DCASE 2020 TASK 3: ENSEMBLE OF SEQUENCE MATCHING NETWORKS FOR DYNAMIC SOUND EVENT LOCALIZATION, DETECTION, AND TRACKING

Technical Report

Thi Ngoc Tho Nguyen[1]*, *Douglas L. Jones*[2], *Woon Seng Gan*[1],

[1] Nanyang Technological University, School of Electrical and Electronic Engineering, Singapore,
{nguyenth003, ewsgan}@ntu.edu.sg
[2] University of Illinois at Urbana-Champaign, Dept. of Electrical and Computer Engineering,
Illinois, USA, {dl-jones}@illinois.edu

ABSTRACT

Sound event localization and detection consist of two subtasks which are sound event detection and direction-of-arrival estimation. While sound event detection mainly relies on time-frequency patterns to distinguish different sound classes, direction-of-arrival estimation uses magnitude or phase differences between microphones to estimate source directions. Therefore, it is often difficult to jointly train two subtasks simultaneously. Our previous sequence matching approach solves sound event detection and direction-of-arrival separately and trains a convolutional recurrent neural network to associate the sound classes with the directions-of-arrival using onsets and offsets of the sound events. This approach achieved better performance than other state-of-the-art networks such as the SELDnet, and the two-stage networks for static sources. Experimental results on the new DCASE dataset for sound event localization, detection, and tracking of multiple moving sound sources showed that the sequence matching network also outperformed the jointly trained SELDnet model. In order to estimate directions-of-arrival of moving sound sources with high spatial resolution, we proposed to separate the directional estimations into azimuth and elevation before passing them to the sequence matching network. We combined several sequence matching networks into ensembles and achieved a sound event detection and localization error of 0.217 compared to 0.466 of the baseline.

Index Terms— CRNN, DCASE, direction-of-arrival estimation, sequence matching network, sound event detection.

1. INTRODUCTION

Sound event localization and detection (SELD) has many applications in urban sound sensing [1], wild life monitoring [2], surveillance [3], autonomous driving [4], and robotics [5]. The SELD task recognizes the sound class, and estimates the direction-of-arrival (DOA), the onset, and offset of a detected sound event [6]. Polyphonic SELD refers to cases where there are multiple sound events overlapping in time. DCASE2020 challenge introduces a new SELD dataset with multiple moving sound sources [7]. Many existing SELD algorithms are frame-based, therefore they extend naturally to the additional task of tracking moving sound sources.

SELD consists of two subtasks, which are sound event detection (SED) and direction-of-arrival estimation (DOAE). In the past decade, deep learning has achieved great success in classifying, tagging, and detecting sound events [8]. The state-of-the-art SED models are often built from convolutional neural networks (CNN) [1], recurrent neural networks (RNN) [9], and convolutional recurrent neural networks (CRNN) [6, 10]. DOAE tasks for small-aperture microphone arrays are often solved using signal processing algorithms such as minimum variance distortionless response (MVDR) beamformer [11], multiple signal classification (MUSIC) [12], and steered-response power phase transform (SRP-PHAT) [13]. To tackle the multi-source cases, many researches exploit the non-stationarity and sparseness of the audio signals to find the single-source time-frequency (TF) regions on the spectrogram to reliably estimate DOAs [14, 15, 16]. Recently, deep learning has also been successfully applied to DOAE tasks [17, 18], and the learning-based DOA models show good generalization to different noise and reverberation levels. However, the angular estimation error is still high for multi-source cases.

