A Model Ensemble Approach for Sound Event Localization and Detection

Qing Wang[1], *Huaxin Wu*[2], *Zijun Jing*[2], *Feng Ma*[2], *Yi Fang*[2],
Yuxuan Wang[1], *Tairan Chen*[1], *Jia Pan*[2], *Jun Du*[*1], *Chin-Hui Lee*[3]

[1] University of Science and Technology of China, Hefei, China
[2] iFLYTEK, Hefei, China
[3] Georgia Institute of Technology, Atlanta, USA

{qingwang2,✉jundu}@ustc.edu.cn, {yxwang1,vea,panjia}@mail.ustc.edu.cn
{hxwu2,zjjing2,fengma,yifang2}@iflytek.com, {chl}@ece.gatech.edu

2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 978-1-7281-6994-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISCSLP49672.2021.9362116

Abstract

In this paper, we propose a model ensemble approach for sound event localization and detection (SELD). We adopt several deep neural network (DNN) architectures to perform sound event detection (SED) and direction-of-arrival (DOA) estimation simultaneously. Generally, the DNN architecture consists of three modules stacked together, i.e, a High-level Feature Representation module, a Temporal Context Representation module, and a Fully-connected module in the end. The High-level Feature Representation module usually contains a series of convolutional neural network (CNN) layers to extract useful local features. The Temporal Context Representation module aims to model longer temporal context dependency in the extracted features. There are two parallel branches in the Fully-connected module with one for SED estimation and the other for DOA estimation. With different combinations of implementation in the High-level Feature Representation module and Temporal Context Representation module, several network architectures are used for the SELD task. At last, a more robust prediction of SED and DOA is obtained by model ensemble and post-processing. Tested on the development and evaluation datasets, the proposed approach achieves promising results and ranks the first place in DCASE 2020 task3 challenge.
Index Terms: sound event localization and detection, deep neural network, model ensemble

1. Introduction

Sound event localization and detection (SELD) aims to recognize individual sound events, identify their temporal activities, and estimate their spatial location when active. An effective SELD approach is able to describe the temporal and spatial characterization of acoustic scenes that can be applied in many areas. Environmental noise types like "footsteps" or "keyboard" can be recognized, tracked, and then suppressed to improve speech quality for video conferences or for robust automatic speech recognition (ASR) [1, 2]. In smart homes and smart cities, the SELD approach can be used for audio surveillance [3].

The SELD task consists of two subtasks, which are sound event detection (SED) and direction-of-arrival (DOA) estimation. Traditional statistical modelling methods for SED include Gaussian mixture model (GMM) - hidden Markov model (HMM) [4], non-negative matrix factorization (NMF) [5], and support vector machines (SVM) [6]. In recent years, neural network architectures have been successfully employed for SED task, including feed-forward neural network (FNN) [7], convolutional neural network (CNN) [8, 9], and recurrent neural network (RNN) [10, 11]. Capsule neural network (CapsNet) were adopted in [12, 13] for Polyphonic SED task to separate individual sound events from their mixture. A recently published model which combined CNN, RNN, and FNN together, referred as the convolutional recurrent neural network (CRNN) [14, 15], achieved state-of-the-art results for SED task.

As for DOA estimation, approaches can be categorized into two kinds: parametric-based and deep neural network (DNN)-based. Parametric-based approaches include multiple signal classification (MUSIC) [16], estimation of signal parameters via rotational invariance technique (ESPRIT) [17], steered-response-power phase transform (SRP-PATH) [18], and so on. Because of the high regression capability of DNNs, they have been employed to estimate sound event direction [19, 20, 21]. In [22], a DOA estimation method was proposed by parametric-based and DNN-based approach.

