. **187**, 108169.

Vesperini, F., Vecchiotti, P., Principi, E., Squartini, S., and Piazza, F. (**2016**). "A neural network based algorithm for speaker localization in a multi-room environment," in *Proceedings of the International Workshop on Machine Learning for Signal Processing*, September 13–16, Salerno, Italy, pp. 1–6.

Vincent, E., and Campbell, D. R. (**2008**). "Roomsimove," GNU Public License, http://homepages.loria.fr/evincent/software/Roomsimove_1.

Vincent, E., Virtanen, T., and Gannot, S. (**2018**). *Audio Source Separation and Speech Enhancement* (John Wiley & Sons, New York).

Vo, B.-N., Mallick, M., Bar-shalom, Y., Coraluppi, S., Osborne, R., Mahler, R., and Vo, B-T. (**2015**). "Multitarget tracking," in *Wiley Encyclopaedia of Electrical and Electronics Engineering* (Wiley, New York).

Wabnitz, A., Epain, N., Jin, C., and Van Schaik, A. (**2010**). "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the ISRA*, August 29–31, Melbourne, Australia, pp. 1–6.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (**1989**). "Phoneme recognition using time-delay neural networks," IEEE Trans. Acoust, Speech, Signal Process. **37**(3), 328–339.

Wang, D., and Chen, J. (**2018**). "Supervised speech separation based on deep learning: An overview," IEEE/ACM Trans. Audio. Speech. Lang. Process. **26**(10), 1702–1726.
Wang, L., Liu, Y., Zhao, L., Wang, Q., Zeng, X., and Chen, K. (**2018**). "Acoustic source localization in strong reverberant environment by parametric Bayesian dictionary learning," Signal Process. **143**, 232–240.
Wang, Q., Du, J., Wu, H.-X., Pan, J., Ma, F., and Lee, C.-H. (**2021**). "A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection," arXiv:2101.02919.
Wang, Q., Wu, H., Jing, Z., Ma, F., Fang, Y., Wang, Y., Chen, T., Pan, J., Du, J., and Lee, C.-H. (**2020**). "The USTC-IFLYTEK system for sound event localization and detection of DCASE 2020 challenge," Technical Report, DCASE 2020 Challenge.
Wang, Z., Zhang, X., and Wang, D. (**2019**). "Robust speaker localization guided by deep learning-based time-frequency masking," IEEE/ACM Trans. Audio. Speech. Lang. Process. **27**(1), 178–188.
Wichern, G., Antognini, J., Flynn, M., Zhu, L. R., McQuinn, E., Crow, D., Manilow, E., and Roux, J. L. (**2019**). "Wham!: Extending speech separation to noisy environments," in *Proceedings of Interspeech*, September 15–19, Graz, Austria.
Woodruff, J., and Wang, D. (**2012**). "Binaural localization of multiple sources in reverberant and noisy environments," IEEE Trans. Audio. Speech. Lang. Process. **20**(5), 1503–1512.
Wu, X., Wu, Z., Ju, L., and Wang, S. (**2021c**). "Binaural audio-visual localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, February 2–9, Vancouver, Canada, pp. 2961–2968.
Wu, Y., Ayyalasomayajula, R., Bianco, M. J., Bharadia, D., and Gerstoft, P. (**2021a**). "Sound source localization based on multi-task learning and image translation network," J. Acoust. Soc. Am. **150**(5), 3374–3386.
Wu, Y., Ayyalasomayajula, R., Bianco, M. J., Bharadia, D., and Gerstoft, P. (**2021b**). "SSLIDE: Sound source localization for indoors based on deep learning," in *Proceedings of the ICASSP*, June 6–11, Toronto, Canada (virtual conference).
Xenaki, A., Bünsow Boldt, J., and Græsbøll Christensen, M. (**2018**). "Sound source localization and speech enhancement with sparse Bayesian learning beamforming," J. Acoust. Soc. Am. **143**(6), 3912–3921.
Xenaki, A., and Gerstoft, P. (**2015**). "Grid-free compressive beamforming," J. Acoust. Soc. Am. **137**(4), 1923–1935.
Xenaki, A., Gerstoft, P., and Mosegaard, K. (**2014**). "Compressive beamforming," J. Acoust. Soc. Am. **136**(1), 260–271.
Xiang, J., Zhang, G., and Hou, J. (**2019**). "Online multi-object tracking based on feature representation and Bayesian filtering within a deep learning architecture," IEEE Access **7**, 27923–27935.
Xiao, X., Zhao, S., Zhong, X., Jones, D. L., Chng, E. S., and Li, H. (**2015**). "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proceedings of the ICASSP*, April 19–24, Brisbane, Australia, pp. 2814–2818.
Xinghao, S., Hu, Y., Zhu, X., and He, L. (**2021**). "Sound event localization and detection based on adaptive hybrid convolution and multi-scale feature extractor," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop)*, November 15–19, Barcelona, Spain, pp. 130–134.
Xu, B., Sun, G., Yu, R., and Yang, Z. (**2013**). "High-accuracy TDOA-based localization without time synchronization," IEEE Trans. Parallel Distrib. Syst. **24**(8), 1567–1576.