To solve SELD problem, Adavanne *et al.* proposed a single-input multiple-output CRNN model called SELDnet that jointly detects sound events and estimates DOAs [6]. The model's loss function is a weighted sum of the individual SED and DOAE loss functions. The SELDnet is also applied to solve SELD for multiple moving sources [19]. Because SED and DOAE requires different acoustic information from the audio inputs, the joint estimation affects the performance of both tasks. To mitigate this problem, Cao *et al.* proposed a two-stage strategy for training SELD models [20]. First, a SED model using a CRNN architecture is trained by minimizing the SED loss function using all the available data. After that, the CNN weights of the SED model is transferred to the DOA model, which has the same architecture as the SED model. The DOA model is trained by minimizing the DOA loss function using only the data that have active sources. The SED outputs are used as masks to select the corresponding DOA outputs. This training scheme significantly improves the performance of the SELD system. However, the DOA model is still dependent on the SED model for detecting the active signals, and the network learns to associate specific sources with specific directions in the training data. In the DCASE 2019 challenge, the top solution trained four separated models for sound activity detection (SAD), SED, single-source and two-source DOAE respectively [21]. This solution heavily used heuristic rules to determine the single-source and two-source segments of the signal to infer the sound classes and DOAs. This approach is highly dependant on the estimation of the number of sources which are not reliable in noisy environments.

*This research was supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2017-T2-2-060.

Our previous research proposed a novel two-step approach that decoupled the learning of the SED and DOAE systems [22]. In the first step, we used Cao's CRNN model [20] to detect the sound events, and a single-source histogram method [15] to estimate the DOAs. In the second step, we trained a CRNN-based sequence matching network (SMN) to match the two output sequences of the event detector and DOA estimator. The motivation of this approach is that overlapping sounds often have different onsets and offsets. By matching the onsets, the offsets, and the active segments in the output sequences of the sound event detector and the DOA estimator, we can associate the estimated DOAs with the corresponding sound classes. This modular and hierarchical approach significantly improves the performance of the SELD task across all the evaluation metrics. We applied our two-step method for the DCASE2020 SELD challenge with several modifications. First, our experimental results showed that spatial features such as generalized cross-correlation phase transform (GCC-PHAT) [20] for microphone format and intensity vector [23] for ambisonic format does not help SED in the moving sources cases. Therefore we only used logmel spectrogram to train SED model. Second, the azimuth and elevation resolution of the DCASE2020 SELD dataset was 1° compared to 10° of the DCASE2019 SELD dataset, therefore the size of the joint 2D single-source histogram significantly increased. The large dimension of the 2D histogram is not optimal for the SMN, therefore we proposed to use 1D histograms for azimuth and elevation separately instead of the joint 2D azimuth-elevation histogram as inputs to the SMN. Third, to boost performance, we combined several SED models into a SED ensemble to train several SMN models, which in turn were combined to form a SMN ensemble. The rest of our paper is organized as follows. Section II describes our SMN network for SELD. Section III presents the experimental results and discussions. Finally, we conclude the paper in Section IV.

2. SEQUENCE MATCHING NETWORK FOR SOUND EVENT LOCALIZATION AND DETECTION

Figure 1 shows the block diagram of a SMN for SELD. The SED network is similar to the one proposed by Cao *et al* [20]. The DOAE module uses a non-learning signal processing approach to robustly estimate the DOAs of sound sources regardless of the sound classes [15]. The output sequences of the SED network and DOAE module are the inputs of the SMN. The SMN uses CNN layers to learn patterns on the azimuth and elevation histogram before concatenating them with the SED inputs. A bidirectional gated recurrent unit (GRU) is used to match the DOA and SED sequences. Fully connected (FC) layers are used to produce the final SELD estimates. The SED subtask is formulated as multi-label multi-class classification. The DOAE subtask is formulated as regression of the Cartesian coordinates on a unit sphere.

2.1. Sound event detection

We use a CRNN-based SED network that uses log-mel spectrogram as input features. Our experimental results show that GCC-PHAT and intensity vector are not helpful for detecting multiple moving sound sources. The DCASE2020 SELD dataset has 14 sound classes with various length in noisy environments. To improve the SED performance, we use various data augmentation such as random cut-out, erasing columns of time steps and rows of frequency bands [24], mixup, and frequency shift.

