

Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks

Sharath Adavanne ^{1b}, Member, IEEE, Archontis Politis ^{1b}, Member, IEEE, Joonas Nikunen ^{1b}, Member, IEEE, and Tuomas Virtanen ^{1b}, Member, IEEE

Abstract—In this paper, we propose a convolutional recurrent neural network for joint sound event localization and detection (SELD) of multiple overlapping sound events in three-dimensional (3-D) space. The proposed network takes a sequence of consecutive spectrogram time frames as input and maps it to two outputs in parallel. As the first output, the sound event detection (SED) is performed as a multi-label classification task on each time frame producing temporal activity for all the sound event classes. As the second output, localization is performed by estimating the 3-D Cartesian coordinates of the direction-of-arrival (DOA) for each sound event class using multi-output regression. The proposed method is able to associate multiple DOAs with respective sound event labels and further track this association with respect to time. The proposed method uses separately the phase and magnitude component of the spectrogram calculated on each audio channel as the feature, thereby avoiding any method- and array-specific feature extraction. The method is evaluated on five Ambisonic and two circular array format datasets with different overlapping sound events in anechoic, reverberant, and real-life scenarios. The proposed method is compared with two SED, three DOA estimation, and one SELD baselines. The results show that the proposed method is generic and applicable to any array structures, robust to unseen DOA values, reverberation, and low SNR scenarios. The proposed method achieved a consistently higher recall of the estimated number of DOAs across datasets in comparison to the best baseline. Additionally, this recall was observed to be significantly better than the best baseline method for a higher number of overlapping sound events.

Index Terms—Sound event detection, direction of arrival estimation, convolutional recurrent neural network.

I. INTRODUCTION

SOUND event localization and detection (SELD) is the combined task of identifying the temporal activities of each sound event, estimating their respective spatial location

Manuscript received July 1, 2018; revised October 27, 2018 and December 4, 2018; accepted December 4, 2018. Date of publication December 7, 2018; date of current version April 11, 2019. This work was supported by the European Research Council under the European Union's H2020 Framework Program through ERC Grant Agreement 637422 EVERYSOUND. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Walter Kellermann. (*Corresponding author: Sharath Adavanne.*)

S. Adavanne, J. Nikunen, and T. Virtanen are with the Signal Processing Laboratory, Tampere University of Technology, Tampere 33720, Finland (e-mail: sharath.adavanne@tut.fi; joonas.s.nikunen@gmail.com; tuomas.virtanen@tut.fi).

A. Politis is with the Department of Signal Processing and Acoustics, Aalto University, Espoo 02150, Finland (e-mail: archontis.politis@aalto.fi).

Digital Object Identifier 10.1109/JSTSP.2018.2885636

trajectories when active, and further associating textual labels with the sound events. Such a method can for example automatically describe social and human activities and assist the hearing impaired to visualize sounds. Robots can employ this for navigation and natural interaction with surroundings [1]–[4]. Smart cities, smart homes, and industries could use it for audio surveillance [5]–[8]. Smart meeting rooms can recognize speech among other events and use this information to beamform and enhance the speech for teleconferencing or for robust automatic speech recognition [9]–[13]. Naturalists could use it for bio-diversity monitoring [14]–[16]. Further, in virtual reality (VR) applications with 360° audio SELD can be used to assist the user in visualizing sound events.

A. Sound Event Detection

The SELD task can be broadly divided into two sub-tasks, sound event detection (SED) and sound source localization. SED aims at detecting temporally the onsets and offsets of sound events and further associating textual labels to the detected events. The sound events in real-life most often overlap with other sound events in time and the task of recognizing all the overlapping sound events is referred as polyphonic SED. The SED task in literature has most often been approached using different supervised classification methods that predict the framewise activity of each sound event class. Some of the classifiers include Gaussian mixture model (GMM) - hidden Markov model (HMM) [27], fully connected (FC) neural networks [28], recurrent neural networks (RNN) [29]–[32], and convolutional neural networks (CNN) [33], [34]. More recently state-of-the-art results were obtained by stacking CNN, RNN and FC layers consecutively, referred jointly as the convolutional recurrent neural network (CRNN) [35]–[39].

Lately, in order to improve recognition of overlapping sound events, several multichannel SED methods have been proposed [39]–[43] and these were among the top performing methods in the real-life SED task of DCASE 2016¹ and 2017² evaluation challenges. More recently, we studied the SED performance on identical sound scenes captured using single, binaural and

¹<http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-sound-event-detection-in-real-life-audio#system-characteristics>

²<http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-sound-event-detection-in-real-life-audio-results#system-characteristics>

first-order Ambisonics (FOA) microphones [35], where the order denotes the spatial resolution of the format and the first order corresponds to four channels. The results showed that the recognition of overlapping sound events improved with increase in spatial sampling, and the best performance was obtained with FOA.

B. Sound Source Localization

Sound source localization is the task of determining the direction or position of a sound source with respect to the microphone. In this paper, we only deal with the estimation of the sound event direction, generally referred as direction-of-arrival (DOA) estimation. The DOA methods in literature can be broadly categorized into parametric- and deep neural network (DNN)-based approaches. Some popular parametric methods are based on time-difference-of-arrival (TDOA) [44], the steered-response-power (SRP) [45], multiple signal classification (MUSIC) [46], and the estimation of signal parameters via rotational invariance technique (ESPRIT) [47]. These methods vary in terms of algorithmic complexity, constraints in array geometry, and model assumptions on the acoustic scenarios. Subspace methods like MUSIC can be applied with different array types and can produce high-resolution DOA estimates of multiple sources. On the other hand, subspace methods require a good estimate of the number of active sources that may be hard to obtain, and they have been found sensitive to reverberant and low signal-to-noise (SNR) scenarios [48].

Recently, DNN-based methods were employed to overcome some of the drawbacks of parametric methods, while being robust towards reverberation and low SNR scenarios. Additionally, implementing the localization task in the DNN framework allows seamless integration into broader DNN tasks such as SELD [20], robots can use it for sound source based navigation and natural interaction in multi-speaker scenarios [1]–[4]. A summary of the most recent DNN-based DOA estimation methods is presented in Table I. All these methods estimate DOAs for static point sources and were shown to perform equally or better than the parametric methods in reverberant scenarios. Further, methods [4], [18], [20], [25] proposed to simultaneously detect DOAs of overlapping sound events by estimating the number of active sources from the data itself. Most methods used a classification approach, thereby estimating the source presence likelihood at a fixed set of angles, while [22], [23] used a regression approach and let the DNN produce continuous output.

All of the past works were evaluated on different array geometries, making a direct performance comparison difficult. Most of the methods estimated full azimuth ('Full' in Table I) using microphones mounted on a robot, circular and distributed arrays, while the rest of the methods used linear arrays thereby estimating only the azimuth angles in a range of 180° . Although few of the existing methods estimated the azimuth and elevation jointly [24], [25], most of them estimated only the azimuth angle [1]–[4], [17]–[20]. In particular, we studied the joint estimation of azimuth and elevation angles in [25], this was enabled by the use of Ambisonic signals (FOA) obtained

using a spherical array. Ambisonics are also known as spherical harmonic (SH) signals in the array processing literature, and they can be obtained from various array configurations such as circular or planar (for 2D capture) and spherical or volumetric (for 3D capture) using an appropriate linear transform of the recordings [49]. The same ambisonic channels have the same spatial characteristics independent of the recording setup, and hence, studies on such hardware-independent formats make the evaluation and results more easily comparable in the future.

Most of the previously proposed DNN-based DOA estimation methods that relied on a single array or distributed arrays of omnidirectional microphones, captured source location information mostly in phase- or time-delay differences between the microphones. However, compact microphone arrays with full azimuth and elevation coverage, such as spherical microphone arrays, rely strongly on the directionality of the sensors to capture spatial information, this reflects mainly in the magnitude differences between channels. Motivated by this fact we proposed to use both the magnitude and phase component of the spectrogram as input features in [25]. Thus making the DOA estimation method [25] generic to array configuration by avoiding method-specific feature extractions like inter-aural level difference (ILD), the inter-aural time difference (ITD), generalized cross-correlation (GCC) or eigenvectors of spatial covariance matrix used in previous methods (Table I).

C. Joint Localization and Detection

In the presence of multiple overlapping sound events, the DOA estimation task becomes the classical tracking problem of associating correctly the multiple DOA estimates to respective sources, without necessarily identifying the source [50], [51]. The problem is further extended for the polyphonic SELD task if the SED and DOA estimation are done separately, resulting in the data association problem between the recognized sound events and the estimated DOAs [13]. One solution to the data association problem is to jointly predict the SED and DOA. In this regard, to the best of the authors' knowledge, [20] is the only DNN-based method which performs SELD. Other works combining SED and parametric DOA estimation include [6], [13], [52], [53]. Lopatka *et al.* [53] used a 3D sound intensity acoustic vector sensor, with MPEG-7 spectral and temporal features along with a support vector machine classifier to estimate DOA along azimuth for five classes of non-overlapping sound events. Butko *et al.* [13] used distributed microphone arrays to recognize 14 different sound events with an overlap of two at a time, using a GMM-HMM classifier, and localized them inside a meeting room using the SRP method. Chakraborty *et al.* [52] replaced SRP-based localization in [13] with a sound-model-based localization, thereby fixing the data association problem faced in [13]. In contrast, Hirvonen [20], extracted the frame-wise spectral power from each microphone of a circular array and used a CNN classifier to map it to eight angles in full azimuth for each sound event class in the dataset. In this output format, the resolution of azimuth is limited to the trained directions and the performance of unseen DOA values is unknown.

