Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network

Nguyen, Thi Ngoc Tho; Gan, Woon-Seng; Ranjan, Rishabh; Jones, Douglas L.

2020

Nguyen, T. N. T., Gan, W.-S., Ranjan R., & Jones, D. L. (2020). Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2626-2637. doi: 10.1109/TASLP.2020.3019646

<https://hdl.handle.net/10356/144539>

<https://doi.org/10.1109/TASLP.2020.3019646>

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: <https://doi.org/10.1109/TASLP.2020.3019646>

Downloaded on 26 Apr 2023 19:05:04 SGT

Robust Source Counting and DOA Estimation Using Spatial Pseudo-Spectrum and Convolutional Neural Network

Thi Ngoc Tho Nguyen*, *Student Member, IEEE*, Woon-Seng Gan, *Senior Member, IEEE*, Rishabh Ranjan, and Douglas L. Jones, *Fellow, IEEE*

Abstract—Many signal processing-based methods for sound source direction-of-arrival estimation produce a spatial pseudo-spectrum of which the local maxima strongly indicate the source directions. Due to different levels of noise, reverberation and different number of overlapping sources, the spatial pseudo-spectra are noisy even after smoothing. In addition, the number of sources is often unknown. As a result, selecting the peaks from these spectra is susceptible to error. Convolutional neural network has been successfully applied to many image processing problems in general and direction-of-arrival estimation in particular. In addition, deep learning-based methods for direction-of-arrival estimation show good generalization to different environments. We propose to use a 2D convolutional neural network with multi-task learning to robustly estimate the number of sources and the directions-of-arrival from short-time spatial pseudo-spectra, which have useful directional information from audio input signals. This approach reduces the tendency of the neural network to learn unwanted association between sound classes and directional information, and helps the network generalize to unseen sound classes. The simulation and experimental results show that the proposed methods outperform other directional-of-arrival estimation methods in different levels of noise and reverberation, and different number of sources.

Index Terms—direction-of-arrival estimation, convolutional neural network, spatial pseudo-spectrum, multi-task learning, multiple sound sources.

I. INTRODUCTION

DIRECTION-OF-ARRIVAL (DOA) estimation of acoustic sources has many applications in wildlife monitoring, surveillance [1], robotics [2], tele-conferencing [3] and hearing aids [4]. Some of the main challenges for DOA estimation in real-life scenes are multiple sources, reverberation, and noise.

This paper focuses on DOA estimation using a small-aperture array in indoor environments under far-field assumption. There are two main classes of algorithms for this focus. The first class is the signal processing (SP) based method, where the geometry of the microphone array is required, and there are some potential mismatches between the theoretical signal model and real-life test data. The second class is the deep learning (DL) based method, where a training dataset with ground truth is required, and there are some potential mismatches between the training set and real-life test sets.

T. N. T. Nguyen, W. S. Gan, and R. Ranjan are with the Department of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore, e-mail: nguyenth003@e.ntu.edu.sg.

D. L. Jones is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL 61801, USA.

A. Signal processing-based methods

The signal processing methods for DOA estimation for small-aperture microphone arrays can be loosely grouped into 4 categories: (i) time-difference-of-arrival (TDOA), (ii) sub-space, (iii) beamforming, and (iv) histogram analysis methods. Sub-space methods such as multiple signal classification (MUSIC) [5] and the estimation of signal parameters via rotational invariance technique (ESPRIT) [6] rely on the spatio-spectral correlation matrix. The number of active sources is often required to estimate a noise sub-space. A beamformer such as delay and sum beamformer [7] and minimum variance distortionless response (MVDR) beamformer [8] algorithmically steers the array toward a particular direction by filtering, weighting and summing the input signals. The histogram analysis methods often compute the DOAs for time-frequency (TF) bins on the spectrograms that contain energy mainly from one source [9], [10]. These estimated DOAs are then populated into a histogram of the angles of arrival.

To reduce the effect of reverberation in DOA estimation, many researches apply the principle of precedence effect [1], [9]. To tackle the multi-source cases, many researches exploit the non-stationarity and sparseness of audio signals to find the single-source TF bins on the spectrograms to reliably estimate DOAs [12], [13].

Majority of the SP-based DOA estimation algorithms produce some form of spatial spectrum of which the local maxima often correspond to the source directions. For example, MUSIC algorithm produces a spatial pseudo-spectrum, beamformers produce power spectra, and histogram methods produces DOA histograms. We refer to these spectra as spatial pseudo-spectra (SPS). If the number of sound sources is unknown, the DOAs are often selected from peaks of the SPS that are above a certain threshold [9], [11]. Under different conditions of noise, reverberation, and different number of sources, the SPS are noisy. As a result, peak detection algorithms that rely on a pre-determined threshold for all of the cases are prone to errors, and severely affect the accuracy of DOA estimation algorithms. Our research aims to mitigate this problem by adding a learning component to SP-based DOA estimation algorithms.

B. Deep learning-based methods

Recently, deep learning (DL) has been successfully applied to the DOA estimation tasks [14]–[18]. The learning-based

DOA models show good generalization to different noise and reverberation levels, but the angular error is still large for multi-source cases. The number of overlapping sources are often limited to 2 or 3 sources. To reduce the effect of reverberation and noise, multi-condition training is often employed, where different room environments and different noise levels are simulated for training. The DL-based DOA models are often trained with spatial features such as generalized cross-correlation with phase transform (GCC-PHAT) [14], phase and magnitude spectrum [16], eigenvectors of the spatial covariance matrix [15], intensity vector for ambisonic format [19], combination of log-mel spectrogram and GCC-PHAT [20]. The common architectures for DOA estimation networks are fully connected (FC) network [14], [15], convolutional neural network (CNN) [17], [18] and convolutional recurrent neural network (CRNN) [16]. CRNNs are able to output DOA estimates at finer time scale than CNNs at the cost of higher computational load. Table I shows some of the existing DL-based DOA estimation methods.

The DL approach often formulates the DOA estimation of multiple sources as multi-label multi-class classification. The continuous range of DOAs of interest are often discretized at lower resolution compared to those of SP-based methods to limit the number of classes. The output of the deep models is the posterior probability of a sound source at each possible direction, which is assumed to be independent of each other. Similar to SP-based methods, in order to determine the DOAs from the output probabilities of the DL models, DL-based methods also require a pre-determined threshold or the number of sound sources known in advance. Adavanne *et al.* [16] selected all the top probabilities that were greater than a pre-determined threshold value (0.5). He *et al.* [18] used a threshold value that optimized precision and recall to select the top probabilities and count the number of sources. Chakrabarty *et al.* [17] assumed the number of sources was known in advance to select the top probabilities. Perotin *et al.* [19] also assumed the number of sources was known and trained different models for different number-of-source cases, i.e. one model for one case. Ranjan *et al.* [21] included an output branch in the model architecture to estimate the number of sources, and used this estimate to select the top probabilities for DOA estimation. However, due to the noisy output probabilities, a magnitude threshold together with some heuristics were required to select the top probabilities in conjunction with the predicted number of sources.