Table 1: A CRNN-based SED network for 14 sound classes

Stage	Layer description
conv1	(conv2d 64 3x3, BN, ReLu) x 2, 2x2 average pooling
conv2	(conv2d 128 3x3, BN, ReLu) x 2, 2x2 average pooling
conv3	(conv2d 256 3x3, BN, ReLu) x 2, 2x2 average pooling
pooling	average pooling frequency dimension
GRU	bidirectional GRU 128
FC	dropout(0.2), FC 14, sigmoid
total parameters	1454122

The SED base network consists of 6 CNN layers, 1 bidirectional GRU layer, and 1 FC layer as shown in Table 1. The SED is formulated as multi-label multi-class classification. We use the raw probability outputs of the SED network as the input to the SMN in step 2. We modify the base SED network in term of pooling size and number of filters to produce several variants. The outputs of these models are averaged to produce an SED ensemble.

2.2. Direction-of-arrival estimation

We use a single-source histogram algorithm proposed in [15] to estimate DOAs. The single-source histogram finds all the time-frequency (TF) bins that contains energy from mostly one source. A TF bin is considered to be a single-source TF bin when it passes all three tests: magnitude, onset, and coherence test. Magnitude test finds the TF bins that are above a noise floor to mitigate the effect of background noise. Onset test finds the TF bins that belong to direct-path signals to reduce the effect of reverberation in the DOA estimation. Coherence test finds the TF bins of which the covariance matrices are approximately rank-1. After all the single-source TF bins are found, the DOA at each bin is computed using the theoretical steering vector of the microphone array [15]. These DOAs are discretized using the required resolution of azimuth and elevation angles. Subsequently, these DOAs are populated into 2 1D histograms, one for azimuth, one for elevation. Our experimental results show that DOA estimation without onset slightly increase the DOA frame recall but slightly increase the DOA error. The overall SELD error is improved without onset detection. A resolution of 5° for both azimuth and elevation are used to estimate the 1D azimuth and 1D elevation histogram. The sizes of the azimuth and elevation histograms for each time frame are 72 and 19, respectively. Fig. 2 shows the estimated azimuth and elevation histograms for a two-source 60-s audio clip. Visually, the single-source histogram algorithm accurately estimates the azimuths with clear onsets and offsets for moving sound sources even for narrow angular distance. The elevation estimates are more blurry than the azimuth estimates.

2.3. Sequence matching network

SMN is a multiple-input multiple-output CRNN. The input features to the SMN are the SED prediction probabilities, 1D azimuth and 1D elevation histograms. The outputs of the SMN are the SED prediction probabilities and the DOA Cartesian coordinate on the unit sphere. Similar to the baseline, our experimental results show that regression using Cartesian coordinate format results in lower DOA errors than polar coordinate format does. The SMN consists of 4 CNN blocks for each of the azimuth and elevation histograms, 2 bidirectional GRU layer, and 4 FC layer as shown in Fig. 1. Table 2 shows the details of the SMN. We train the base SMN with different input lengths of 4, 6, 8, 10, and 15 seconds and combines these