TABLE I

SUMMARY OF DNN BASED DOA ESTIMATION METHODS IN THE LITERATURE. THE AZIMUTH AND ELEVATION ANGLES ARE DENOTED AS ‘AZI’ AND ‘ELE’, DISTANCE AS ‘DIST’, ‘X’ AND ‘Y’ REPRESENT THE DISTANCE ALONG THE RESPECTIVE CARTESIAN AXIS. ‘FULL’ REPRESENTS THE ESTIMATION IN THE COMPLETE RANGE OF THE RESPECTIVE FORMAT, AND ‘REGRESSION’ REPRESENTS THE CLASSIFIER ESTIMATION TYPES

Approach	Input feature	Output format	Sources	DNN	Array	SELD
Chakrabarty et al. [17, 18]	Phase spectrum	azi	1, multiple	CNN	Linear	×
Yalta et al. [3]	Spectral power	azi (Full)	1	CNN Resnet	Robot	×
Xiao et al. [19]	GCC	azi (Full)	1	FC	Circular	×
Takeda et al. [1, 2]	Eigen vectors of spatial covariance matrix	azi (Full)	1, 2	FC	Robot	×
He et al. [4]	GCC	azi (Full)	Multiple	CNN	Robot	×
Hirvonen [20]	Spectral power	azi (Full) for each class	Multiple	CNN	Circular	✓
Yiwere et al. [21]	ILD, cross-correlation	azi and dist	1	FC	Binaural	×
Ferguson et al. [22]	GCC, cepstrogram	azi and dist (regression)	1	CNN	Linear	×
Vesperini et al. [23]	GCC	x and y (regression)	1	FC	Distributed	×
Sun et al. [24]	GCC	azi and ele	1	PNN	Cartesian	×
Adavanne et al. [25]	Phase and magnitude spectrum	azi and ele (Full)	Multiple	CRNN	Generic	×
Roden et al. [26]	ILD, ITD, phase and magnitude spectrum	azi, ele and dist (separate NN)	1	FC	Binaural	×
Proposed	Phase and magnitude spectrum	azi and ele (Full, regression) for each class	Multiple	CRNN	Generic	✓

For larger datasets with a higher number of sound events and increased resolution along azimuth and elevation directions, this approach results in a large number of output nodes. Training such a DNN with a large number of output nodes where the number of positive class labels per frame is one or two with respect to a high number of negative class labels poses challenges of an imbalanced dataset. Additionally, training such a large number of classes requires a huge dataset with enough examples for each class. On the other hand, this output format allows the network to simultaneously recognize more than one instance of the same sound event in a given time frame, at different locations.

D. Contributions of this Paper

In general, the number of existing SELD methods is limited [6], [13], [20], [52], [53], with only one published DNN-based approach [20]. On the other hand, there are several DNN-based methods in the literature for the SELD sub-tasks of SED and DOA estimation. Yet, there is no comprehensive work published that studies the various choices affecting the performance of these DNN-based SED, DOA and SELD methods, compare them with multiple competitive baselines, and evaluate them over a wide range of acoustic conditions. Besides, with respect to the SELD task, the existing methods [6], [13], [52], [53] localize up to one or maximum two overlapping sound events and do not scale to a higher number of overlapping sources. Further, the only DNN-based SELD method [20] localizes sound events exclusively at a predefined grid of directions and requires a large number of output classes for a higher number of sound event labels and increased spatial resolution. Additionally, all the above SELD approaches use method-specific features and hence not independent of input array structure.

In contrast to existing SELD methods, this paper presents novelty in two broad areas: the proposed SELD method, and the exhaustive evaluation studies presented. The novelty of the proposed SELD method is as follows. It is the first method that addresses the problem of localizing and recognizing more

than two overlapping sound events simultaneously and tracking their activity with respect to time. The proposed method is able to localize sources at any azimuth and elevation angles while being robust to unseen spatial locations, reverberation, and ambiance. Further, the method itself is generic enough to learn to perform SELD from any input array structure. Specifically, as our method, we propose to use the polyphonic SED output [39] as a confidence measure for choosing the DOAs estimated in a regression manner. By this approach, we not only extend the state-of-the-art polyphonic SED performance [39] for polyphonic SELD but also tackle the data-association problem faced due to the polyphony in SELD tasks [13]. As the second broad area of novelty, we present the performance of the proposed method with respect to various design choices made such as the DNN architecture, input feature and DOA output format. Additionally, we also present the comprehensive results of the proposed method with respect to six baselines (two SED, three DOA estimation, and one SELD baseline) evaluated on seven datasets with different acoustic conditions (anechoic and reverberant scenarios with simulated and real-life impulse responses), array configurations (Ambisonic and circular array) and the number of overlapping sound events.

In order to facilitate reproducibility of research, the proposed method and all the datasets used have been made publicly available.³ Additionally, the real-life impulse responses used to simulate datasets have also been published to enable users to experiment with custom sound events.

The rest of the paper is organized as follows. In Section II, we describe the proposed SELD method and the training procedure. In Section III, we describe the datasets, the baseline methods, the metrics and the experiments carried out for evaluating the proposed method. The experimental results on the evaluation datasets are presented, compared with baselines and discussed in Section IV. Finally, we summarize the conclusions of the work in Section V.

³<https://github.com/sharathadavanne/seld-net>

II. METHOD

The block diagram of the proposed method for SELD is presented in Fig. 1(a). The input to the method is the multichannel audio. The phase and magnitude spectrograms are extracted from each audio channel and are used as separate features. The proposed method takes a sequence of features in consecutive spectrogram frames as input and predicts all the sound event classes active for each of the input frames along with their respective spatial location, producing the temporal activity and DOA trajectory for each sound event class. In particular, a CRNN is used to map the feature sequence to the two outputs in parallel. At the first output, SED is performed as a multi-label classification task, allowing the network to simultaneously estimate the presence of multiple sound events for each frame. At the second output, DOA estimates in the continuous 3D space are obtained as a multi-output regression task, where each sound event class is associated with three regressors that estimate the 3D Cartesian coordinates x , y and z of the DOA on a unit sphere around the microphone. The SED output of the network is in the continuous range of $[0, 1]$ for each sound event in the dataset, and this value is thresholded to obtain a binary decision for the respective sound event activity as shown in Fig. 1(b). Finally, the respective DOA estimates for these active sound event classes provide their spatial locations. The detailed description of the feature extraction and the proposed method is explained in the following sections.

A. Feature Extraction

The spectrogram is extracted from each of the C channels of the multichannel audio using an M -point discrete Fourier transform (DFT) on Hamming window of length M and 50% overlap. The phase and magnitude of the spectrogram are then extracted and used as separate features. Only the $M/2$ positive frequencies without the zeroth bin are used. The output of the feature extraction block in Fig. 1(a) is a feature sequence of T frames, with an overall dimension of $T \times M/2 \times 2C$, where the $2C$ dimension consists of C magnitude and C phase components.

B. Neural Network Architecture

The output of the feature extraction block is fed to the neural network as shown in Fig. 1(a). In the proposed architecture the local shift-invariant features in the spectrogram are learned using multiple layers of 2D CNN. Each CNN layer has P filters of $3 \times 3 \times 2C$ (as in [25]) dimensional receptive fields acting along the time-frequency-channel axis with a rectified linear unit (ReLU) activation. The use of filter kernels spanning all the channels allows the CNN to learn relevant inter-channel features required for localization, whereas the time and frequency dimensions of the kernel allows learning relevant intra-channel features suitable for both the DOA and SED tasks. After each layer of CNN, the output activations are normalized using batch normalization [54], and the dimensionality is reduced using max-pooling (MP_i) along the frequency axis, thereby keeping the sequence length T unchanged. The output after the final CNN layer with P filters is of dimension $T \times 2 \times P$, where the

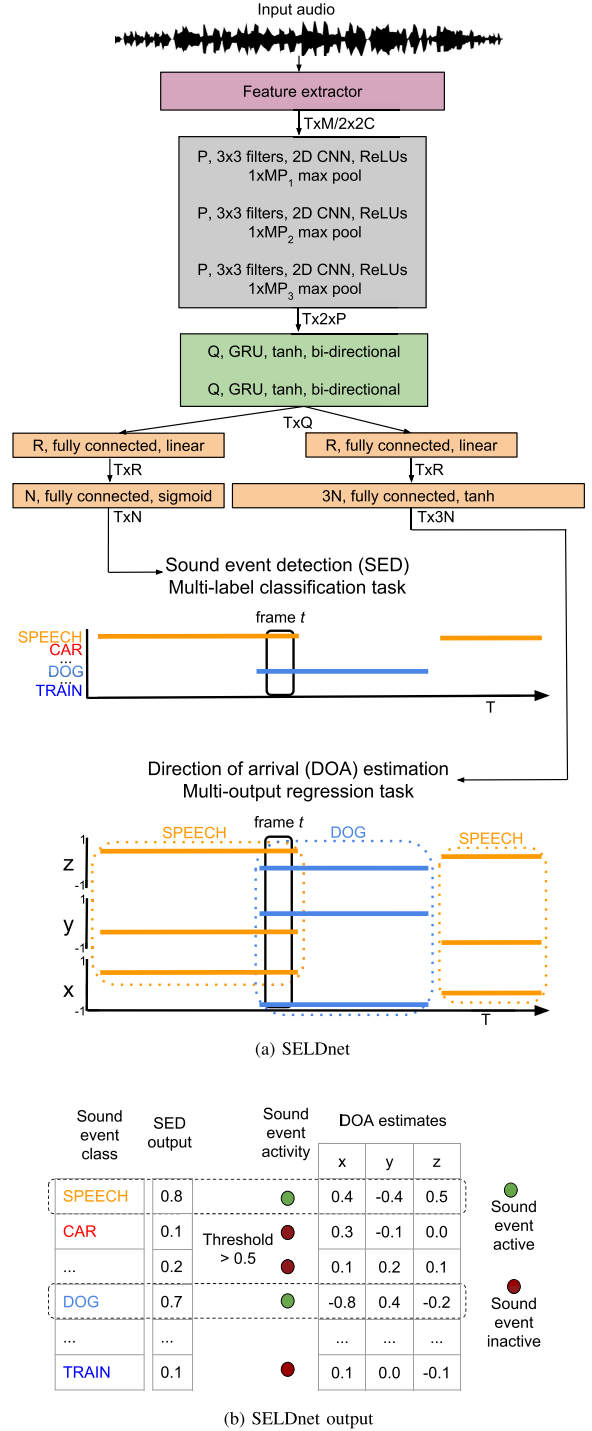


Fig. 1. a) The proposed SELDnet and b) the frame-wise output for frame t in Figure a). A sound event is said to be localized and detected when the confidence of the SED output exceeds the threshold.

reduced frequency dimension of 2 is a result of max-pooling across CNN layers (see Section IV).