C. Evaluation metrics

The most common metric to evaluate DOA estimation is the DOA error [16], which is calculated as the mean absolute difference between the reference and the estimated DOA. Hungarian algorithm [22] is later proposed to account for cases when the number of estimated DOAs is different from the number of reference DOAs [16]. In addition, some researches use a recall metric to measure the percentage of the number of directions that are correctly estimated [16], or an accuracy metric that measured the percentage of predictions that had DOA errors less than a pre-determined threshold [17], [18].

D. The proposed method

The SPS produced by SP-based methods contain useful and relevant directional information. It is relatively independent of the spectral contents of the input signals. However, when the number of sources and the level of noise and reverberation increase, the magnitude of the SPS reduces and the local maxima are less well-defined. In addition, if the number of sound sources is unknown, which is often the case in real-life scenarios, a single threshold to determine the source directions for all the cases is not optimal for all cases. On the other hand, the DL-based methods have great capacity to accommodate different acoustics scenes. However, the DL models struggles to learn directional information from audio signal and their performances are strongly dependent on a large amount of training data. Due to large model capacities, there is also a risk of unwanted association between the sound classes and the directional information in the DL models. We would like to combine the strengths of both SP-based and DL-based methods to robustly estimate DOAs of multiple sound sources in noisy and reverberant environments when the number of sound sources is unknown. We propose a method that use SPS as input features to a CNN model to estimate the DOAs. The SPS are used as input features to the CNN to reduce the tendency of the neural network to learn unwanted association between the sound classes and the directional information, and help the network generalize to unseen sound classes that are not used during training. Furthermore, other SP-based methods such as single-source time-frequency (TF) selection method to improve the quality of the SPS can be seamlessly incorporated with the SP-based SPS feature extractor. CNN is used to learn complex environments with different levels of noise, reverberation and different number of sources to select the peaks on the SPS. Before us, Adavanne *et al.* [16] proposed a two-stage DL-approach for DOA estimation that estimate SPS as intermediate outputs which had lower performance than our proposed approach.

In addition, majority of the SP-based methods for DOA estimations average the SPS across both frequency and time to produce a 1 directional (1D) SPS along a directional axis, which removes the temporal information of the sound events. In real-life scenarios, many overlapping sources often have different onsets, offsets. Therefore, we propose to use 2 directional (2D) short-time SPS that preserve temporal information as inputs to a CNN model to robustly estimate DOAs in multiple-sources cases. The output SPS of SP-based algorithms are averaged across the frequency dimension but not the time dimension to produce a 2D short-time SPS, which will increase the degrees of freedom for the directional estimation.

The DL-based methods for DOA estimation often formulate the problem as multi-label, multi-class classification. The number of discretized DOAs is equal to the number of classes. The number of classes is often much larger than the number of active sound sources. Therefore, the output predictions are mostly zeros. As a results, there is an imbalance in the loss of active sources and inactive sources. We proposed to use a weighted cross-entropy loss as the DOA estimation loss to

TABLE I
COMPARISON OF OUR PROPOSED METHODS WITH SOME EXISTING DL-BASED DOA ESTIMATION METHODS

Approach	# of Sources	Input Feature	Network Architecture	Predict # of Sources	Top-Probabilities Selection
Xiao <i>et al.</i> [14]	1	GCC-PHAT coefficients	Fully connected	No	Largest peak
Takeda <i>et al.</i> [15]	≤ 2	MUSIC eigenvector	Fully connected	No	Threshold
Adavanne <i>et al.</i> [16]	≤ 3	Magnitude and phase spectrum	CRNN	No	Threshold = 0.5
Chakrabarty <i>et al.</i> [17]	≤ 2 at frame level	phase spectrum	CNN	No	Assume # of source is known
He <i>et al.</i> [18]	≤ 2	GCC-PHAT	CNN	No	Threshold
Perotin <i>et al.</i> [19]	≤ 2	Intensity vector (ambisonic format)	CNN	No	Assume # of sources is known
Our proposed methods	≤ 4	Short-time Spatial Pseudo-spectrum	CNN	Yes	Use predicted # of sources

mitigate this problem.

Finally, because the the number of sound sources in real-life scenarios are often unknown, we proposed to use multi-task learning to estimate the number of sound sources together with the DOAs. During the inference, the number-of-sources estimates are used to select the top DOA prediction probabilities. As a result, a threshold to select top DOA probabilities is not required. Furthermore, multi-task learning of similar tasks helps regularizing the DL-model and reduces over-fitting [23].

In summary, the main contribution of the paper is a multi-task CNN network with weighted cross-entropy loss that use 2D SPS as input feature to robustly estimate DOAs of multiple sound sources in different noisy and reverberant environments even when the number of sources is unknown. The proposed approach combines strengths of both SP and DL domains. We evaluate the proposed DOA estimation algorithm in different acoustic environments up to 4 sources. We also introduce two more evaluation metrics to comprehensively evaluate the baseline and proposed methods. The proposed methods outperform both the state-of-the-art SP and DL-based methods for DOA estimation using our simulated and recorded test sets.

The rest of our paper is organized as follows. Section II presents our proposed DOA estimation algorithm. Section III describes the dataset and evaluation metrics that are used for experiments. Section IV briefly introduces some state-of-the-art DOA algorithms that are compared with our proposed methods. Section V shows the experimental results and discussion. Finally, we conclude the paper in Section VI.

II. DOA ESTIMATION USING SPATIAL PSEUDO-SPECTRUM AND CONVOLUTIONAL NEURAL NETWORK

Figure 1 shows the block diagram of our proposed DOA estimation algorithm, which consists of two blocks: the 2D short-time SPS feature extraction and the 2D CNN DOA estimation algorithm. We can both interpret the deep neural network as a robust peak detection algorithm for the SPS, and the SPS algorithm as the feature extractor for the deep neural network. The SPS algorithm computes a 2D short-time SPS using MUSIC or histogram analysis method. The CNN processes the short-time SPS to estimate the number of sources and the DOAs at all the directions of interest. The final DOAs are selected from the top probabilities using the estimated number of sources. Furthermore, an optional block to select single-source (SS) time-frequency (TF) bins, which contain energy from mostly one source, can be inserted before

the SPS algorithm [9], [10]. Our proposed method is applied to a planar circular omnidirectional microphone array to estimate 2D DOAs of multiple sound sources on the azimuth plane.

A. Short-time spatial pseudo-spectrum estimation

We propose to use short-time 2D SPS, which have an additional time dimension, to increase the degrees of freedom in the SPS to detect more sources correctly.

We use two algorithms, MUSIC and histogram analysis, to extract the SPS and examine the performance of each method in this paper. Because the number of sound sources is not known in advance, to simplify the computation of MUSIC spectra, we assume the W-disjoint orthogonality (WDO) condition [24], in which each TF bin is dominated by a single speaker, to compute the MUSIC SPS [12]:

$$P_{MUSIC}(\theta) = \frac{1}{a^H(\theta)Q_nQ_n^H a(\theta)}, \quad (1)$$

where $a(\theta)$ is the steering vector and Q_n is the noise subspace. The dimension of the noise space is $M - 1$, where M is the number of microphones. In real-life scenarios, WDO condition, which is mainly used for speech signal, does not hold for all the TF bins and results in suboptimal MUSIC spectra at those violated bins.