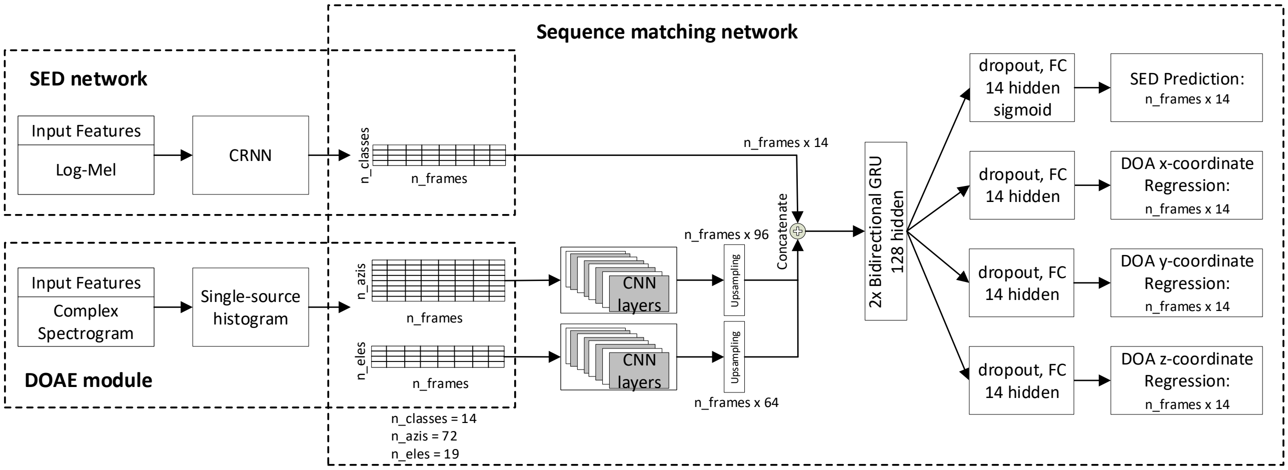


Figure 1: Block diagram of the two-step sound event localization and detection. Step 1: SED network and DOA module generate SED and DOA output sequences (1D azimuth and elevation histograms for each time step). Step 2: Sequence matching network matches the sound classes, azimuths and elevations for detected sound events. n_{frames} is the number of time frames of one training samples, $n_{classes}$ is the number of sound classes, n_{azis} is the number of azimuths, and n_{eles} is the number of elevations.

Table 2: A CRNN-based SMN network. The table entries are not in sequence. Refer to Fig.1

Stage	Layer description
azi conv1	(conv2d 16 3x3, BN, ReLu) x 2, 2x2 average pooling
azi conv2	(conv2d 32 3x3, BN, ReLu) x 2, 2x2 average pooling
azi conv3	(conv2d 64 3x3, BN, ReLu) x 2, 2x2 average pooling
azi conv3	(conv2d 96 3x3, BN, ReLu) x 2, 2x2 average pooling
azi pooling	average pooling angle dimension and upsampling
ele conv1	(conv2d 8 3x3, BN, ReLu) x 2, 2x2 average pooling
ele conv2	(conv2d 16 3x3, BN, ReLu) x 2, 2x2 average pooling
ele conv3	(conv2d 32 3x3, BN, ReLu) x 2, 2x2 average pooling
ele conv3	(conv2d 64 3x3, BN, ReLu) x 2, 2x2 average pooling
ele pooling	average pooling angle dimension and upsampling
concatenate	SED , azimuth feature, elevation feature
GRU	(bidirectional GRU 128) x 2
SED FC	dropout(0.2), FC 14, sigmoid
DOA-x FC	dropout(0.2), FC 14
DOA-y FC	dropout(0.2), FC 14
DOA-z FC	dropout(0.2), FC 14
total parameters	829427

different models into a SMN ensemble by averaging the SED and DOA outputs.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

We used the FOA format of the DCASE2020 SELD dataset [7] for the challenge. The SELD development dataset consists of 600 one-minute audio clips divided into training, validation, and test set of size 400, 100, and 100 clips, respectively. All 600 clips were used to train models for evaluation. There are 14 sound classes. The sound durations are between 0.3 and 15 seconds. The azimuth and elevation ranges are $[-180°, 180°)$ and $[-45°, 45°]$, respectively. We used azimuth and elevation resolutions of $5°$.

3.1. Evaluation metrics

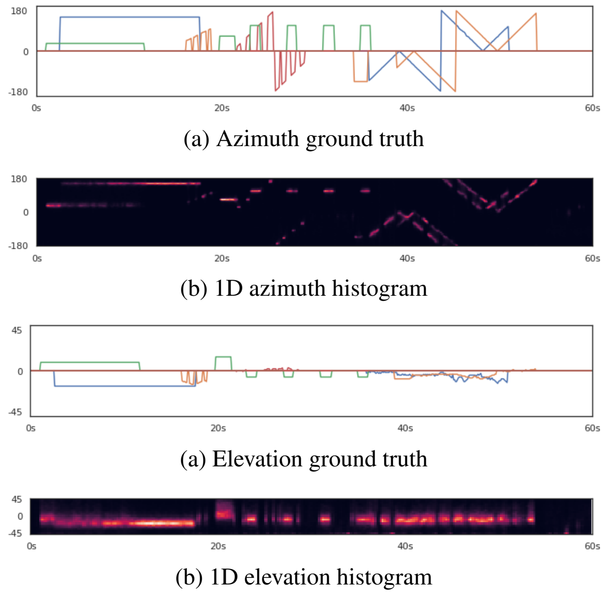
The SELD task was evaluated for SED and DOAE subtask separately in the 2019 SELD challenge. This year, a new evaluation metrics that take into account the correct association between sound classes and DOA are introduced [25]. A sound event is considered correct detection if it has correct class prediction and its estimated DOA is less than $20°$ from the DOA ground truth. The DOA metrics are computed for each class before averaging. Since we solved SED and DOAE separately before joining them, both 2019 and 2020 evaluation metrics were used in our experiments. The 2019 version was used to evaluate the performance of SED networks and DOAE modules separately. The 2020 version was used to evaluate the performance of the SMNs.