The output activation from CNN is further reshaped to a T frame sequence of length $2P$ feature vectors and fed to bidirectional RNN layers which are used to learn the temporal context information from the CNN output activations. Specifically, Q nodes of gated recurrent units (GRU) are used in each layer

with tanh activations. This is followed by two branches of FC layers in parallel, one each for SED and DOA estimation. The FC layers share weights across time steps. The first FC layer in both the branches contains R nodes each with linear activation. The last FC layer in the SED branch consists of N nodes with sigmoid activation, each corresponding to one of the N sound event classes to be detected. The use of sigmoid activation enables multiple classes to be active simultaneously. The last FC layer in the DOA branch consists of $3N$ nodes with tanh activation, where each of the N sound event classes is represented by 3 nodes corresponding to the sound event location in x , y , and z , respectively. For a DOA estimate on a unit sphere centered at the origin, the range of location along each axis is $[-1, 1]$, thus we use the tanh activation for these regressors to keep the output of the network in a similar range.

We refer to the above architecture as SELDnet. The SED output of the SELDnet is in the continuous range of $[0, 1]$ for each class, while the DOA output is in the continuous range of $[-1, 1]$ for each axes of the sound class location. A sound event is said to be active, and its respective DOA estimate is chosen if the SED output exceeds the threshold of 0.5 as shown in Fig. 1(b). The network hyperparameters are optimized based on cross-validation as explained in Section III-D1.

C. Training Procedure

In each frame, the target values for each of the active sound events in the SED branch output are one while the inactive events are zero. Similarly, for the DOA branch, the reference DOA x , y , and z values are used as targets for the active sound events and $x = 0, y = 0$, and $z = 0$ is used for inactive events. A binary cross-entropy loss is used between the SED predictions of SELDnet and reference sound class activities, while a mean square error (MSE) loss is used for the DOA estimates of the SELDnet and the reference DOA. By using the MSE loss for DOA estimation in 3D Cartesian coordinates we truly represent the distance between two points in space. The distance between two points (x_1, y_1, z_1) and (x_2, y_2, z_2) in 3D space is given by \sqrt{SE} , where $SE = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2$, while the MSE between the same points is given by $SE/3$. Thus the MSE loss is simply a scaled version of the distance in 3D space, and reducing the MSE loss implies the reduction in the distance between the two points.

Theoretically, the advantage of using Cartesian coordinates instead of azimuth and elevation for regression can be observed when predicting DOA in full azimuth and/or full elevation. The angles are discontinuous at the wrap-around boundary (for example the $-180^\circ, 180^\circ$ boundary for azimuth), while the Cartesian coordinates are continuous. This continuity allows the network to learn better. Further experiments on this are discussed in Section III-D.

We train the SELDnet with a weighted combination of MSE and binary cross-entropy loss for 1000 epochs using Adam optimizer with default parameters as used in the original paper [55]. Early stopping is used to control the network from over-fitting to training split. The training is stopped if the SELD score (Section III-C) on the test split does not improve for 100 epochs.

TABLE II
SUMMARY OF DATASETS

Audio format	Sound scene	Impulse response	Dataset acronym	Train/Test, notes
Ambisonic (four channel)	Anechoic	Synthetic	ANSYN	240/60
	Reverberant		RESYN	
		Real life	REAL	600/150
	REALBIG		600/150, ambiance	
Circular array (eight channel)	Anechoic	Synthetic	CANSYN	240/60
	Reverberant		CRESYN	

The network was implemented using Keras library [56] with TensorFlow [57] backend.

III. EVALUATION

A. Datasets

The proposed SELDnet is evaluated on seven datasets that are summarized in Table II. Four of the datasets are synthesized with artificial impulse responses (IR), that consists of anechoic and reverberant scenarios virtually recorded both with a circular array and in the Ambisonics format. Three of the datasets are synthesized with real-life impulse responses, recorded with a spherical array and encoded into the Ambisonics format. All the datasets consist of stationary point sources each associated with a spatial coordinate. The synthesis procedure in all the datasets consists of mixing isolated sound event instances at different spatial locations, since this allows producing the reference event locations and times of activity for evaluation and training of the methods.

1) *TUT Sound Events 2018 - Ambisonic, Anechoic and Synthetic Impulse Response (ANSYN) Dataset*: This dataset consists of spatially located sound events in an anechoic environment synthesized using artificial IRs. It comprises three subsets: no temporally overlapping sources ($O1$), maximum two temporally overlapping sources ($O2$) and maximum three temporally overlapping sources ($O3$). Each of the subsets consists of three cross-validation splits with 240 training and 60 testing FOA format recordings of length 30 s sampled at 44100 Hz. The dataset is generated using the 11 isolated sound event classes from the DCASE 2016 task 2 dataset [58] such as speech, coughing, door slam, page-turning, phone ringing and keyboard. Each of these sound classes has 20 examples, of which 16 are randomly chosen for the training set and the rest four for the testing set, amounting to 176 examples from 11 classes for training, and 44 for testing. During synthesis of a recording, a random collection of examples are chosen from the respective set and are randomly placed in a spatial grid of 10° resolution along azimuth and elevation, such that two overlapping sound events are separated by 10° , and the elevation is in the range of $[-60^\circ, 60^\circ]$. In order to have a variability of amplitude, the sound events are randomly placed at a distance ranging from 1 to 10 m with 0.5 m resolution from the microphone. More details regarding the synthesis can be found in [25].

2) *TUT Sound Events 2018 - Ambisonic, Reverberant and Synthetic Impulse Response (RESYN) Dataset*: This dataset is synthesized with the same details as the ambisonic ANSYN

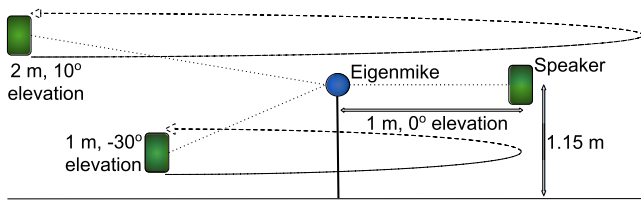


Fig. 2. Recording real-life impulse responses for sound scene generation. A person walks around the Eigenmike⁴ holding a Genelec loudspeaker⁵ playing a maximum length sequence at different elevation angles and distances.

dataset, with the only difference being that the sound events are spatially placed within a room using the image source method [59]. Specifically, the microphone is placed at the center of the room, and the sound events are randomly placed around the microphone, with their distance ranging from 1 m from the microphone to the respective end of the room at 0.5 m resolution. The three cross-validation splits of each of the three subsets $O1$, $O2$ and $O3$ are generated for a moderately reverberant room of size $10 \times 8 \times 4$ m (Room 1), with reverberation times 1.0, 0.8, 0.7, 0.6, 0.5, and 0.4 s per each octave band, and 125 Hz–4 kHz band center frequencies. Additionally, to study the performance in mismatched reverberant scenarios, testing splits are generated for two different sized rooms: room 2 that is 80% the volume ($8 \times 8 \times 4$ m) and reverberation-time per band of room 1, and room 3 that is 125% the volume ($10 \times 10 \times 4$ m) and reverberation-time per band of room 1. In order to remove any ambiguity while comparing the performance difference of room 1 with room 2 and 3, we keep the sound events and their respective spatial locations in room 2 and 3 identical to the testing split of room 1. But the individual sound events whose distance from the microphone exceeded the room size were reassigned a new distance within the room. Further details on the reverberant synthesis can be read in [25].

3) *TUT Sound Events 2018 - Ambisonic, Reverberant and Real-life Impulse Response (REAL) Dataset*: In order to study the performance of SELDnet in a real-life scenario, we generated a dataset by collecting impulse responses from a real environment using the Eigenmike⁴ spherical microphone array. For the IR acquisition, we used a continuous measurement signal as in [60]. The measurement was done by slowly moving a Genelec G Two loudspeaker⁵ continuously playing a maximum length sequence around the array in circular trajectory in one elevation at a time, as shown in Fig. 2. The playback volume was set to be 30 dB greater than the ambient sound level. The recording was done in a corridor inside the university with classrooms around it.

The moving-source IRs were obtained by a freely available tool from CHiME challenge [61] which estimates the time-varying responses in STFT domain by forming a least-squares regression between the known measurement signal and the far-field recording independently at each frequency. The IR for any azimuth within one trajectory can be analyzed by assuming block-wise stationarity of acoustic channel. The average angular

speed of the loudspeaker in the measurements was $6^\circ/s$ and we used a block size of 860 ms (81 STFT frames with analysis frame size of 1024 with 50% overlap and sample rate $F_s = 48$ kHz) for estimation of IR of length 170 ms (16 STFT frames).

The IRs were collected at elevations -40° to 40° with 10° increments at 1 m from the Eigenmike and at elevations -20° to 20° with 10° increments at 2 m. For the dataset creation, we analyzed the DOA of each time frame using MUSIC and extracted IRs for azimuthal angles at 10° resolution (36 IRs for each elevation). The IR estimation tool [61] was applied independently on all 32 channels of the Eigenmike.

In order to synthesize the sound scene from the estimated IRs, we used isolated real-life sound events from the urbansound8k dataset [62]. This dataset consists of 10 sound event classes such as: air conditioner, car horn, children playing, dog barking, drilling, engine idling, gunshot, jackhammer, siren and street music. Among these, we did not include children playing and air conditioner classes since these can also occur in our ambient recording which we use as background recording in dataset REALBIGAMB (Section III-A5). From the sound examples in urbansound8k, we only used the ones marked as foreground in order to have clean isolated sound events. Similarly to the other datasets used in this paper, we used the splits 1, 8 and 9 provided in the urbansound8k as the three cross-validation splits. These splits were chosen as they had a good number of examples for all the chosen sound event classes after selecting only the foreground examples. The final selected examples varied in length from 100 ms to 4 s and amount to 15671.5 seconds from 4542 examples.

During the sound scene synthesis, we randomly chose a sound event example and associated it with a random distance among the collected ones, azimuth and elevation angle. The sound event example was then convolved with the respective IR for the given distance, azimuth and elevation to spatially position it. Finally, after positioning all the sound events in a recording we converted this multichannel audio to FOA format. The transform of the microphone signals to FOA was performed using the tools published in [63]. In total, we generated 300 such 30 s recordings in a similar fashion as ANSYN and RESYN with 240 of them earmarked for training and 60 for testing. Similar to the ANSYN recordings we also generated three subsets $O1$, $O2$ and $O3$ with a different number of overlapping sound events.