The SS histogram method [9] shows that the largest eigenvector $\mathbf{u}$ of the signal covariance matrix at each SS TF bin is a scaled version of the theoretical steering vector $\mathbf{a}(\mathbf{f}, \theta)$ at the source direction θ and the frequency f . The DOA θ at that SS TF bin is directly calculated by equating the theoretical steering vector $\mathbf{a}(\theta)$ to the largest eigenvector $\bar{\mathbf{u}}$ that is normalized by its first element [9]. The research in [9] used a 4-microphone array. We extend this method to a planar M -microphone array. Let microphone 1 be the reference microphone. The steering vector of a planar omnidirectional microphone array at frequency f and DOA θ is

$$\mathbf{a}(f, \theta) = \begin{pmatrix} 1 \\ \exp(-i2\pi f\tau_{12}) \\ \vdots \\ \exp(-i2\pi f\tau_{1M}) \end{pmatrix}, \quad (2)$$

where i is the imaginary unit and τ_{1j} is the TDOA between microphone j and the reference microphone:

$$\tau_{1j} = \frac{1}{c} \left[(\boldsymbol{\zeta}_j - \boldsymbol{\zeta}_1)^T \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix} \right], \quad (3)$$

Fig. 1. Block diagram of the proposed direction-of-arrival estimation algorithm. The spatial pseudo-spectrum algorithm processes the multi-channel audio input and produces a 2D short-time spatial spectrum. The 2D convolutional neural network inputs the image-like spectrum and output the estimated number of sources and the DOA posterior probabilities.

Fig. 2. Block diagram of the SS TF bin selection algorithm.

where c is the speed of sound, ζ_j and ζ_1 are the coordinate of microphone j and 1. The value of τ_{1j} can be found using the j^{th} element of the normalized largest eigenvector:

$$\hat{\tau}_{1j} = -\frac{1}{2\pi f} \angle \bar{u}_j \quad (4)$$

Equating Eq. (4) and Eq. (3) we have:

$$\underbrace{\begin{bmatrix} (\zeta_2 - \zeta_1)^T \\ \vdots \\ (\zeta_M - \zeta_1)^T \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}}_{\xi} = -\underbrace{\frac{c}{2\pi f}}_{\mathbf{b(f)}} \underbrace{\begin{bmatrix} \angle \bar{u}_2 \\ \vdots \\ \angle \bar{u}_M \end{bmatrix}}_{\mathbf{b(f)}}. \quad (5)$$

We solve for θ using the method of least squares: $\xi = \mathbf{A} \setminus \mathbf{b(f)}$. The estimated DOAs from all the TF bins are quantized and populated into a histogram.

In order not to use the WDO assumption, we select the SS TF bins using an algorithm shown in Fig. 2 [9]. The algorithm finds all the TF bins that contains energy from mostly one source by performing magnitude, onset, and coherence of tests on all the TF bins of the input spectrogram. Magnitude test finds TF bins that are above a noise floor to mitigate the effect of background noise. Onset test finds TF bins that belong to direct-path signals to reduce the effect of reverberation. Coherence test finds TF bins of which the covariance matrices are approximately rank-1. After all the SS TF bins are found, we compute the SPS of SS TF bins and aggregate them across the frequency dimension.

Figure 3 shows the SPS of a 4-source case using real-recorded data. The dimension of the single-source SPS for 2D DOA estimation is $K \times T$, where K is the number of discretized azimuths and T is the number of time steps. The figures on the left column are the short-time SPS with additional temporal information of the active durations. The figures on the right column are the aggregated 1D SPS of

which the local maxima indicate the DOAs. The peaks of the histograms are well-defined even for small gaps between two sources and sharper than the peaks of the MUSIC SPS. The SS TF selection produces cleaner short-time SPS. When the number of sources and levels of noise and reverberation change, the number of true SS TF bins vary and the coherence-to-diffuse ratios in the TF bins also vary. As a consequence, both the 1D and 2D SPS have high variability. In the next section, we present a CNN to reliably estimate the DOAs from these short-time SPS.

B. DOA estimation algorithm using convolutional neural network, multi-task learning, and weighted cross-entropy

For DOA estimation using DL-approach, VGG-like CNN is a common architecture of choice [17], [18]. Recurrent layers such as bidirectional gated recurrent unit (GRU) [16], [20] and bidirectional long short-term memory (LSTM) [19] are used to produced DOA estimates at finer time-step. In this research, we estimate DOAs for each 2-second block of data. Therefore, we chose a VGG-like CNN to learn the patterns of the image-like 2D short-time SPS. Fig. 4 shows the proposed network architecture, which consists of 10 CNN layers and 4 fully connected (FC) layers. The model has two output branches to estimate the number of sources and the directional probabilities. The azimuth range between 0° and 360° is discretized with a resolution of 5° . As a result, there are a total of $K = 72$ classes for DOA estimation. The 2-second long input signal consists of 173 frames. Therefore, the size of the input spectrum is 173×72 . Our experimental results show that average pooling is better than max pooling for DOA estimation. The dropout in front of the FC layers helps to reduce the model's over-fitting. The input SPS are normalized by a scalar mean and a scalar standard deviation that are estimated using all training SPS along both time and angular dimensions. The estimation of the number of sources is formulated as multi-class classification, and a softmax activation is used to compute the output probabilities. The output prediction for the number of sources can only be one value between 0 and N inclusively, where N is the pre-defined maximum number of sources. We set $N = 4$ in this study.

Fig. 3. Real-recorded SPS of a 4-source case using 7-mic circular array. Left column: short-time SPS. Right column: aggregated SPS across time and frequency. The dotted lines denote the ground-truth DOAs at 50° , 95° , 205° , and 230° .

Cross-entropy loss for multi-class classification is:

$$ndoa_loss = -\frac{1}{L} \sum_{l=1}^L \sum_{n=0}^N y_n^l \log \hat{y}_n^l, \quad (6)$$

where L is the number of samples and $l = 0 \dots L$ is the sample index, $n = 0 \dots N$ is the class index, $y_n^l \in \{0, 1\}$ is the label of class n of l^{th} sample, $\hat{y}_n^l \in [0, 1]$ is the model's estimated probability if the number of sources is n .

The DOA estimation is formulated as multi-label multi-class classification, and sigmoid activation is used to compute the DOA probabilities. Because the number of the active DOAs ($\leq N = 4$) is much smaller than the total number of possible DOAs $K = 72$, weighted binary cross-entropy loss [25] is used to increase the recall rate:

$$doa_loss = -\frac{1}{L} \sum_{l=1}^L \sum_{k=0}^K (\alpha y_k^l \log \hat{y}_k^l + (1 - y_k^l) \log(1 - \hat{y}_k^l)), \quad (7)$$

where K is the number of DOA classes, i.e. the number of possible DOAs, and $k = 0 \dots K$ is the DOA class index, $y_k^l \in [0, 1]$ is the ground-truth DOA of class k , $\hat{y}_k^l \in [0, 1]$ is the model's estimated probability of a source at direction k ,

Fig. 4. Multi-task CNN architecture to estimate DOAs and the number of sources.

and α is the positive weight. We set $\alpha = 100$ in this study. $y_k^l \in [0, 1]$ instead of $y_k^l \in \{0, 1\}$ is because we use label smoothing, which will be described later on. The total loss of the multi-task network is

$$loss = \gamma ndoa_loss + (1 - \gamma) doa_loss, \quad (8)$$

where γ is a weighting constant.