The SELD task, which was held for the first time in DCASE 2019 challenge, is focused on a combined task of SED and DOA estimation, instead of treating them as two separate tasks. To solve SELD problem, the author used CNN to jointly estimate both spatial location and audio content type [23]. Adavanne *et al.* adopted CRNN model for SELD of multiple overlapping sound events in three-dimensional (3-D) space [24]. In [25], the authors proposed a two-stage strategy for SELD task and achieved great improvements. The top solution in DCASE 2019 Challenge used four CRNN single output models to predict the number of active sources, DOA of single source, DOA of two sources, and the types of sound events, respectively [26]. However, this approach is highly dependent on the accuracy for predicting the number of active sources which is not reliable in noisy acoustic scenes.

In this paper, we present a model ensemble approach for the SELD task in DCASE 2020 challenge [27]. The main difference from DCASE 2019 challenge is that joint metrics [28], namely location-dependent detection and class-dependent localization, are adopted for SED and DOA evaluation. Since the performance of SED and DOA estimation affects each other, we think it is appropriate to jointly perform SED and DOA estimation. Data augmentation methods are used to expand the official dataset, which will be discussed in detail in another paper. The SELD models we adopt consist of a High-level Feature Representation module, a Temporal Context Representation module, and a Fully-connected module. Residual neural network (ResNet) [29] and Xception model [30] achieve great performance in image recognition, and are used to learn useful local features. Bidirectional gated recurrent unit (GRU) and factorized time delay neural network (TDNN-F) [31] used in ASR are adopted to model longer temporal context dependency in the audio signal. Two parallel branches of fully-connected

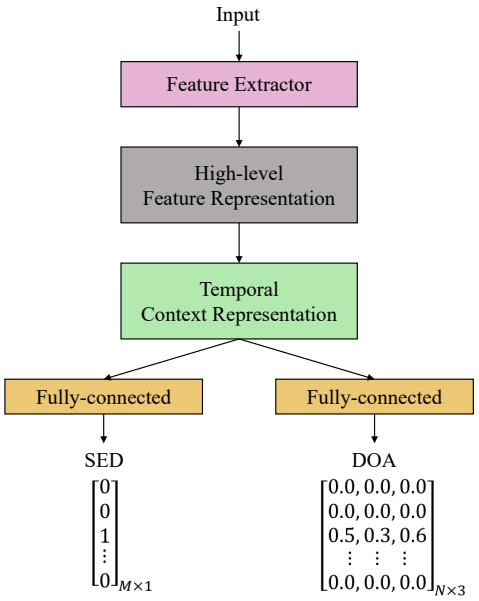


Figure 1: *An overview of the DNN architecture for SELD, where N denotes sound event classes. M is equal to N if sigmoid activation is used in the output layer of SED branch, whereas M is equal to $N + 1$ if softmax activation function is used.*

layers are employed for SED and DOA respectively. Model ensemble and post-processing strategies are finally used to generate a more accurate SED and DOA estimation.

The rest of the paper is organized as follows. In Section 2, the proposed approach is described in detail, including feature extraction, network architecture, model ensemble and post-processing. Evaluation results on development dataset is shown in Section 3. Conclusions are summarized in Section 4.

2. Proposed Approach

An overview of the DNN architecture for SELD task is shown in Figure 1. The input is multichannel audio signal. Before fed into the DNN, feature extraction is performed to obtain acoustic features. Similar to the official baseline SELDnet [24], the DNN takes a sequence of features which will be described in detail in the following subsection 2.1 and predicts active sound event classes along with their respective spatial locations. The DNN architecture consists of three modules stacked together, i.e, a High-level Feature Representation module marked in grey, a Temporal Context Representation module marked in green, and a Fully-connected module marked in yellow in the end. We employ different models for the High-level Feature Representation and Temporal Context Representation modules, which will be discussed in detail in the following subsection 2.2. The Fully-connected module contains two parallel branches for SED which is obtained as a multiclass-multilabel classification task and DOA estimation which is performed as a multioutput regression task. The SED output is thresholded to get the active sound events, whereas the corresponding DOA estimations are referred to as their spatial locations.