4) *TUT Sound Events 2018 - Ambisonic, Reverberant and Real-life Impulse Response big (REALBIG) Dataset*: In order to study the performance of SELDnet with respect to the size of the dataset, we generated for each of three ambisonic REAL subsets a 750 recordings REALBIG subset of 30 s length, with 600 for training and 150 for testing.

5) *TUT Sound Events 2018 - Ambisonic, Reverberant, Real-life Impulse Response and Ambiance big (REALBIGAMB) Dataset*: Additionally, to simulate a real sound-scene we recorded 30 min of ambient sound to use as background noise in the same location as the IR recordings without changing the setup. We mixed randomly chosen segments of the recorded ambiance at three different SNRs: 0, 10 and 20 dB for each of the three ambisonic REALBIG subsets and refer to it as

⁴<https://mhacoustics.com/products>

⁵<https://www.genelec.com/home-speakers/g-series-active-speakers>

TABLE III
BASELINE AND PROPOSED METHOD SUMMARY

Task	Acronym	Notes	Datasets evaluated
SED	SEDnet [39]	Single channel	All
	MSEDnet [39]	Multichannel	
DOA	MUSIC*	Azi and ele	All except CANSYN and CRESYN
	DOAnet [25]	Azi and ele	
	AZInet [18]	Azi	CANSYN and CRESYN
SELD	HIRnet [20]	Azi	All
	SELDnet-azi	Azi	
	SELDnet	Azi and ele	

*Parametric, all other methods are DNN based

REALBIGAMB subsets. The ambiance used for the testing set was kept separate from the training set.

6) *TUT Sound Events 2018 - Circular array, Anechoic and Synthetic Impulse Response (CANSYN) Dataset*: To study the performance of SELDnet on generic array configurations, similarly to the SELD baseline method [20] (Section III-B3), we synthesized the ANSYN recordings for a circular array of radius 5 cm with eight omnidirectional microphones at 0, 45, 90, 135, 180, 225, 270, 315°, and the array plane parallel to the ground, and refer to it as CANSYN. It is an exact replica of the ANSYN dataset in terms of the synthesized sound events except for the microphone array setup, and hence the number of channels. Similar to ANSYN, the CANSYN dataset has three subsets with a different number of overlapping sound events each with three cross-validation splits.

7) *TUT Sound Events 2018 - Circular array, Reverberant and Synthetic Impulse Response (CRESYN) Dataset*: Similar to the CANSYN dataset, we synthesize the circular array version of ambisonic RESYN room 1 dataset, referred as CRESYN. During synthesis, the circular microphone array is placed in the center of the room, and the array plane parallel to the floor.

B. Baseline Methods

The SELDnet is compared with six different baselines as summarized in Table III: two SED baselines (single- and multichannel), three DOA baselines (parametric and DNN-based), and a SELD baseline.

1) *SED Baseline*: The SED capabilities of the proposed SELDnet is compared with the existing state-of-the-art multichannel SED method [39], referred here as MSEDnet. MSEDnet is easily scalable to any number of input audio channels and won [38] the recently concluded real-life SED task in DCASE 2017 [64]. In particular, it won the top two positions among 34 submissions, first using single-channel mode (referred as SEDnet) and a close second using multichannel mode. The SED performance of SELDnet is compared with both the single- and the multichannel modes of MSEDnet.

In the original MSEDnet implementation [39] the input is a sequence of log mel-band energy (40-bands) frames, that are mapped to an equal-length sequence of sound event activities. The SED metrics (Section III-C) for MSEDnet did not change much on using phase and magnitude components of the STFT spectrogram instead of log mel-band energies. Hence, in order to have a one-to-one comparison with SELDnet, we use the

phase and magnitude components of the STFT spectrogram for MSEDnet in this paper. We train the MSEDnet for 500 epochs and use early stopping when SED score (Section III-C) stops improving for 100 epochs.

2) *DOA Baseline*: The DOA estimation performance of the SELDnet is evaluated with respect to three baselines. As a parametric baseline, we use MUSIC [46] and as DNN-based baselines, we use the recently proposed DOAnet [25] that estimates DOAs in 3D and [18] that estimates only the DOA azimuth angle referred as AZInet.

i) *MUSIC*: is a versatile high-resolution subspace method that can detect multiple narrowband source DOAs and can be applied to generic array setups. It is based on a subspace decomposition of the spatial covariance matrix of the multichannel spectrogram. For a broadband estimation of DOAs, we combine narrowband spatial covariance matrices over three frames and frequency bins from 50 to 8000 Hz. The steering vector information required to produce the MUSIC pseudo-spectrum from which the DOAs are extracted is adapted to the recording system under use, meaning uniform circular array steering vectors for CANSYN and CRESYN datasets, and real SH vectors for all the other ambisonic datasets.

MUSIC requires a good estimate of the number of active sound sources in order to estimate their DOAs. In this paper, we use MUSIC with the number of active sources taken from the reference of the dataset. Hence, the DOA estimation results of MUSIC can be considered as the best possible for the given dataset and serve as a benchmark for DOA estimation with and without the knowledge of the number of active sources. For a detailed description on MUSIC and other subspace methods, the reader is referred to [65], while for application of MUSIC to SH signals similar to this work, please refer to [66].

ii) *DOAnet*: Among the recently proposed DNN-based DOA estimation methods listed in Table I, the only method that attempts DOA estimation of multiple overlapping sources in 3D space is the DOAnet [25]. Hence, DOAnet serves as a suitable baseline to compare against the DOA estimation performance of the proposed SELDnet. DOAnet is based on a similar CRNN architecture, the input to which is a sequence of multichannel phase and magnitude spectrum frames. It considers DOA estimation as a multi-label classification task by directional sampling with a resolution of 10° along azimuth and elevation and estimating the likelihood of a sound source being active in each of these points.

iii) *AZInet*: Among the DOA-only estimation methods listed in Table I, apart from the DOAnet [25], methods [18] and [4] are the only ones which attempt simultaneous DOA estimation of overlapping sources. Since [4] is evaluated on a dataset collected using microphones mounted on a humanoid robot, it is difficult to replicate the setup. Hence in this paper, we use the AZInet evaluated on a linear array in [18] as the baseline. The AZInet is a CNN-based method that uses the phase component of the spectrogram of each channel as input, and maps it to azimuth angles in the range 0° to 180° at 5° resolution as a multi-label classification task. AZInet uses only the phase spectrogram since the dataset evaluated on employs omnidirectional microphones, which for compact arrays and sources in the far-field, preserve

directional information in inter-channel phase differences. Thus, although the evaluation in [18] was carried out on a linear array, the method is generic to any omnidirectional array under these conditions. Further, in order to have a direct comparison, we extend the output of AZInet to full-azimuth with 10° resolution and reduce the output of SELDnet to generate only the azimuth, i.e., we only estimate x and y coordinates of the DOA (SELDnet-azi). To enable this full-azimuth estimation we use the circular array with omnidirectional microphones datasets CANSYN and CRESYN.

3) *SELD Baseline (HIRnet)*: The joint SED and DOA estimation performance of SELDnet is compared with the method proposed by Hirvonen [20], hereafter referred to as HIRnet. The HIRnet was proposed for a circular array of omnidirectional microphones, hence we compare its performance only on the CANSYN and CRESYN datasets. HIRnet is a CNN-based network that uses the log-spectral power of each channel as the input feature and maps it to eight angles in full azimuth for each of the two classes (speech and music) as a multi-label classification task. More details about HIRnet can be found in [20]. In order to have a direct comparison with SELDnet-azi, we extend HIRnet to estimate DOAs at a 10° resolution for each of the sound event classes in our testing datasets.

C. Evaluation Metrics

The proposed SELDnet is evaluated using individual metrics for SED and DOA estimation. For SED, we use the standard polyphonic SED metrics, error rate (ER) and F-score calculated in segments of one second with no overlap as proposed in [67], [68]. The segment-wise results are obtained from the frame-level predictions of the classifier by considering the sound events to be active in the entire segment if it is active in any of the frames within the segment. Similarly, we obtain labels for one-second segments of reference from its frame-wise annotation, and calculate the segment-wise ER and F-scores. Mathematically, the F-score is calculated as follows:

$$F = \frac{2 \cdot \sum_{k=1}^K TP(k)}{2 \cdot \sum_{k=1}^K TP(k) + \sum_{k=1}^K FP(k) + \sum_{k=1}^K FN(k)}, \quad (1)$$

where the number of true positives $TP(k)$ is the total number of sound event classes that were active in both reference and predictions for the k th one-second segment. The number of false positives in a segment $FP(k)$ is the number of sound event classes that were active in the prediction but were inactive in the reference. Similarly, $FN(k)$ is the number of false negatives, i.e. the number of sound event classes inactive in the predictions but active in the reference.

The ER metric is calculated as

$$ER = \frac{\sum_{k=1}^K S(k) + \sum_{k=1}^K D(k) + \sum_{k=1}^K I(k)}{\sum_{k=1}^K N(k)}, \quad (2)$$

where, for each one-second segment k , $N(k)$ is the total number of active sound event classes in the reference. Substitution $S(k)$ is the number of times an event was detected but given the wrong level, this is obtained by merging the false negatives and false positives without individually correlating which false positive

substitutes which false negative. The remaining false positives and false negatives, if any, are counted as insertions $I(k)$ and deletions $D(k)$ respectively. These statistics are mathematically defined as follows:

$$S(k) = \min(FN(k), FP(k)), \quad (3)$$

$$D(k) = \max(0, FN(k) - FP(k)), \quad (4)$$

$$I(k) = \max(0, FP(k) - FN(k)). \quad (5)$$

An SED method is jointly evaluated using the F-score and ER metric, and an ideal method will have an F-score of one (reported as percentages in Table) and ER of zero. More details regarding the F-score and ER metric can be read in [67], [68].