From the outputs of the network, the number of sources is determined as

$$ndoa = \arg \max_{0 \leq n \leq N} P(n|X), \quad (9)$$

where $P(n|X)$ is the posterior probability of the number of sources, given the input spatial spectrum X . From the number of active source $ndoa$, we select the top $ndoa$ probabilities of the DOA posterior probabilities. There is no threshold required for this post-processing process.

We use label-smoothing [18] to encode the ground-truth DOAs to regularize the network. The ground-truth DOA is encoded as a Gaussian with the maximum value centered at the true DOA instead of one-hot encoding:

$$p_\theta = \exp((\theta - \theta_{gt})^2 / \sigma^2), \quad (10)$$

where θ_{gt} is the ground-truth DOA, σ is the standard deviation that controls the spread of the curve. Fig. 5 show an example of the encoded ground-truth DOAs with 3 sources at 110° , 220° , and 300° . We use $\sigma = 2.5^\circ$. In addition to label-smoothing, we use random-cutout augmentation for the SPS [26] to reduce over-fitting. Random-cutout has been adopted as a common data augmentation for sound classification. It randomly set a rectangular area on the SPS to zeros.

Fig. 5. Encoding of ground truth DOAs using label smoothing.

TABLE II
DATASETS FOR EVALUATION

Dataset (length in hours)	Purposes
Simulated training set (11.1)	Train DL-based models Select thresholds for SP-based algorithms
Simulated validation set (2.7)	Select hyperparameters for DL-based models
Simulated test sets (15.6)	Evaluate DL and SP-based algorithms
Real-recorded test sets (15)	Evaluate DL and SP-based algorithms

III. EVALUATION

The proposed DOA estimation methods are evaluated and examined using a circular planar UMA-8 microphone array [27] to estimate 2D DOAs of multiple sound sources on the azimuth plane. The UMA-8 array is commercially available and ready-to-use with universal-serial-bus connection. It has 1 microphone in the center and 6 microphones on the perimeter. The radius of the array is 8.55 cm, thus the aliasing frequency is 4011 Hz. The microphones of the UMA-8 array are of microelectro-mechanical systems (MEMS) type so their frequency responses match each other and no calibration is required before using. We generated both simulated and real-recorded datasets to benchmark the proposed algorithm against other state-of-the-art SP and DL-based DOA estimation algorithms. We also introduced new evaluation metrics to better assess the advantages and disadvantages of each algorithm.

A. Datasets

Table II shows the list of datasets that are used to train, validate, and evaluate the performance of the proposed DOA estimation methods. It is time-consuming and labour-intensive to collect real-recorded data in different background noises and room environments to train DL models. Therefore, we generated simulated datasets to train the DL models for DOA estimation. We also recorded data from different real environments to evaluate the performances of the baseline and the proposed methods. Our experimental results showed that models trained with simulated data were able to generalize well to real-recorded data. The SP-based methods for DOA estimation do not require training data. However, they require a pre-determined threshold for peak selection. For a fair comparison with the DL algorithms, we used the simulated training set to select the best thresholds for the SP-based algorithms according to some criteria. Speech was used to generate simulated datasets, while both sound and speech were

used to record test sets to evaluate the generalization of the proposed algorithms to unseen data types.

Pyroomacoustic module [28] was used to generate room impulse responses (RIRs) for the simulated training, validation, and test datasets. The simulated room height was fixed at 2.7 m. The room width and length were uniformly sampled from the interval [3, 7] m and [4, 10] m respectively. The room reflection coefficients were selected such that the reverberation time RT60 was from a normal distribution with mean of 0.5 seconds, standard deviation of 0.6 seconds, and $RT60 \geq 0.2$ seconds. The simulated UMA-8 array was at a height of 1 m and randomly placed in the room such that it was at least 1.5 m away from the walls. The distances from the sound sources to the array were randomly selected such that the sound sources were at least 1 m away from the array. The number of sources was sampled uniformly between 0 and 4 sources. The DOAs were discretized between 0° and 360° with a resolution of 5° . The DOAs were randomly selected such that the polar distance between two sources was at least 10° . Clean speech was randomly selected from Librispeech dataset [29] to be convolved with the simulated RIRs. The Librispeech dataset includes two folders "dev-clean" and "test-clean". The "dev-clean" data were used for simulated training set. Half of the "test-clean" data were used for the simulated validation set, and the other half were used for the simulated test set. Real-recorded background noises from 3 different indoor locations (1 hour per location) were recorded using UMA-8 array and added into the simulated signals. The signal to noise ratios (SNR) were sampled uniformly between -5 dB and 20 dB. The real background noises were used to generate simulated data instead of simulated noise to reduce the mismatch between the simulated and real-recorded data.

To generate real-recorded test sets, we played back speech and sound from Genelec loudspeakers in 3 different rooms and record the signals using the UMA-8 array. Speech was randomly selected from "test-clean" section of Librispeech dataset. Environmental sound was selected from the DCASE 2016 isolated event dataset [30], which contained 11 sound classes for indoor environments: *clearing throat*, *coughing*, *door knock*, *door slam*, *drawer*, *laughter*, *keyboard*, *keys*, *page turning*, *phone ringing*, and *speech*. A turntable was used to rotate the microphone array with a step size of 5° . The dimensions of the 3 rooms were $9 \times 6 \times 2.7$ m³, $5 \times 3 \times 2.7$ m³, and $12 \times 6 \times 2.7$ m³. Their estimated reverberation times were approximately 300 ms, 200 ms, and 400 ms respectively. Fig. 6 shows the experimental set-up of the microphone array and the speakers in the first room. The recorded audio signals were randomly combined to artificially synthesize multi-source cases up to 4 sources. Each audio clip was randomly scaled with a factor that was randomly sampled from normal distribution of mean 1 and standard deviation of 0.4. The real-recorded background noises were added to the synthesized clips with SNR values sampled uniformly between -10 dB and 10 dB to increase the difficulty of the test data. The real-recorded background noises were used for 0 source cases in both simulated and real-recorded datasets. The length of one input sample was 2 seconds. The sizes of the simulated training set, simulated validation set, simulated test set, real-

Fig. 6. Experimental setup to record data from loudspeakers.

recorded speech test set, and real-recorded sound test set were 20000, 10000, 28000, 12000, and 12000 samples respectively.

B. Evaluation metrics

To evaluate the performance of the proposed DOA estimation method, we used the DOA error metric proposed in [16], which used Hungarian algorithm [22] to compute the minimum difference in degrees between the estimated DOAs and the ground truth DOAs in spherical coordinate even when the number of estimated DOAs was different from the ground truth number. The Hungarian algorithm looks at all possible pairwise assignments between the estimated DOAs and the ground-truth DOAs and returns the smallest distance sum. The pairwise distance between an estimated azimuth θ_e and an assigned ground truth azimuth θ_{gt} is $|\theta_e - \theta_{gt}|$. The DOA error is

$$doa_error = \frac{1}{\sum_{i=1}^L n_e^i} \sum_{i=1}^L H(\theta_e^i, \theta_{gt}^i), \quad (11)$$

where L is the number of test samples, n_e^i is the number of estimated DOAs of sample i , θ_e^i and θ_{gt}^i are the list of estimated and ground truth DOAs of sample i respectively, and $H(\cdot)$ is the Hungarian algorithm.