Xu, P., Arcondoulis, E. J. G., and Liu, Y. (**2021a**). "Acoustic source imaging using densely connected convolutional networks," Mech. Syst. Signal Process. **151**, 107370.
Xu, Y., Ban, Y., Delorme, G., Gan, C., Rus, D., and Alameda-Pineda, X. (**2021b**). "Transcenter: Transformers with dense queries for multiple-object tracking," arXiv:2103.15145.
Xue, W., Tong, Y., Zhang, C., Ding, G., He, X., and Zhou, B. (**2020**). "Sound event localization and detection based on multiple DoA beamforming and multi-task learning," in *Proceedings of the ICSP*, September 12–15, Shanghai, China (virtual conference).
Xue, W., Ying, T., Chao, Z., and Guohong, D. (**2019**). "Multi-beam and multi-task learning for joint sound event detection and localization," Technical Report, DCASE 2019 Challenge.
Yalta, N., Nakadai, K., and Ogata, T. (**2017**). "Sound source localization using deep learning models," J. Robot. Mechatron. **29**(1), 37–48.
Yalta, N., Sumiyoshi, Y., and Kawaguchi, Y. (**2021**). "The Hitachi DCASE 2021 Task 3 system: Handling directive interference with self attention layers," Technical Report, DCASE 2021 Challenge.
Yang, W.-H., Chan, K.-K., and Chang, P.-R. (**1994**). "Complex-valued neural network for direction of arrival estimation," Electron. Lett. **30**(7), 574–575.
Yang, B., Li, X., and Liu, H. (**2021a**). "Supervised direct-path relative transfer function learning for binaural sound source localization," in *Proceedings of the ICASSP*, June 6–11, Toronto, Canada (virtual conference), pp. 825–829.
Yang, B., Liu, H., and Li, X. (**2021b**). "Learning deep direct-path relative transfer function for binaural sound source localization," IEEE/ACM Trans. Audio. Speech. Lang. Process. **29**, 3491–3503.
Yang, Z., Li, J., Stoica, P., and Xie, L. (**2018**). "Sparse methods for direction-of-arrival estimation," Academic Press Library Signal Process. **7**, 509–581.
Yang, Z., and Xie, L. (**2015**). "Enhancing sparsity and resolution via reweighted atomic norm minimization," IEEE Trans. Signal Process. **64**(4), 995–1006.
Yasuda, M., Koizumi, Y., Saito, S., Uematsu, H., and Imoto, K. (**2020**). "Sound event localization based on sound intensity vector refined by DNN-based denoising and source separation," in *Proceedings of the ICASSP*, May 4–8, Barcelona, Spain (virtual conference), pp. 651–655.
Yiwere, M., and Rhee, E. J. (**2017**). "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," Int. J. Eng. Res. Appl. **12**(22), 12384–12389.
Youssef, K., Argentieri, S., and Zarader, J. (**2013**). "A learning-based approach to robust binaural sound localization," in *Proceedings of the IEEE/RSJ IROS*, November 3–7, Tokyo, Japan, pp. 2927–2932.
Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (**2017**). "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proceedings of the ICASSP*, March 5–9, New Orleans, LA, pp. 241–245.
Zea, E., and Laudato, M. (**2021**). "On the representation of wavefronts localized in space-time and wavenumber-frequency domains," JASA Express Lett. **1**(5), 054801.
Zermini, A., Yu, Y., Xu, Y., Wang, W., and Plumbley, M. D. (**2016**). "Deep neural network based audio source separation," in *Proceedings of the IMA*, May 18–20, Birmingham, UK.
Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (**2018**). "Mixup: Beyond empirical risk minimization," arXiv:1710.09412.
Zhang, J., Ding, W., and He, L. (**2019a**). "Data augmentation and priori knowledge-based regularization for sound event localization and detection," Technical Report, DCASE 2019 Challenge.
Zhang, Y., Wang, S., Li, Z., Guo, K., Chen, S., and Pang, Y. (**2021**). "Data augmentation and class-based ensembled CNN-Conformer networks for sound event localization and detection," Technical Report, DCASE 2021 Challenge.
Zhang, Y., and Yang, Q. (**2021**). "A survey on multi-task learning," IEEE Trans. Knowl. Data Eng. (published online).
Zhang, Y., Ye, Z., Xu, X., and Hu, N. (**2014**). "Off-grid DOA estimation using array covariance matrix and block-sparse Bayesian learning," Signal Process. **98**, 197–201.
Zhang, W., Zhou, Y., and Qian, Y. (**2019b**). "Robust DoA estimation based on convolutional neural network and time-frequency masking," in *Proceedings of Interspeech*, September 15–19, Graz, Austria, pp. 2703–2707.
Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (**2020**). "A comprehensive survey on transfer learning," Proc. IEEE **109**(1), 43–76.
Zotter, F., and Frank, M. (**2019**). "*Ambisonics: A practical 3D audio theory for recording*," in *Studio Production, Sound Reinforcement, and Virtual Reality* (Springer Nature, New York).