The predicted DOA estimates (x_E, y_E, z_E) are evaluated with respect to the reference (x_G, y_G, z_G) used to synthesize the dataset, utilizing the central angle $\sigma \in [0, 180]$. The σ is the angle formed by (x_E, y_E, z_E) and (x_G, y_G, z_G) at the origin in degrees, and is given by

$$\sigma = 2 \cdot \arcsin \left(\frac{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}{2} \right) \cdot \frac{180}{\pi}, \quad (6)$$

where, $\Delta x = x_G - x_E$, $\Delta y = y_G - y_E$, and $\Delta z = z_G - z_E$. The DOA error for the entire dataset is then calculated as

$$DOA\ error = \frac{1}{D} \cdot \sum_{d=1}^D \sigma((x_G^d, y_G^d, z_G^d), (x_E^d, y_E^d, z_E^d)) \quad (7)$$

where D is the total number of DOA estimates across the entire dataset, and $\sigma((x_G^d, y_G^d, z_G^d), (x_E^d, y_E^d, z_E^d))$ is the angle between d -th estimated and reference DOAs.

Additionally, in order to account for time frames where the number of estimated and reference DOAs are unequal, we report the frame recall, calculated as $TP/(TP + FN)$ in percentage, where true positives TP is the total number of time frames in which the number of DOAs predicted is equal to reference, and false negatives FN is the total number of frames where the predicted and reference DOA are unequal.

The DOA estimation method is jointly evaluated using the DOA error and the frame recall, and an ideal method has a frame recall of one (reported as percentages in Table) and DOA error of zero.

During the training of SELDnet, we perform early stopping based on the combined SELD score calculated as

$$SELD\ score = (SED\ score + DOA\ score)/2, \quad (8)$$

where

$$SED\ score = (ER + (1 - F))/2, \quad (9)$$

$$DOA\ score = (DOA\ error/180 + (1 - frame\ recall))/2, \quad (10)$$

and an ideal SELD method will have an SELD score of zero. In the proposed method, the localization performance is dependent on the detection performance. This relation is represented by the frame recall metric of DOA estimation. As a consequence, the SELD score which is comprised of frame recall metric in addition to the SED metrics can be seen to weigh the SED performance more than DOA.

D. Experiments

The SELDnet is evaluated in different dimensions to understand its potential and drawbacks. The experiments carried out with different datasets in this regard are explained below.

1) *SELDnet Architecture and Model Parameter Tuning*: A wide variety of architectures with different combinations of CNN, RNN and FC layers are explored on the ANSYN O2 subset with frame length $M = 1024$ (23.2 ms). Additionally, for each architecture, we tune the model parameters such as the number of CNN, RNN, and FC layers (0 to 4) and nodes (in the set of [16, 32, 64, 128, 256, 512]). The input sequence length is tuned in the set of [32, 64, 128, 256, 512], the DOA and SED branch output loss weights in the set of [1, 5, 50, 500], the regularization (dropout in the set of [0, 0.1, 0.2, 0.3, 0.4, 0.5], L1 and L2 in the set of [0, 10^{-1} , 10^{-2} , 10^{-3} , 10^{-4} , 10^{-5} , 10^{-6} , 10^{-7}]) and the CNN max-pooling in the set of [2, 4, 6, 8, 16] for each layer. The best set of parameters are the ones which give the lowest SELD score on the three cross-validation splits of the dataset. After finding the best network architecture and configuration, we tune the input audio feature parameter M by varying it in the set of [512, 1024, 2048]. Simultaneously the sequence length is also changed with respect to M such that the input audio length is kept constant (1.49 s obtained from the first round of tuning). We perform fine-tuning of model parameters for different M and sequence length values, this time only the number of CNN, RNN and FC nodes are tuned in a small range (neighboring nodes in the set of [16, 32, 64, 128, 256, 512]) to identify the optimum parameters. Similar fine-tuning is repeated for other datasets.

2) *Selecting SELDnet Output Format*: The output format for polyphonic SED in the literature has become standardized to estimating the temporal activity of each sound class using frame-wise binary numbers [31]–[34]. On the other hand, the output formats for DOA estimation are still being experimented with as seen in Table I. Among the DOA estimation methods using regression mode, there are two possible output formats, predicting azimuth and elevation, and predicting x, y, z coordinates of the DOA on the unit sphere. In order to identify the best output format among these two, we evaluate the SELDnet for both. During this evaluation, only the output weight parameter of the model is fine-tuned in the set of [1, 5, 50, 500]. Additionally, for a regression-based model, the default output i.e. the DOA target when the event is not active should be chosen carefully. In this study, we chose the default DOA output to be 180° in azimuth and 60° in elevation (the datasets do not contain sound events for these DOA values), and $x = 0$, $y = 0$ and $z = 0$ for default Cartesian outputs. The chosen default Cartesian coordinates are equidistant from all the possible DOA values. On the other hand, there are no such equidistant azimuth and elevation values. Hence we chose the default values (180° , 60°) to be in a similar range as the true DOA values.

3) *Continuous DOA Estimation and Performance on Unseen DOA Values*: Theoretically, the advantage of using a regression-based DOA estimator over a classification-based one is that the network is not limited to a set of DOA angles, but it can operate as a high-resolution continuous DOA estimator. To study this,

we train the SELDnet on ANSYN subsets whose sound events are placed on an angular grid of 10° resolution along azimuth and elevation, and test the model on a dataset where the angular grid is shifted by 5° along azimuth and elevation while keeping the temporal location unchanged. This shift makes the DOA values of the testing split unseen, and correctly predicting the DOAs will prove that the regression model can estimate the DOAs in a continuous space. Additionally, it also proves the robustness of the SELDnet to predict unseen DOA values.

4) *Performance on Mismatched Reverberant Dataset*: Parametric DOA estimation methods are known to be sensitive to reverberation [48]. In this regard, we first evaluate the performance of SELDnet on the simulated (RESYN), and real-life (REAL, REALBIG, and REALBIGAMB) reverberant datasets and further compare the results with the parametric baseline MUSIC.

DNN based methods are known to fail when the training and testing splits come from different domains. For example, the performance of a DNN trained on anechoic dataset would be poor on a reverberant testing dataset. This performance can only be improved by training the DNN on a similar reverberant dataset as the testing dataset. On the other hand, it is impractical to train such a DNN for every existing room-dimension, its surface material distribution, and the reverberation times associated with it. In this regard, it would be ideal if the proposed method is robust to a moderate mismatch in reverberant conditions so that a single model can be used for a range of comparable room configurations. Motivated by this, we study the sensitivity of SELDnet on moderately mismatched reverberant data. Specifically, we train the SELDnet with RESYN room 1 dataset and test it on RESYN room 2 and 3 datasets that are mismatched in terms of volume and reverberation times as described in Section III-A2.

5) *Performance on the Size of the Dataset*: We study the performance of SELDnet on two datasets, REAL, and REALBIG that are similar in content, but different in size.

6) *Performance with Ambiance at Different SNR*: The performance of SELDnet with respect to different SNRs (0, 10 and 20 dB) of the sound event is studied on the REALBIGAMB dataset.

7) *Generic to Array Structure*: SELDnet is a generic method that learns to localize and recognize sound events from any array structure. This additionally implies that the SELDnet will continue to work in the desired manner if the configuration of the array such as individual microphone response, microphone spacing and the number of microphones remains the same between the training and testing set. If the array configuration changes between the training and testing set, then the SELDnet will have to be retrained for the new array configuration.

In order to prove that the SELDnet is applicable to any array configuration and not just dependent on the Ambisonics format, SELDnet is evaluated on a circular array. In comparison to the Ambisonic format, the chosen circular array has a different number of microphones, each placed on a single plane, and with an omnidirectional response. Further, we compare the SELDnet performance with dataset compatible baselines such as SEDnet, MSEDnet, HIRnet, and AZInet. Since the HIRnet and AZInet baselines methods are proposed for estimating azimuth only, we

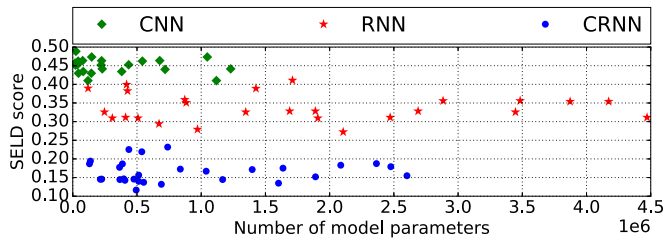


Fig. 3. SELD score for ANSYN *O2* dataset for different CNN, RNN and CRNN architecture configurations.

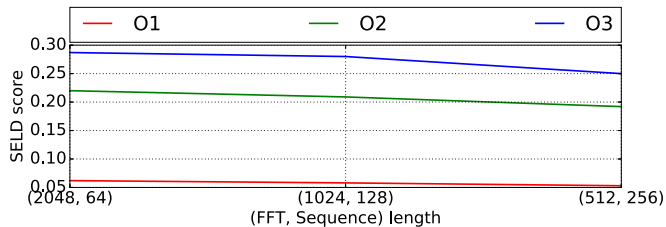


Fig. 4. SELD score for ANSYN datasets for different combinations of FFT length and input sequence length in frames.

compare the results with the SELDnet-azi version. Additionally, we also report the performance of using SELDnet with DOA estimation in x, y, z axis on CANSYN and CRESYN datasets.

In general, for all our experiments the only difference between the training and testing splits is the mutually exclusive set of sound examples. Apart from experiment III-D3 the training and testing splits contains the same set of spatial locations i.e., azimuth and elevation angles at 10° resolution amounting to 468 spatial locations ($= 36$ azimuth angles $\times 13$ elevation angles). But the distance of the sound event at each of this 468 spatial locations is an added variable. For example, in the anechoic case, a sound event can be placed anywhere between 1-10 m at 0.5 m resolution. This variable amounts to 8892 spatial locations ($= 468 \times 19$ distance positions) that are being coarsely grouped to 468 locations. This complexity is stretched further in experiment III-D3 where the testing split sound event examples and their spatial locations both are different from the training split.

IV. RESULTS AND DISCUSSION

1) *SELDnet Architecture and Model Parameter Tuning*: The SELD scores obtained with hyper-parameter tuning of different CNN, RNN, and CRNN configurations as explained in Section III-D1 are visualized with respect to the number of model parameters in Fig. 3. CNN in this figure refers to a SELDnet architecture which had no RNN layers but just CNN and FC layers. Similarly, RNN refers to SELDnet without CNN layers, while CRNN refers to SELDnet with CNN, RNN and FC layers. This experiment was carried out on ANSYN *O2* dataset. The CRNN architecture was seen to perform the best followed by the RNN architecture.