In addition, we also evaluated the usefulness of DOA estimation algorithms using recall, precision, and F1 score metrics. To use these metrics, the DOA estimation can be formulated as a classification task. The number of classes is equal to the number of discretized angles. However, in practice, it is not very useful to directly apply the formulas of recall, precision, and F1 score to the output of a DOA estimation algorithm. For example, if the ground-truth DOA is 20° , a DOA estimate of 15° is a wrong classification, but still a relevant estimate with error of 5° . To accommodate the adjacency property of the DOA estimation problem, we used a tolerance to determine the relevancy of the estimation results. A DOA estimate that is within a tolerance of the ground truth is considered a relevant estimate with some angle error. A DOA ground truth that is within a tolerance of a DOA estimate is considered a relevant ground truth with some error. With the

introduction of tolerance, DOA recall is defined as the ratio between the number of relevant ground truth DOAs and the total number of ground truth DOAs; DOA precision is defined as the ratio between all the relevant estimated DOAs and the total number of estimated DOAs; and DOA F1 score is the geometric mean of the DOA recall and the DOA precision. We used a tolerance of 10° to compute recall, precision and F1 score. The DOA recall is similar to accuracy that were used in [17], [18]

IV. EXPERIMENTS

We compared our proposed DOA estimation methods with several state-of-the-art SP and DL-based methods for DOA estimation. We used the simulated training set to train the proposed DOA estimation methods and the state-of-the-art DL-based methods. The model hyperparameters were selected using simulated validation set. All the DL-based DOA algorithms used the same model architecture that was shown in Section II-B. The simulated training set was used to select the required threshold for the SP-based DOA estimation methods.

A. Baseline methods

We denote our proposed approach that use CNN to estimate DOAs from short-time MUSIC SPS and histogram SPS as **MUSIC-CNN** and **HIST-CNN** respectively. If single-source TF bin selection was used, a prefix **SS** was added to the short form of the algorithm. If a threshold was required by the algorithm to determine DOAs, a suffix **T** was added to the short form of the algorithm. Our paper focuses on different spatial input features for DOA estimation, therefore, we used the same network architecture as shown in Section II-B with different state-of-the-art spatial input features for DOA estimation. The list of the baselines is as follows:

- **MUSIC-T, SS-MUSIC-T**: use MUSIC algorithm [12] to estimate DOAs; the number of source is not known in advance, a threshold is determined from the simulated training set to optimize F1 score,
- **HIST-T, SS-HIST-T**: use histogram method [9] to estimate DOAs, a threshold is determined from the simulated training set to optimize F1 score,
- **MP-CNN-T**: use magnitude and phase spectrogram [16] as input features to train a CNN model to estimate DOAs,
- **GCC-CNN-T**: use GCC-PHAT [18] as input features to train a CNN model to estimate DOAs,
- **LG-CNN-T**: use a combination of log-mel spectrogram and GCC-PHAT as input features [20], which is the current state-of-the-art input feature for DOA estimation, to train a CNN model to estimate DOAs,
- **LG-CNN**: use log-mel spectrogram and GCC-PHAT as input features [20] to train a CNN models to estimate DOAs and the number of sources.

A threshold to select top DOA probabilities was tuned for MP-CNN-T, GCC-CNN-T, and LG-CNN-T to maximize F1 scores on the validation set. The estimated number of sources of the MUSIC-CNN, SS-MUSIC-CNN, HIST-CNN, SS-MUSIC-CNN, and LG-CNN models were used to select the largest probabilities in the estimated DOA output of the

network. All of the learning-based methods used random-cutout for data augmentation, label smoothing to encode the ground truths and weighted binary cross-entropy to increase the recall rate. The magnitude and phase spectrogram, log-mel spectrogram, and GCC-PHAT were normalized by their mean and standard deviation vector along the frequency dimension for each channel separately.

B. Hyper-parameters

We used a sampling rate of 44.1 kHz, window length of 1024 samples, hop length of 512 samples, Hann window, 1024 FFT points, 96 mel filters, GCC-PHAT length of 96 lags [20]. The length of one input sample was 2 seconds, which translated to $T = 173$ frames. The spatial aliasing frequency of the UMA8 array is around 4 kHz, so we selected SS TF bins between 50 Hz and 4 kHz. The azimuths were discretized with a resolution of 5° , which resulted in a total of $K = 72$ angles. The weighting constant γ was set to 0.1. Adam optimizer was used to train the convolutional neural network. We trained the DL-based DOA network for 80 epochs with the learning rate set to 0.001 for the first 50 epochs and reduced by 10% for each subsequent epoch. The same network architecture and training procedure were used for the DL-based and the proposed methods, except that the network architecture for MP-CNN-T, GCC-CNN-T, and LG-CNN-T had one output branch to estimate DOA instead of two output branches. The input feature sizes for MUSIC-CNN, SS-MUSIC-CNN, HIST-CNN, and SS-HIST-CNN were $1 \times 173 \times 72$, while the input feature sizes for MP-CNN-T was $14 \times 173 \times 72$, the input feature size for GCC-CNN-T was $21 \times 173 \times 72$, and the input feature sizes for LG-CNN-T and LG-CNN were $28 \times 173 \times 72$.

V. RESULTS AND DISCUSSION

We compared our proposed methods with 8 baselines. All of these methods were carefully tuned for our simulated and recorded data to produce a fair comparison among different spatial input features. We first evaluated the performance of DOA estimation, then we evaluated the performance of the number-of-sources estimation. Finally, we examined the required training data for some of the learning-based approaches and their processing time.

A. Experimental results for DOA estimation

The performances of the baselines and our proposed methods on the simulated and real-recorded test sets are shown in Table III. Among the SP-based methods, the histogram-based methods (HIST and SS-HIST) were better than the MUSIC-based methods (MUSIC and SS-MUSIC). The SS TF bin selection improved the performance of the histogram-based methods across all of the evaluation metrics on both simulated and real-recorded data. The performances of the SP-based methods on the real-recorded test set were lower than those on the simulated test set. The best algorithm on both simulated and real-recorded test sets was single-source histogram analysis (SS-HIST). In general, the DL-based methods for DOA estimation performed better than the SP-based counterparts