3.2. Hyper-parameters and training procedure

Hyper-parameters for processing raw audio signals are sampling rate of 24 kHz, window length of 1024 samples, hop length of 300 samples (12.5 ms), Hann window, and 1024 FFT points. 128 mel bands were used to extract log-mel features. For the single-source histogram estimation, we used magnitude signal-to-noise ratio of 1.5 for the magnitude test, and a condition number of 5 for the coherence test. Adam optimizer was used to train the SED networks and the SMNs. We train the SED network for 50 epochs with the learning rate set to 0.001 for the first 30 epochs and reduced by 10% for each subsequent epoch until it reaches 0.0001. We train the SMN for 60 epochs with the learning rate set to 0.001 for the first 30 epochs and reduced by 10% for each subsequent epoch.

3.3. Our challenge submissions

We combined the outputs of 4 SED models to form an SED ensemble. This SED ensemble are used to train 6 SMN models using the same SMN base network with different input lengths. The outputs



(a) Azimuth ground truth

(b) 1D azimuth histogram

(a) Elevation ground truth

(b) 1D elevation histogram

Figure 2: 1D azimuth and elevation histograms of a two-source audio clip. The classes are color coded in the ground truths. Pictures viewed better with color

Table 3: Submissions for DCASE2020 SELD challenge.

Submission name	Descriptions
SMN-EN-1	ensemble of 4 SMN models, SED threshold = 0.3
SMN-EN-2	ensemble of 5 SMN models, SED threshold = 0.3
SMN-EN-3	ensemble of 6 SMN models, SED threshold = 0.3
SMN-EN-4	ensemble of 6 SMN models, use best SED threshold on the test set

of the 6 SMN models are combines to form several SMN ensembles. The best 4 SMN ensembles evaluated using the provided test set are selected for submissions. The details of the 4 submissions are as shown in Table 3.

3.4. SELD baselines and SMNs

We compared the performance of our submissions with the following methods.:

- **Baseline**: A CRNN-based network called SELDnet that jointly train SED and DOAE [6], with log-mel and spatial input features and masking loss for DOA [20],
- **SED-base**: the base model for SED as shown in Section 2.1,
- **SED-en**: ensemble of 4 different SED models which are variants of SED-base,
- **SS-hist**: single-source histogram for DOAE estimation. The DOA are selected as highest peaks of the joint 2D azimuth-elevation histogram that above a certain threshold,
- **SMN-base**: the SMN network for SELD as shown in Section 2.3. This network take the class prediction probabilities of the SED-en and 1D histograms of the single-source histogram method as input features,

Table 4: SELD development results using validation set

Methods	Metrics	ER	F	DE	FR	SELD
Baseline	2019	0.53	64.3	19.5°	68.4	0.327
Baseline	2020	0.72	39.1	24.0°	64.3	0.455
SED-base	2019	0.239	85.0	NA	NA	NA
SED-EN	2019	0.180	88.9	NA	NA	NA
SS-hist	2019	NA	NA	6.6°	74.7	NA
SMN-base	2019	0.196	88.3	10.6°	77.7	0.149
SMN-base	2020	0.305	76.2	11.7°	88.4	0.181
SMN-EN-1	2020	0.292	77.3	10.3°	88.7	0.172
SMN-EN-2	2020	0.291	77.3	10.4°	**88.8**	0.172
SMN-EN-3	2020	**0.290**	77.4	10.2°	88.7	**0.171**
SMN-EN-4	2020	**0.290**	**77.6**	**10.1°**	88.7	**0.171**