The optimum model parameters across the ANSYN subsets after hyper-parameter tuning the CRNN architecture was found to have three layers of CNN with 64 nodes each (P in Fig. 1(a)),

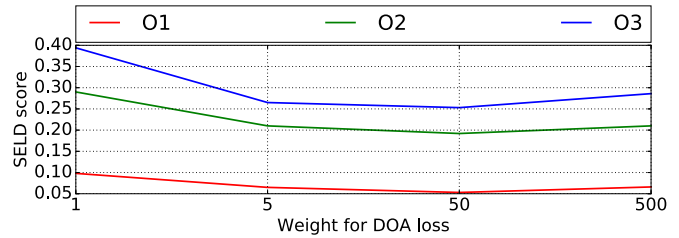


Fig. 5. SELD score for ANSYN datasets with respect to different weights for DOA output.

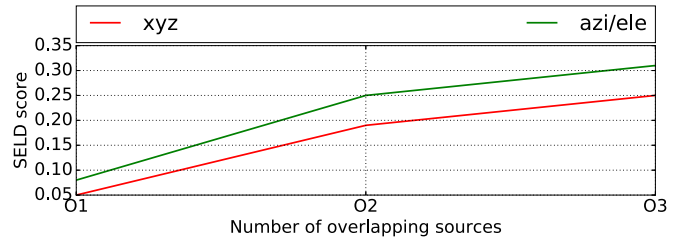


Fig. 6. SELD score for ANSYN datasets with respect to DOA output formats.

followed by two layers of GRU with 128 nodes each (Q in Fig. 1(a)), and one FC layer with 128 nodes (R in Fig. 1(a)). The max-pooling over frequency after each of the three CNN layers (MP_i in Fig. 1(a)) was $(8, 8, 2)$. This configuration had about 513,000 parameters.

Further, the SELDnet was seen to perform best with no regularization (dropout, or L1 or L2 regularization of weights). A frame length of $M = 512$ and sequence length of 256 frames was seen to give the best results across ANSYN subsets (Fig. 4). Furthermore, on tuning the sequence length with frame length fixed ($M = 512$), the best scores were obtained using 512 frames (2.97 s). Sequences longer than this could not be studied due to hardware restrictions. For the output weights, DOA output weighted 50 times more than SED output was seen to give the best results across subsets (Fig. 5).

On fine-tuning the SELDnet parameters obtained with ANSYN dataset for RESYN subsets, the only parameter that helped improve the performance was using a sequence length of 256 instead of 512, leaving the total number of network parameters unchanged at 513,000. Similar configuration gave the best results for CANSYN and CRESYN datasets.

Model parameters identical to ANSYN dataset were observed to perform the best on the REAL subsets. The same parameters were also used for the study of REALBIG and REALBIGAMB subsets.

2) *Selecting SELDnet Output Format*: In the output data formats study, it was observed that using the Cartesian x, y, z format in place of azimuth/elevation angle was truly helping the network learn better across datasets as seen in Fig. 6. This suggests that the discontinuity at the angle wrap-around boundary actually reduces the performance of DOA estimation and hence the SELD score.

3) *Continuous DOA Estimation and Performance on Unseen DOA Values*: The input and outputs of SELDnet trained on ANSYN *O1* and *O2* subsets for a respective 1000 frame test

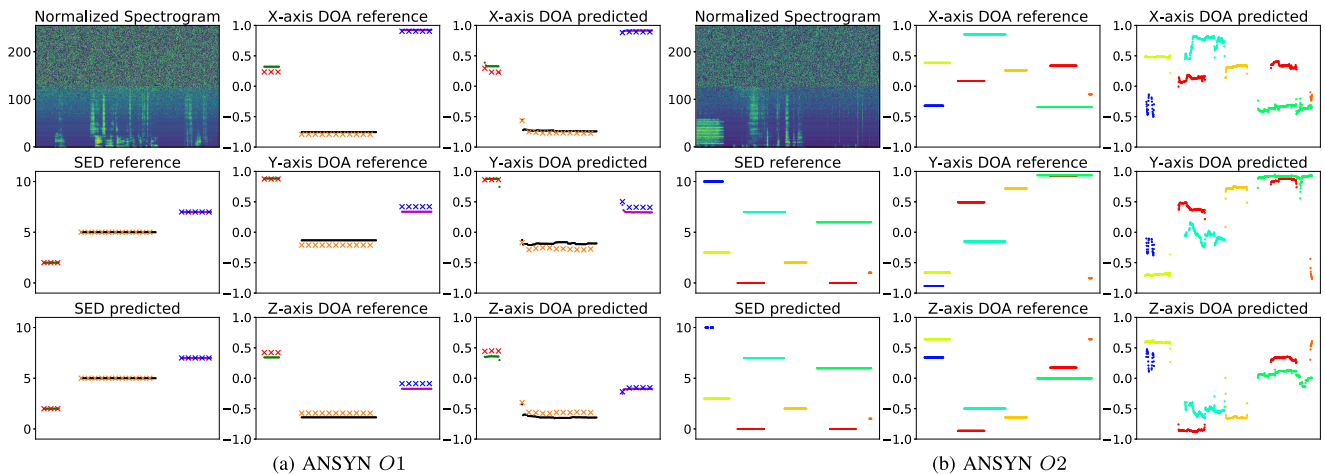


Fig. 7. SELDnet input and outputs visualized for ANSYN $O1$ and $O2$ datasets. The horizontal-axis of all sub-plots for a given dataset represents the same time frames, the vertical-axis for spectrogram sub-plot represents the frequency bins, vertical-axis for SED reference and prediction sub-plots represents the unique sound event class identifier, and for the DOA reference and prediction sub-plots, it represents the distance from the origin along the respective axes. The bold lines visualize both the reference labels and predictions of DOA and SED for ANSYN $O1$ and $O2$ datasets, while the \times markers in Fig. 7(a) visualize the results for testing split with unseen DOA values (shifted by 5° along azimuth and elevation).

sequence are visualized in Fig. 7. The Figure represents each sound event class and its associated DOA outputs with a unique color. In the case of ANSYN $O1$, we see that the network predictions of SED and DOA are almost perfect. In the case of unseen DOA values (\times markers), the network predictions continue to be accurate. This shows that the regression mode output format helps the network learn continuous DOA values, and further that the network is robust to unseen DOA values. In case of ANSYN $O2$, the SED predictions are accurate, while the DOA estimates, in general, are seen to vary around the respective mean reference value. In this work, the DOA and SED labels for a single sound event instance are considered to be constant for the entire duration even though the instance has inherent magnitude variations and silences within. From Fig. 7(b) it seems that these variations and silences are leading to fluctuating DOA estimates, while the SED predictions are unaffected. In general, we see that the proposed method successfully recognizes, localizes in time and space, and tracks multiple overlapping sound events simultaneously.

Table IV presents the evaluation metric scores for the SELDnet and the baseline methods with ANSYN and RESYN datasets. In the SED metrics for the ANSYN datasets, the SELDnet performed better than the best baseline MSEDnet for $O1$ subset while MSEDnet performed slightly better for $O2$ and $O3$ subsets. With regard to DOA metrics, the SELDnet is significantly better than the baseline DOAnet in terms of frame recall. This improvement in frame recall is a direct result of using SED output as a confidence measure for estimating DOA, thereby extending state-of-the-art SED performance to SELD. Although the frame recall of DOAnet is poor, its DOA error for $O1$ and $O2$ subsets is observed to be lower than SELDnet. The DOA error of the parametric baseline MUSIC with the knowledge of the number of sources is seen to be the best among the evaluated methods for $O2$ and $O3$ subsets.

4) *Performance on Mismatched Reverberant Dataset:* From Table IV results on RESYN room 1 subsets, we see that the performance of parametric method MUSIC is poor in comparison

to SELDnet in reverberant conditions. The SELDnet is seen to perform significantly better than the baseline DOAnet in terms of frame recall, although the DOAnet has lower DOA error for $O1$ and $O2$ subsets. The SED metrics of SELDnet are comparable if not better than the best baseline performance of MSEDnet. Further, on training the SELDnet on room 1 dataset and testing on moderately mismatched reverberant room 2 and 3 datasets the SED and DOA metric trends remain similar to the results of room 1 testing split. That is, the SELDnet has higher frame recall, the DOAnet has better DOA error, the MUSIC performs poorly, and the SED metrics of SELDnet are comparable to MSEDnet. These results prove that the SELDnet is robust to reverberation in comparison to baseline methods and performs seamlessly on moderately mismatched room configurations.

Fig. 8 visualizes the confusion matrices for the estimated number of sound event classes per frame by SELDnet. For example in Fig. 8(c) the SELDnet correctly estimated the number of sources to be two in 76% (true positive percentage) of the frames which had two sources in the reference. In context, the frame recall value used as a metric to evaluate DOA estimation represents this confusion matrix in one number. From the confusion matrices, we observe that the percentage of true positives drops with higher number of sources, and this drop is even more significant in the reverberant scenario. But, in comparison to the frame recall metric of the baseline DOAnet in Table IV, the SELDnet frame recall is significantly better for higher number of overlapping sound events, especially in the reverberant conditions.