on the simulated data. The DL-based algorithms had lower DOA error, and much higher recall. However the recall rate dropped significantly on the real-recorded test set compared to the recall rate on the simulated test set. This reduction can be explained by the mismatches between the real-recorded data and the simulated data, which included differences between the real and simulated environments, as well as differences between microphones used to record the LibriSpeech dataset and the microphone array used to record the real test set. We tried to reduce these mismatches by using real-recorded background noise to generate simulated data. In reality, the data mismatches are inevitable because of different number of sound sources, different room environments, and different background noise conditions. In addition, the frequency responses of simulated microphone signals are different from the frequency responses of the real microphones. The MP-CNN-T model that estimated DOAs from the raw audio input features (magnitude and phase spectrogram) has the highest degree of deterioration in performance, whereas the GCC-CNN-T models that use intermediate input features (GCC-PHAT) are more robust. Models that use GCC-PHAT together with raw audio feature log-mel is less robust than the model that only uses GCC-PHAT. The performance of LG-CNN was slightly lower than LG-CNN-T. Multi-task learning had to jointly optimize both the losses of DOA estimation and the number-of-source estimation. Therefore, it potentially caused lower performance for each task. The best DL-based model on both test sets was GCC-CNN-T, which used GCC-PHAT as input feature, and its best threshold to select top DOA probabilities was determined from the simulated validation set. GCC-CNN-T achieved the lowest DOA error, and highest F1 score among all the DL-based model. The proposed methods outperformed the DL-based methods and the SP-based methods on both test sets. In addition, the proposed methods were able to generalize well to the real-recorded test set, with smallest drop in performances compared to other methods. It is interesting to note that single-source TF bin selection did not help the performances of those proposed methods as it did for the SP-based methods. The best proposed algorithm was HIST-CNN, which used short-time histogram and CNN to estimate the number of sources and DOAs. On the simulated test set, it achieved lowest DOA error of 1.4° , highest F1 score of 0.973. On the real-recorded test set, it achieved third lowest DOA error of 2.8° , highest F1 score of 0.903. Both the F1 scores were significantly larger than those of SP and DL-based methods.

Table IV examines the performances of the baselines and the proposed methods on the real recorded test set according to different data types: speech and sound. Environmental sounds were not used during training, and these data were used to test the generalization of all the DOA algorithms to unseen data types. Because environmental sound is less sparse than speech [31], there are more overlapping in the TF domain, which leads to fewer SS TF bins. As a result, the performances of the SP-based methods on the recorded-sound test set were lower than those on the recorded-speech test set. The performance gaps of the DL-based methods between the recorded-speech and recorded sound test sets were larger

TABLE III
DOA ESTIMATION RESULTS USING SIMULATED AND REAL-RECORDED TEST SETS. SHADED CELLS INDICATE THE BEST PERFORMANCE WITHIN EACH ALGORITHM GROUP

Group	Method	Simulated test set				Recorded test set (speech + sound)			
		error	recall	precision	F1	error	recall	precision	F1
SP-based methods	MUSIC	7.4°	0.667	0.838	0.743	8.7°	0.622	0.820	0.707
	SS-MUSIC	6.4°	0.686	0.857	0.762	4.8°	0.663	0.904	0.765
	HIST	2.0°	0.853	0.980	0.912	3.2°	0.746	0.934	0.829
	SS-HIST	2.4°	0.874	0.973	0.921	1.3°	0.779	0.989	0.871
DL-based methods	MP-CNN-T	3.2°	0.930	0.953	0.941	3.8°	0.468	0.973	0.632
	GCC-CNN-T	1.9°	0.944	0.979	0.961	2.7°	0.785	0.966	0.866
	LG-CNN-T	1.9°	0.940	0.982	0.961	3.2°	0.753	0.957	0.843
	LG-CNN	2.5°	0.927	0.977	0.951	3.5°	0.747	0.968	0.843
Proposed methods	MUSIC-CNN	2.2°	0.948	0.960	0.954	2.9°	0.816	0.954	0.880
	SS-MUSIC-CNN	4.4°	0.927	0.890	0.908	3.7°	0.809	0.931	0.866
	HIST-CNN	1.4°	0.969	0.978	0.973	2.8°	0.860	0.950	0.903
	SS-HIST-CNN	2.5°	0.947	0.967	0.957	1.8°	0.837	0.977	0.902

than those of the SP-based methods. These results showed that there was some degree of association between the signal types and the estimated DOAs when using DL-based methods for DOA estimation. When the test signals were of different classes, the method performed poorly. Among all the metrics, the recall rate dropped the fastest. The best of the SP-based methods, SS-HIST, outperformed the best of the DL-based methods, GCC-CNN-T on the recorded-sound test set. Our proposed methods outperformed both the SP and DL-based methods on the recorded-sound test set by a large margin. This showed that estimating DOAs from SPS was more robust to the signal mismatches between the test data and train data compared to the DL-based methods that used magnitude and phase spectrum or GCC-PHAT as input features.

Table V shows the DOA estimation results using recorded test sets for different number of sources. As the number of sources increased, the performance of all the methods decreased as expected. However, the proposed methods consistently outperformed all other methods and the performance gaps between the proposed methods and other methods widened when the number of source increased. These results showed that the proposed methods were more robust to the multiple-sources scenarios than other DOA estimation methods. The SS-HIST-CNN has slightly higher DOA errors but higher F1 scores than the SP-based counterpart. When there were 4 sources, the HIST-CNN achieved the highest F1 scores of 0.881 with moderate DOA error of 3.1°.

B. Experimental results for number-of-source estimation

Table VI shows the experimental results for the estimation of the number of sources. All the proposed methods outperformed both the SP-based and DL-based algorithms on the ability to correctly estimate the number of sources. HIST-CNN achieved the highest accuracy of 0.717, whereas the best DL-based method, GCC-CNN-T, achieved an accuracy of 0.604, and the best SP-based method, SS-HIST, achieved an accuracy of 0.562. Fig. 7 shows the confusion matrix for the estimation

Fig. 7. Confusion matrix for the estimation of the number of sources using HIST-CNN method

Fig. 8. Performance vs amount of training data

of the number of sources using the HIST-CNN method. The proposed methods worked well when the number of sources was small. As the number of sources increased, the proposed methods underestimated more often than overestimated the number of sources, which in turn affected the recall more than the precision of the DOA estimation.

C. Amount of training data

Figure 8 shows the DOA error and F1 score on the recorded test set for the GCC-CNN-T, LG-CNN, SS-HIST-CNN, and HIST-CNN when the number of simulated training samples varied from 5000 to 20000. Even with small number of training samples, the proposed methods (HIST-CNN and SS-

TABLE IV
DOA ESTIMATION RESULTS USING RECODED-SPEECH AND RECORDED-SOUND TEST SETS. SHADED CELLS INDICATE THE BEST PERFORMANCE WITHIN EACH ALGORITHM GROUP

Group	Method	Recorded-speech test set				Recorded-sound test set			
		error	recall	precision	F1	error	recall	precision	F1
SP-based methods	MUSIC	6.9°	0.635	0.855	0.729	10.3°	0.609	0.786	0.686
	SS-MUSIC	4.1°	0.690	0.918	0.788	5.6°	0.636	0.889	0.741
	HIST	2.3°	0.786	0.962	0.865	4.1°	0.706	0.905	0.793
	SS-HIST	1.3°	0.807	0.988	0.889	1.2°	0.750	0.990	0.853
DL-based methods	MP-CNN-T	2.7°	0.607	0.984	0.751	5.8°	0.328	0.954	0.489
	GCC-CNN-T	2.1°	0.864	0.978	0.917	3.5°	0.707	0.952	0.811
	LG-CNN-T	2.3°	0.856	0.973	0.911	4.3°	0.651	0.936	0.768
	LG-CNN	3.2°	0.841	0.969	0.901	3.8°	0.653	0.967	0.780
Proposed methods	MUSIC-CNN	2.2°	0.855	0.968	0.925	3.7°	0.747	0.939	0.832
	SS-MUSIC-CNN	2.8°	0.864	0.963	0.911	4.7°	0.753	0.897	0.819
	HIST-CNN	2.2°	0.911	0.955	0.932	3.3°	0.810	0.944	0.872
	SS-HIST-CNN	1.5°	0.883	0.983	0.930	2.2°	0.791	0.971	0.872