2.1. Feature extraction

A new dataset TAU-NIGENS Spatial Sound Events 2020 [32] is released for the SELD task in DCASE 2020 challenge. Two different 4-channel spatial sound formats, namely first-order Ambisonics (FOA) and tetrahedral microphone array (MIC), are opted from the synthesized sound recordings. We extract two features for each of the two datasets, FOA and MIC. The multichannel audio signal is sampled at 24 kHz. Using a short-time Fourier transform (STFT) with a hamming window of length 1024 samples and a 50% overlap, linear spectrogram for each channel is extracted. Then 64 log mel-band feature is extracted for both datasets. The second feature is format-specific. For FOA dataset acoustic intensity vector (IV) computed at each of the 64 mel-bands is extracted and for MIC dataset generalized cross-correlation with phase transformation (GCC-PATH) computed in each of the 64 mel-bands is extracted similar to [25]. Finally, there are 4 channels of log mel features and 3 channels of IV features, hence up to 7 feature maps for FOA signals. For MIC signals, there are 4 channels of log mel features and 6 channels of GCC-PATH features, hence up to 10 feature maps. We use both FOA and MIC datasets, 17 input feature maps are used to train the DNN architecture.

2.2. Network architecture

The output of feature extraction is fed into the DNN architecture as shown in Figure 1. Usually the High-level Feature Representation module consists of a series of CNN blocks, each CNN block usually having a 2D convolution layer followed by a batch normalization process, a rectified linear unit (ReLU), and a max-pooling operation. As ResNet [29] and Xception [30] show great advantages in image processing, we adopt the modified versions of them in the High-level Feature Representation module to learn local shift-invariant features in the SELD task. The detailed parameters of ResNet and Xception are shown in Figure 2(a) and Figure 2(b), respectively.

The output of the High-level Feature Representation module is then fed into the Temporal Context Representation module in order to learn the temporal structures within sound events. We use two bidirectional RNN layers with each containing 128 GRU cells to exploit the full context information of an input audio. Besides RNN, TDNN-F, which is a factored form of TDNN, is also an efficient architecture for temporal context modelling and performs well in ASR [31]. The TDNN-F architecture used in SELD task consists of several CNN layers with dilated convolution, which allows it to learn longer receptive field in temporal context. The detailed parameters of TDNN-F is shown in Figure 3.

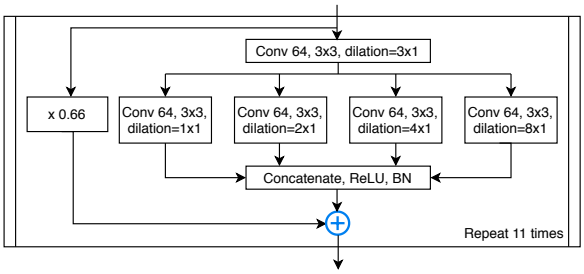
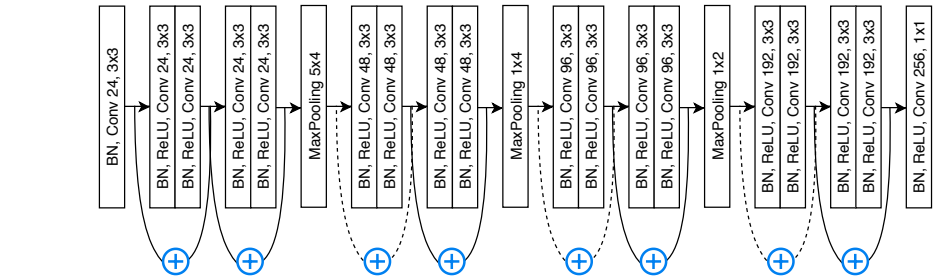
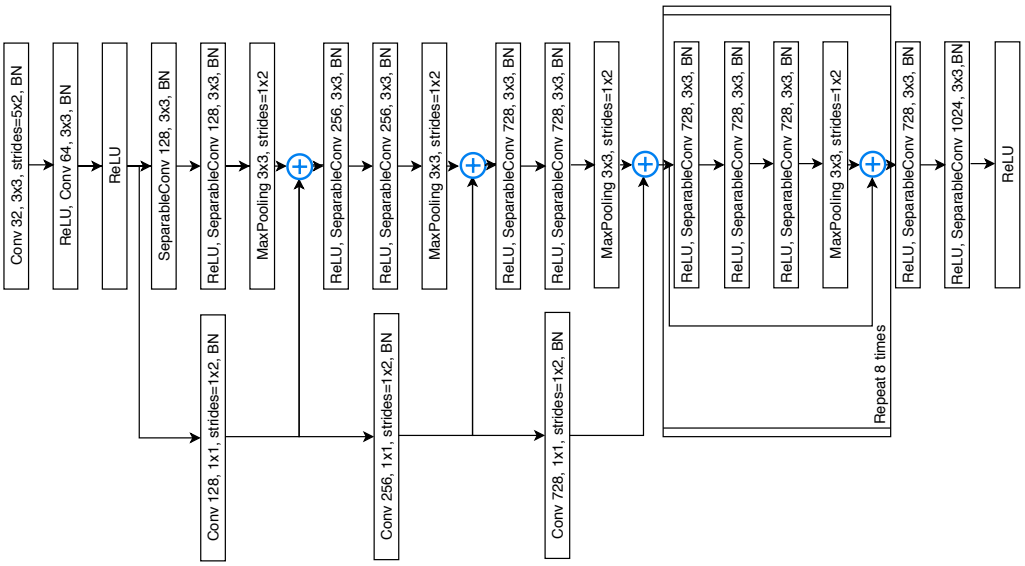