5) *Performance on the Size of the Dataset:* The overall performance of SELDnet on REAL dataset (Table V) reduced in comparison to ANSYN and RESYN datasets. The baseline MSEDnet is seen to perform better than SELDnet in terms of SED metrics. Similar performance drop on real-life datasets has also been reported on SED datasets in other studies [37]. With regard to DOA metrics, the frame recall of SELDnet continues to be significantly better than DOAnet, while the DOA error of DOAnet is lower than SELDnet. The performance of MUSIC is

TABLE IV
 SED AND DOA ESTIMATION METRICS FOR ANSYN AND RESYN DATASETS. THE RESULTS FOR THE RESYN ROOM 2 AND 3 TESTING SPLITS WERE OBTAINED FROM CLASSIFIERS TRAINED ON RESYN ROOM 1 TRAINING SET. BEST SCORES FOR SUBSETS IN BOLD

SED metrics	Overlap	ANSYN			RESYN Room 1			RESYN Room 2			RESYN Room 3		
		1	2	3	1	2	3	1	2	3	1	2	3
SELDnet	ER	0.04	0.16	0.19	0.10	0.29	0.32	0.11	0.33	0.35	0.13	0.32	0.34
	F	97.7	89.0	85.6	92.5	79.6	76.5	91.6	79.5	75.8	89.8	79.1	75.5
MSEDnet [39]	ER	0.10	0.13	0.17	0.17	0.28	0.29	0.19	0.30	0.26	0.18	0.29	0.30
	F	94.4	90.1	87.2	89.1	79.1	75.6	88.3	78.2	74.2	86.5	80.5	76.1
SEDnet [39]	ER	0.14	0.16	0.18	0.18	0.28	0.30	0.19	0.32	0.28	0.21	0.32	0.33
	F	91.9	89.1	86.7	88.2	76.9	74.1	87.6	76.4	73.2	85.1	78.2	75.6
DOA metrics													
SELDnet	DOA error	3.4	13.8	17.3	9.2	20.2	26.0	11.5	26.0	33.1	12.1	25.4	31.9
	Frame recall	99.4	85.6	70.2	95.8	74.9	56.4	96.2	78.9	61.2	95.9	78.2	60.7
DOAnet [25]	DOA error	0.6	8.0	18.3	6.3	11.5	38.4	3.4	6.9	-	4.6	10.9	-
	Frame recall	95.4	42.7	1.8	59.3	15.8	1.2	46.2	14.3	-	49.7	14.1	-
MUSIC	DOA error	4.1	7.2	15.8	40.2	47.1	50.5	45.7	58.1	74.0	48.3	60.6	75.6

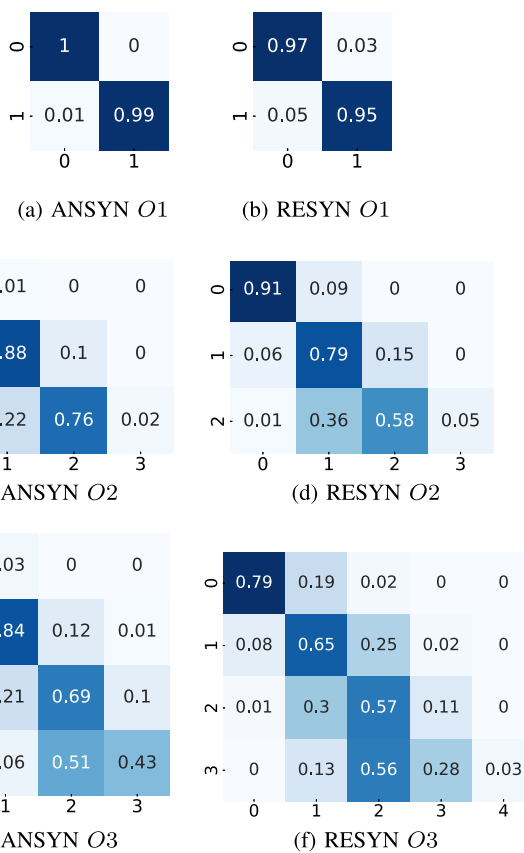


Fig. 8. Confusion matrix for the number of sound event classes estimated to be active per frame by the SELDnet for ANSYN and RESYN datasets. The horizontal axis represents the SELDnet estimate, and the vertical axis represents the reference.

seen to be poor in comparison to both DOAnet and SELDnet. With the larger REALBIG dataset the SELDnet performance was seen to improve. A similar study was done with larger ANSYN and RESYN datasets, where the results were comparable with that of smaller datasets. This shows that the datasets with real-life IR are more complicated than synthetic IR datasets, and having larger real-life datasets helps the network learn better.

6) *Performance with Ambiance at Different SNR*: In presence of ambiance, SELDnet was seen to be robust for 10 and

20 dB SNR REALBIGAMB datasets (Table V). In comparison to the SED metrics of REALBIG dataset with no ambiance, the SELDnet performance on $O1$ subsets of 10 dB and 20 dB ambiance is comparable, while a small drop in performance was observed with the respective $O2$ and $O3$ subsets. Whereas, the performance was seen to drop considerably for the 0 dB SNR dataset. With respect to DOA error, the SELDnet performed better than MUSIC but poorer than DOAnet across datasets, on the other hand, SELDnet gave significantly higher frame recall than DOAnet. From the insight of SELDnet performance on REAL dataset (Section IV), the more complex the acoustic scene the larger the dataset size required to learn better. Considering that the SELDnet is jointly estimating the DOA along with SED in a challenging acoustic scene with ambiance the SELDnet performance can potentially improve with larger datasets.

7) *Generic to Array Structure*: The results on circular array datasets are presented in Table VI. With respect to SED metrics, the SELDnet-azi performance is seen to be better than the best baseline MSEDnet for all subsets of CRESYN dataset, while MSEDnet is seen to perform better for $O2$ and $O3$ subsets of CANSYN dataset. Similarly, in the case of DOA metrics, the SELDnet-azi has better frame recall than the best baseline method AZInet across datasets (except for CANSYN $O1$). Whereas, AZInet has lower DOA error than SELDnet-azi. Between SELDnet and SELDnet-azi, even though the frame recall is in the same order the DOA error of SELDnet-azi are lower than SELDnet. This shows that estimating DOA in 3D (x, y, z) is challenging using a circular array. Overall, the SELDnet is shown to perform consistently across different array structures (Ambisonic and circular array), with good results in comparison to baselines.

The usage of SED output as a confidence measure for estimating the number of DOAs in the proposed SELDnet is shown to improve the frame recall significantly and consistently across the evaluated datasets. On the other hand, the DOA error obtained with SELDnet is consistently higher than the classification based baseline DOA estimation methods [18], [25]. We believe that this might be a result of the regression-based DOA estimation approach in SELDnet not having completely learned the full mapping between input feature and the continuous DOA space. The investigation of which is

TABLE V
SED AND DOA ESTIMATION METRICS FOR REAL, REALBIG AND REALBIGAMB DATASETS. BEST SCORES FOR SUBSETS IN BOLD

	Overlap	REAL			REALBIG			REALBIGAMB 20dB			REALBIGAMB 10dB			REALBIGAMB 0dB		
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
SED metrics																
SELDnet	ER	0.40	0.49	0.53	0.37	0.42	0.50	0.34	0.46	0.52	0.37	0.49	0.52	0.46	0.58	0.59
	F	60.3	53.1	51.1	65.4	61.5	56.5	65.6	58.5	55.0	66.3	55.4	53.3	57.9	48.6	49.0
MSEDnet [39]	ER	0.35	0.38	0.41	0.34	0.39	0.38	0.35	0.40	0.41	0.38	0.43	0.42	0.48	0.56	0.54
	F	66.2	61.6	59.5	67.3	61.8	61.9	66.0	61.6	60.1	63.2	58.7	59.3	54.5	49.3	51.3
SEDnet [39]	ER	0.38	0.42	0.43	0.38	0.43	0.44	0.39	0.42	0.43	0.41	0.44	0.46	0.51	0.61	0.57
	F	64.6	61.5	57.2	68.0	62.4	62.4	65.7	60.1	59.2	62.7	56.3	56.9	52.6	46.0	50.4
DOA metrics																
SELDnet	DOA error	26.6	33.7	36.1	23.1	31.3	34.9	25.4	32.5	36.1	27.2	32.5	36.1	30.7	33.7	36.7
	Frame recall	64.9	41.5	24.6	68.0	45.2	28.3	69.1	42.8	25.8	66.9	40.0	27.3	62.5	35.2	23.4
DOAnet [25]	DOA error	6.3	20.1	25.8	7.5	17.8	22.9	6.3	18.9	25.78	8.0	20.1	24.1	14.3	24.1	27.5
	Frame recall	46.5	11.5	2.9	44.1	12.5	3.1	34.7	11.6	3.2	42.1	13.5	3.3	30.1	10.5	2.8
MUSIC	DOA error	36.3	49.5	54.3	35.8	49.6	53.8	54.5	56.1	61.3	51.6	54.5	62.6	41.9	47.5	62.3

TABLE VI
SED AND DOA ESTIMATION METRICS FOR CANSYN AND CRESYN DATASETS. BEST SCORES FOR SUBSETS IN BOLD

	Overlap	CANSYN			CRESYN		
		1	2	3	1	2	3
SED metrics							
SELDnet	ER	0.11	0.18	0.19	0.13	0.22	0.30
	F score	93.0	86.6	85.3	90.4	82.2	78.0
SELDnet-azi	ER	0.08	0.19	0.24	0.06	0.18	0.20
	F score	94.7	87.5	83.8	96.3	87.9	85.6
MSEDnet [39]	ER	0.09	0.18	0.16	0.12	0.22	0.26
	F score	94.6	89.0	86.7	92.7	83.7	80.7
SEDnet [39]	ER	0.15	0.21	0.20	0.18	0.26	0.25
	F score	91.4	87.3	84.7	90.5	84.3	82.8
HIRnet [20]	ER	0.41	0.45	0.62	0.43	0.46	0.50
	F score	60.0	54.9	58.8	59.3	60.2	58.6
DOA metrics							
SELDnet	DOA error	29.5	31.3	34.3	28.4	33.7	41.0
	Frame recall	97.9	78.8	67.0	96.4	75.7	60.7
SELDnet-azi	DOA error	7.5	14.4	19.6	5.2	13.2	18.4
	Frame recall	98.0	82.1	66.2	98.5	82.3	70.6
HIRnet [20]	DOA error	5.2	16.3	33.0	7.4	18.6	43.3
	Frame recall	60.2	35.9	18.4	56.9	20.5	10.7
AZInet [18]	DOA error	1.2	4.0	7.4	2.3	6.9	9.7
	Frame recall	99.4	80.5	60.5	97.3	65.2	44.8

planned for future work. In general, a classification only or a classification-regression based SELD approach can be chosen based on the required frame recall, DOA error, resolution of DOA labels, training split size, and robustness to unseen DOA values and reverberation.

V. CONCLUSION

In this paper, we proposed a convolutional recurrent neural network (SELDnet) to simultaneously recognize, localize and track sound events with respect to time. The localization is done by estimating the direction of arrival (DOA) on a unit sphere around the microphone using 3D Cartesian coordinates. We tie each sound event output class in the SELDnet to three regressors to estimate the respective Cartesian coordinates. We show that using regression helps estimating DOA in a continuous space, and also estimating unseen DOA values accurately. On the other hand, estimating a single DOA for each sound event class does not allow recognizing multiple instances of the same class overlapping. We plan to tackle this problem in our future work.