TABLE V
DOA ESTIMATION RESULTS FOR VARIOUS NUMBER OF SOURCES USING RECORDED TEST SET

Method	1 source		2 sources		3 sources		4 sources	
	error	F1	error	F1	error	F1	error	F1
MUSIC	2.4°	0.796	11.5°	0.775	10.7°	0.705	8.5°	0.634
SS-MUSIC	1.0°	0.895	5.4°	0.833	5.8°	0.758	5.4°	0.686
HIST	0.7°	0.886	4.4°	0.894	3.9°	0.835	2.9°	0.770
SS-HIST	0.2°	0.990	0.8°	0.930	1.5°	0.865	1.7°	0.808
MP-CNN-T	5.1°	0.777	4.1°	0.709	3.5°	0.626	3.3°	0.549
GCC-CNN-T	1.4°	0.915	2.6°	0.911	2.9°	0.872	3.1°	0.823
LG-CNN-T	2.5°	0.875	3.2°	0.889	3.4°	0.852	3.3°	0.801
LG-CNN	2.0°	0.956	3.3°	0.889	3.7°	0.836	3.9°	0.791
MUSIC-CNN	0.7°	0.938	2.5°	0.914	3.3°	0.878	3.5°	0.846
SS-MUSIC-CNN	0.7°	0.905	2.9°	0.903	3.9°	0.865	5.1°	0.834
HIST-CNN	1.1°	0.927	2.6°	0.927	3.1°	0.905	3.1°	0.881
SS-HIST-CNN	0.3°	0.975	1.2°	0.943	2.0°	0.898	2.6°	0.861

TABLE VI
THE-NUMBER-OF-SOURCES ESTIMATION RESULTS USING RECORDED TEST SET

Method	Accuracy
MUSIC	0.359
SS-MUSIC	0.395
HIST	0.493
SS-HIST	0.562
MP-CNN-T	0.354
GCC-CNN-T	0.604
LG-CNN-T	0.564
LG-CNN	0.565
MUSIC-CNN	0.652
SS-MUSIC-CNN	0.612
HIST-CNN	0.717
SS-HIST-CNN	0.696

HIST-CNN) achieved good performance on both DOA error and F1 score compared to the DL-based methods (GCC-CNN-T and LG-CNN). This results suggested that it was more effective to learn the DOA information from the short-time SPS, which succinctly contained DOA information, than to learn the DOA information from GCC-PHAT or log-mel GCC-PHAT features.

D. Processing time

The advantages of the proposed methods come at the cost of higher computation. Compared to the SP-based methods, a 2D CNN for peak detection was more expensive than a threshold-based peak picking procedure. The SPS algorithms were more computationally expensive if SS TF bin selection was not used because the DOA estimation had to be done for all of the TF bins instead of only a handful of selected bins. Table VII shows the average processing time to extract spatial features for DOA estimation from a 2-second audio input of a 7-microphone array. The specifications of the computer that were used for processing data were Intel i9 3.3 GHz 10

Core 20 Threads with 64GB DDR4 RAM. The fastest feature extraction method was magnitude and phase spectrogram. Extracting short-time SPS using HIST and MUSIC methods took the longest time. However, extracting short-time SPS using SS-HIST and SS-MUSIC were still slightly faster than the state-of-the-art spatial features GCC-PHAT and log-mel GCC-PHAT for DOA estimation. The performances of SS-HIST-CNN were lower than those of HIST-CNN and MUSIC-CNN but higher than those of GCC-CNN and LG-CNN. In addition, the input size to the CNN of the proposed methods were much smaller than those state-of-the-art spatial features. Therefore, it was faster to train the model and do the inference using the proposed methods than the DL-based methods.

E. Visual evaluation with real recordings of human speakers

To evaluate the effectiveness of the proposed method in real recordings with human speaker, we recorded a short audio clip with two human subjects, one male speaker and one female speaker. The data was recorded in a residential

TABLE VII
AVERAGE PROCESSING TIME TO EXTRACT FEATURES FROM 2-SECOND
AUDIO SEGMENT OF 7-MICROPHONE ARRAY.

Feature extraction method	Average processing time (s)
magnitude and phase spectrogram	0.04
GCC-PHAT	0.33
logmel + GCC-PHAT	0.34
MUSIC	1.19
SS-MUSIC	0.29
HIST	1.20
SS-HIST	0.30

Fig. 9. DOA estimation results of two human speakers using SS-HIST-CNN

room with dimension of $3 \times 4 \times 2.7 \text{ m}^3$. This room was not used to record data for the real-recorded dataset. The male speaker was static at 300° . The female speaker started to speak at 10^{th} second at 240° for a duration of 10 seconds and slowly move to 120° . The audio clip was divided into chunks of 2 seconds with overlap of 0.5 seconds. Fig. 9 show the DOA estimation results using the SS-HIST-CNN method. No temporal smoothing was done on the results. The estimated number of sources were used to select the peaks. We can see that the proposed method SS-HIST-CNN was able to detect two speakers even in dynamic situations with a few mistakes. These mistakes can be potentially mitigated by applying some post-processing step.

VI. CONCLUSION

DOA estimation has many applications in real-life. In order for the algorithm to be effective, there is a need for an end-to-end evaluation from feature extraction, directional estimation, and DOA selections. In this paper, we propose to use a multi-task CNN network to estimate the number of sources and DOAs from short-time SPS. The short-time SPS contain useful DOA information with temporal information and decouple the DOA cues from other signal characteristics. CNN models are able to learn patterns on the SPS in various conditions. The proposed methods combine the strengths of both SP and DL domain. They are more robust to different data types, different rooms environments, and different number of sources. The SPS can be extracted using several methods such as MUSIC or histogram analysis. A SS TF bin selection algorithm can be added to speed up the calculation. One important contribution of this research is that we do not make any assumption about the number of sources. Instead, the estimated number

of sources is used to select the final DOAs without using a fixed threshold. The experimental results showed that our proposed methods significantly outperformed other methods in estimating the number of sources and DOAs. The proposed methods were verified and tested for 2D DOA estimation using a planar omnidirectional array. These methods can also be extended to 3D DOA estimation by vectorizing both azimuth and elevation angles. As a result, one dimension of the short-time SPS is still the DOAs of interest, and the other dimension is time.

From the signal processing point of view, 2D CNN is a useful tool for peak detection. From the deep learning point of view, the effectiveness of different spatial input features for DOA estimation were analysed in this study. We compared our proposed short-time SPS input features with the state-of-the-art spatial features for DOA estimation such as magnitude and phase spectrogram, GCC-PHAT, log-mel and GCC-PHAT. The experimental results showed that the short-time SPS consistently outperformed other features across all the evaluation metrics, and increased the generalization of the DOA estimation network to unseen conditions, while kept the processing time lower than those of GCC-PHAT, log-mel and GCC-PHAT. To our knowledge, this is the first time short-time SPS are used as the input features to a deep network to solve DOA estimation problem.