Table 5: SELD development results using test set

Methods	Metrics	ER	F	DE	FR	SELD
Baseline	2019	0.54	60.9	20.4°	66.6	0.345
Baseline	2020	0.72	37.4	22.8°	60.7	0.466
SED-base	2019	0.299	80.7	NA	NA	NA
SED-EN	2019	0.278	81.6	NA	NA	NA
SS-hist	2019	NA	NA	8.5°	73.2	NA
SMN-base	2019	0.272	81.4	11.3°	77.8	0.186
SMN-base	2020	0.381	69.4	13.5°	81.5	0.237
SMN-EN-1	2020	0.356	71.5	**12.0°**	81.9	0.222
SMN-EN-2	2020	0.357	71.4	12.1°	82.0	0.221
SMN-EN-3	2020	0.359	71.2	12.1°	82.0	0.223
SMN-EN-4	2020	**0.349**	**71.9**	12.1°	**82.7**	**0.217**

3.5. SELD experimental results

The SELD development results of the validation and test set using both the 2019 and 2020 evaluation metrics consistently showed that our SMN-base and SMN ensembles outperformed the baseline SELDnet by a large margin. The 2020 metrics penalized the mismatching between sound classes and their DOA estimates, therefore their scores were lower than those of the 2019 metrics. Using the official 2020 evaluation metrics, the SED error rates and the DOA errors of the SMN-base reduced almost by half compared to those of the baseline. On the test set, the F1 score of the SMN-base is 69.4% compared to 37.4% of the baseline, and the DOA frame recall of the SMN-based is 82.7% compared to 60.7% of the baseline.

The SED-en model slightly improved the SED error rate and F1 score compared to those of SED-base model. Likewise, the SMN-en models improved the SELD performance across all the metrics compared to the individual SMN-base model. All the SMN-en ensembles had similar performance. We observed that all the models performed much better on the validation set compared to the test set. A close examination showed that the SED performance for the *male-shouting* class on the test set is particularly poor.

4. CONCLUSION

We submitted 4 ensembles for the DCASE 2020 SELD challenge. We combined several SMNs into these ensembles. We found that the SMN approach significantly outperformed the SELDnet baseline. In the SMN approach, we solved SED and DOAE separately to optimize the performance of each tasks. After that, a CRNN-based SMN was used to match the onsets, offsets, sound classes and DOAs.

5. ACKNOWLEDGMENT

This research was supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2017-T2-2-060.

6. REFERENCES

[1] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, March 2017.

[2] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: A survey and a challenge," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2016, pp. 1–6.

[3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, Jan 2016.

[4] M. K. Nandwana and T. Hasan, "Towards smart-cars that can listen: Abnormal acoustic event detection on the road." in *INTERSPEECH*, 2016, pp. 2968–2971.

[5] 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.

[6] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, March 2019.

[7] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv e-prints: 2006.01919*, 2020. [Online]. Available: https://arxiv.org/abs/2006.01919

[8] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[9] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.

[10] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[11] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug 1969.

[12] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[13] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 581–585, 2014.

[14] 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.

[15] N. T. N. Tho, 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.

[16] A. Griffin, D. Pavlidi, M. Puigt, and A. Mouchtaris, "Real-time multiple speaker doa estimation in a circular microphone array based on matching pursuit," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Aug 2012, pp. 2303–2307.

[17] 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.

[18] 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.

[19] S. Adavanne, A. Politis, and T. Virtanen, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 20–24.

[20] Y. Cao, Q. 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)*, 2019. [Online]. Available: https://arxiv.org/abs/1905.00268

[21] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," DCASE2019 Challenge, Tech. Rep., June 2019.

[22] T. N. Tho Nguyen, D. L. Jones, and W. Gan, "A sequence matching network for polyphonic sound event localization and detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 71–75.

[23] 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.

[24] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[25] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, Oct 2019, accepted.