The usage of SED output as a confidence measure to estimate DOA was seen to extend the state-of-the-art SED performance

to SELD resulting in a higher recall of DOAs. With respect to the estimated DOA error, although the classification based baseline methods had poor recall they had lower DOA error in comparison to the proposed regression based DOA estimation. The proposed SELDnet uses phase and magnitude spectrogram as the input feature. The usage of such non-method-specific feature makes the method generic and easily extendable to different array structures. We prove this by evaluating on datasets of Ambisonic and circular array format. The proposed SELDnet is shown to be robust to reverberation, low SNR scenarios and unseen rooms with comparable room-sizes. Finally, the overall performance on dataset synthesized using real-life impulse response (IR) was seen to drop in comparison to artificial IR dataset, suggesting the need for larger real-life training datasets and more powerful classifiers in future.

ACKNOWLEDGMENT

The authors would like to thank CSC-IT Center for Science, Finland, for computational resources.

REFERENCES

- [1] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 405–409.
- [2] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2016, pp. 603–609.
- [3] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *J. Robot. Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.
- [4] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. Int. Conf. Robot. Autom.*, 2018, pp. 74–79.
- [5] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, 2016, Art. no. 52.
- [6] C. Grobler, C. Kruger, B. Silva, and G. Hancke, "Sound based localization and identification in industrial environments," in *Proc. 43rd Annu. Conf. IEEE Ind. Electron. Soc.*, 2017, pp. 6119–6124.
- [7] P. W. Wessels, J. V. Sande, and F. V. der Eerden, "Detection and localization of impulsive sound events for environmental noise assessment," *J. Acoust. Soc. Amer.*, vol. 141, no. 5, 2017, Art. no. 3886.
- [8] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 279–288, Jan. 2016.
- [9] C. Busso *et al.*, "Smart room: Participant and speaker localization and identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2005, pp. ii–1117.

- [10] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 1997, pp. 187–190.
- [11] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Hoboken, NJ, USA; Wiley, 2009.
- [12] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1120–1124, Sep. 2014.
- [13] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *Proc. Euro. Signal Process. Conf.*, 2011, pp. 1317–1321.
- [14] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.
- [15] T. A. Marques *et al.*, "Estimating animal population density using passive acoustics," *Biol. Rev. Cambridge Philosophical Soc.*, vol. 88, no. 2, pp. 287–309, 2012.
- [16] B. J. Furnas and R. L. Callas, "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," *J. Wildlife Manage.*, vol. 79, no. 2, pp. 325–337, 2014.
- [17] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017.
- [18] S. Chakrabarty and E. A. P. Habets, "Multi-speaker localization using convolutional neural network trained with noise," in *Proc. Neural Inf. Process. Syst.*, 2017.
- [19] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2814–2818.
- [20] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Proc. Audio Eng. Soc. Conv. 138*, 2015, paper 9294.
- [21] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," *Int. J. Appl. Eng. Res.*, vol. 12, no. 22, pp. 12384–12389, 2017.
- [22] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2386–2390.
- [23] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2016, pp. 1–6.
- [24] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Trans. Ind. Electron.*, vol. 29, no. 1, pp. 6403–6413, Aug. 2017.
- [25] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. Euro. Signal Process. Conf.*, 2018.
- [26] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in *Proc. Deutsche Jahrestagung für Akustik*, 2015.
- [27] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *Proc. Euro. Signal Process. Conf.*, 2010, pp. 1267–1271.
- [28] E. Çakır, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi-label deep neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2015, pp. 1–7.
- [29] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 6440–6444.
- [30] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Proc. Detection Classification Acoust. Scenes Events*, 2016.
- [31] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. L. Roux, and K. Takeda, "Duration-controlled LSTM for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 11, pp. 2059–2070, Nov. 2017.
- [32] M. Zöhrer and F. Pernkopf, "Virtual adversarial training and data augmentation for acoustic event detection with gated recurrent neural networks," in *Proc. INTERSPEECH*, 2017, pp. 493–497.
- [33] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 559–563.
- [34] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," in *Proc. INTERSPEECH*, 2016.
- [35] S. Adavanne, A. Politis, and T. Virtanen, "Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2018.
- [36] H. Lim, J. Park, K. Lee, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," in *Proc. Detection Classification Acoust. Scenes Events*, 2017.
- [37] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.
- [38] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," in *Proc. Detection Classification Acoust. Scenes Events*, 2017.
- [39] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 771–775.
- [40] I.-Y. Jeong, S. Lee, Y. Han, and K. Lee, "Audio event detection using multiple-input convolutional neural network," in *Proc. Detection Classification Acoust. Scenes Events*, 2017.
- [41] J. Zhou, "Sound event detection in multichannel audio LSTM network," in *Proc. Detection Classification Acoust. Scenes Events*, 2017.
- [42] R. Lu and Z. Duan, "Bidirectional GRU for sound event detection," in *Proc. Detection Classification Acoust. Scenes Events*, 2017.
- [43] A. Temko, C. Nadeu, and J.-I. Biel, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07," in *Multimodal Technologies for Perception of Humans*. New York, NY, USA: Springer, 2008.
- [44] Y. Huang, J. Benesty, G. Elko, and R. Mersereau, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.
- [45] M. S. Brandstein and H. F. Silverman, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, pp. 80–82.
- [46] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.
- [47] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.
- [48] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. New York, NY, USA: Springer, 2001.
- [49] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*. Springer, 2007, vol. 348.
- [50] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, no. 3, pp. 216–228, 2007.
- [51] J. Traa and P. Smaragdis, "Multiple speaker tracking with the Factorial Von Mises-Fisher filter," in *Proc. IEEE Int. Workshop Mach. Learn. for Signal Process.*, 2014, pp. 1–6.
- [52] R. Chakraborty and C. Nadeu, "Sound-model-based acoustic source localization using distributed microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 619–623.
- [53] K. Lopatka, J. Kotus, and A. Czyzewski, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools Appl. J.*, vol. 75, no. 17, pp. 10407–10439, 2016.
- [54] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
- [55] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [56] F. Chollet, "Keras v2.0.8," 2015. Accessed: May 7, 2018. [Online]. Available: <https://github.com/fchollet/keras>
- [57] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Accessed: May 7, 2018. [Online]. Available: <https://www.tensorflow.org/>
- [58] E. Benetos, M. Lagrange, and G. Lafay, "Sound event detection in synthetic audio," 2016. Accessed: May 7, 2018. [Online]. Available: https://archive.org/details/dcace2016_task2_train_dev

- [59] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
- [60] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 325–328.
- [61] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.
- [62] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 1041–1044.
- [63] A. Politis, "Microphone array processing for parametric spatial audio techniques," Ph.D. dissertation, Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland, 2016.
- [64] A. Mesaros *et al.*, "DCASE 2017 challenge setup: tasks, datasets and baseline system," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2017.
- [65] B. Ottersten, M. Viberg, P. Stoica, and A. Nehorai, "Exact and large sample maximum likelihood techniques for parameter estimation and detection in array processing," in *Radar Array Processing (Springer Series in Information Sciences)*. Berlin, Germany: Springer, 1993.
- [66] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *J. Acoust. Soc. Amer.*, vol. 132, no. 1, pp. 261–270, 2012.
- [67] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, pp. 1–17, 2016.
- [68] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. Plumbley, and D. Ellis, Eds. New York, NY, USA: Springer, 2018, ch. 6.



Sharath Adavanne received the M.Sc. degree in information technology from Tampere University of Technology (TUT), Tampere, Finland, in 2011. Since 2016, he has been working toward the Ph.D. degree with the Laboratory of Signal Processing, TUT. From 2011 to 2016, he worked in the industry solving problems related to music information retrieval, speech recognition, audio fingerprinting, and general audio content analysis. His current research interest include the application of machine learning based methods for real-life auditory scene analysis.



Archontis Politis received the M.Eng. degree in civil engineering Aristotle University, Thessaloniki, Greece, and the M.Sc. degree in sound and vibration studies from the Institute of Sound and Vibration Studies (ISVR), Southampton, U.K., in 2006 and 2008, respectively, and the Doctor of Science degree on the topic of parametric spatial sound recording and reproduction from Aalto University, Espoo, Finland, in 2016. From 2008 to 2010, he worked as a graduate acoustic consultant with Arup Acoustics, U.K., and as a Researcher in a joint collaboration between Arup Acoustics and the Glasgow School of Arts, on architectural auralization using spatial sound techniques. He was an intern with the Audio and Acoustics Research Group of Microsoft Research, during summer of 2015. He is currently a Postdoctoral Researcher with Aalto University. His research interests include spatial audio technologies, virtual acoustics, array signal processing, and acoustic scene analysis.



Joonas Nikunen received the M.Sc. degree in signal processing and communications engineering and the Ph.D. degree in Signal Processing from Tampere University of Technology (TUT), Tampere, Finland, in 2010 and 2015, respectively. He is currently a Postdoctoral Researcher with TUT focusing on sound source separation with applications on spatial audio analysis, modification, and synthesis. His other research interests include microphone array signal processing, 3-D/360 audio in general and machine and deep learning for source separation.



Tuomas Virtanen (SM'15) received the M.Sc. and Doctor of Science degrees in information technology from Tampere University of Technology (TUT), Tampere, Finland, in 2001 and 2006, respectively. He is currently a Professor with the Laboratory of Signal Processing, TUT, where he is leading the Audio Research Group. He has also been working as a Research Associate with Cambridge University Engineering Department, Cambridge, U.K. He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques, and their application to noise-robust speech recognition and music content analysis. Recently, he has done significant contributions to sound event detection in everyday environments. In addition to the above-mentioned topics, his research interests include content analysis of audio signals in general and machine learning. He has authored more than 150 scientific publications on the above topics, which have been cited more than 6000 times. He was a recipient of the IEEE Signal Processing Society 2012 Best Paper Award for his article "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria" as well as three other best paper awards. He is a member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society, an Associate Editor for the IEEE/ACM TRANSACTION ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and a recipient of the ERC 2014 Starting Grant.