In addition, we also unify several evaluation metrics for DOA estimations and introduce two new metrics that have been widely used in deep learning community to evaluate performance of an algorithm from a classification perspective. The detailed evaluation metrics provide interesting and valuable insight into the performances of the proposed methods as well as other SP and DL-based DOA estimation methods. Experimental results showed that, with the same objective function to maximize F1 score, the DL-based DOA algorithms were excellent in optimizing DOA precision but had lower DOA recall when data mismatch occurred. The SP-based DOA algorithms generally had lower DOA precision than the DL-based algorithms. The DL-based models that use more intermediate spatial features such as GCC-PHAT has lower degrees of deterioration in performance than models that use features such as magnitude and phase spectrogram. Our proposed methods consistently outperformed other SP and DL-based DOA estimation methods using our simulated and recorded datasets. The best proposed method based on the evaluation metrics was the HIST-CNN. When processing time was taken into account, the SS-HIST-CNN was a better choice. The performance of the SS-HIST-CNN was higher than those of SP and DL-based methods, while its computation time was much lower than the HIST-CNN and slightly lower than other DL-based methods.

ACKNOWLEDGEMENT

This research was conducted at Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU), which is a collaboration between Singapore Telecommunications Limited (Singtel) and Nanyang Technological University (NTU) that is funded by the Singapore Government through

the Industry Alignment Fund - Industry Collaboration Projects Grant.

The authors would like to thank Mr. Kenneth Ooi and Mr. Sathish Jayabalan for helping us with the data collection for the experiment.

REFERENCES

- [1] P. Transfeld, U. Martens, H. Binder, T. Schypior, and T. Fingscheidt, "Acoustic event source localization for surveillance in reverberant environments supported by an event onset detection," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 2629–2633.
- [2] J. M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *IEEE International Conference on Robotics and Automation, ICRA'04*, vol. 1. IEEE, 2004, pp. 1033–1038.
- [3] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 1997, pp. 187–190 vol.1.
- [4] T. V. Bogaert, E. Carette, and J. Wouters, "Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna," *International Journal of Audiology*, vol. 50, no. 3, pp. 164–176, 2011.
- [5] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.
- [6] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.
- [7] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.
- [8] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug 1969.
- [9] T. N. T. Nguyen, S. K. Zhao, and D. L. Jones, "Robust doa estimation of multiple speech sources," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2287–2291.
- [10] D. Pavlidis, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, Oct 2013.
- [11] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2009, pp. 2027–2032.
- [12] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.
- [13] B. Rafaely and D. Kolossa, "Speaker localization in reverberant rooms based on direct path dominance test statistics," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 6120–6124.
- [14] X. Xiao, S. K. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Z. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.
- [15] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 603–609.
- [16] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 1462–1466.
- [17] S. Chakrabarty and E. A. P. Habets, "Multi-speaker doa estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, March 2019.
- [18] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 74–79.
- [19] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, March 2019.
- [20] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pp. 30–34, 2019. [Online]. Available: <https://arxiv.org/abs/1905.00268>
- [21] R. Ranjan, S. Jayabalan, T. N. T. Nguyen, and W. S. Gan, "Sound events detection and direction of arrival estimation using residual net and recurrent neural networks," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pp. 214–218, 2019.
- [22] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [24] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2002, pp. I–529–I–532.
- [25] Weighted cross-entropy with logits. 2020. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/nn/weighted_cross_entropy_with_logits. Accessed: 2019-09-30.
- [26] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 13001–13008, 2020.
- [27] "Uma-8 usb mic array - v2.0," <https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array>, accessed: 2019-09-30.
- [28] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 351–355.
- [29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.
- [30] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcse 2016 challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 2, pp. 379–393, 2018.
- [31] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, May 2019.

Thi Ngoc Tho Nguyen is a Ph.D. student at the Nanyang Technological University (NTU) in Singapore. Prior to joining NTU, she worked at University of Illinois Research Center in Singapore for five years as a research engineer. Her research interests are audio signal processing, deep learning, microphone array signal processing, and real-time processing.

Woon-Seng Gan received his BEng (1st Class Hons) and PhD degrees, both in Electrical and Electronic Engineering from the University of Strathclyde, UK in 1989 and 1993 respectively. He is currently a Professor of Audio Engineering in the School of Electrical and Electronic Engineering in Nanyang Technological University. He also served as the Head of the Information Engineering Division in the School of Electrical and Electronic Engineering in Nanyang Technological University from 2011-2014. He is currently the Director, Centre for In-

focomm Technology @ NTU, which was formed in 2011 and brings together expert in the area of Information and Communication Engineering.

His research has been concerned with the connections between the physical world, signal processing and sound control, which resulted in the practical demonstration and licensing of spatial audio algorithms, directional sound beam, and active noise control for headphones. He has published more than 330 international refereed journal and conference papers, and has translated his research into 6 granted US/Singapore patents. He had co-authored three books on Subband Adaptive Filtering: Theory and Implementation (John Wiley, 2009); Embedded Signal Processing with the Micro Signal Architecture, (Wiley-IEEE, 2007); and Digital Signal Processors: Architectures, Implementations, and Applications (Prentice Hall, 2005). In 2017, he won the APSIPA Sadaoki Furui Prize Paper Award. He is a Fellow of the Audio Engineering Society(AES), a Fellow of the Institute of Engineering and Technology(IET), and a Senior Member of the IEEE. He also served as an Associate Editor of the IEEE/ACM Transaction on Audio, Speech, and Language Processing (TASLP; 2012-15); IEEE Signal Processing Letters (SPL; 2016-2019); and Associate Technical Editor of the Journal of Audio Engineering Society (JAES; 2013-).

Rishabh Ranjan received his B.Tech degree in Electronics & Communication Engineering from IIIT Hyderabad, India and Ph.D. degree in electrical and electronic engineering, Nanyang Technological University (NTU), Singapore, in 2015. Currently, he is working as senior researcher at Hitachi Asia Ltd., developing novel solutions in the field of social innovations, applying people centric solutions using artificial intelligence. His research interests include machine learning, signal processing, 3D audio, active noise control. His research work has resulted in

more than 20 international refereed publications including IEEE/ACM Trans. Audio, Speech, and Language Processing (TASLP), IEEE Potential, JAES, ICASSP, Applied Acoustics, and Internoise.

Douglas L. Jones received the BSEE, MSEE, and Ph.D. degrees from Rice University in 1983, 1986, and 1987, respectively. During the 1987-1988 academic year, he was at the University of Erlangen-Nuremberg in Germany on a Fulbright postdoctoral fellowship. Since 1988, he has been with the University of Illinois at Urbana-Champaign, where he is currently a Professor in Electrical and Computer Engineering, Neuroscience, the Coordinated Science Laboratory, and the Beckman Institute. He was on sabbatical leave at the University of Washington

in Spring 1995 and at the University of California at Berkeley in Spring 2002. In the Spring semester of 1999 he served as the Texas Instruments Visiting Professor at Rice University. He is an author of two DSP laboratory textbooks, and was selected as the 2003 Connexions Author of the Year. He is a Fellow of the IEEE. He served on the Board of Governors of the IEEE Signal Processing Society from 2002-2004. His research interests are in digital signal processing, including nonstationary signal analysis, adaptive processing, multisensor data processing, OFDM, and various applications such as low-power implementations, biology and neuroengineering, and advanced hearing aids and other audio systems.