Figure 3: *TDNN-F architecture used in the Temporal Context Representation module.*

The Temporal Context Representation module is followed by two parallel branches of fully-connected (FC) layers, where the first FC layer in both branches contains 512 neurons with linear activation. We adopt two ways to perform SED. One



(a) ResNet



(b) Xception

Figure 2: *a) ResNet and b) Xception architectures used in the High-level Feature Representation module.*

way is similar to the official baseline where the last FC layer in the SED branch contains N neurons with sigmoid activation, and N is equal to the sound event classes. Another way is to predict $N + 1$ classes in the last FC layer with an additional silence class, and use softmax activation. Since the number of overlapping sound events is up to two, if overlapping sound events are active, the probabilities of corresponding labels of SED branch are both set to 0.5. For DOA branch, the last FC layer contains $3N$ neurons with tanh activation for multioutput regression, corresponding to the Cartesian coordinates (x, y, z) of all sound event classes as shown in Figure 1.

2.3. Loss function

As for the loss function, we solve the SED task as a multilabel classification with a binary cross-entropy (BCE) loss. Suggested by the second-best team [25] in DCASE 2019 challenge, a masked mean square error (MSE) loss is adopted for the DOA estimation performed as a multioutput regression. The masked MSE loss is computed based on the ground truth activations of each sound event class, hence not contributing to the training when the sound event is not active. The SED classification loss and DOA regression loss are combined for joint optimization during training with a weight [1, 10].

2.4. Model ensemble

It is well known that fusing the outputs of several models trained with different model architectures usually help to improve system performance over individual models. Here we perform model ensemble by using the weighted mean of the outputs predicted by different DNN architectures as the final result. Both the outputs of SED and DOA branches are averaged with weights. Different DNN models are assigned with different weights.

2.5. Post-processing

Instead of using a global threshold for all sound events, we adopt a dynamic threshold for the ensemble result. An optimal threshold is chosen for each sound event on the validation set.

3. Development Results

3.1. Experimental setup

The proposed approach is evaluated on TAU-NIGENS Spatial Sound Events 2020 [32] which contains 600 60-second audio recordings with a 24 kHz sampling rate. These recordings are split into six folds with 4 folds for training, 1 fold for validation and 1 fold for testing. We use several data augmentation

methods to expand the official dataset which will be discussed in detail in another paper later. Totally, there are 14 sound classes of the spatialized events: Alarm, Crying baby, Crash, Barking dog, Running engine, Female scream, Female speech, Burning fire, Footsteps, Knocking on door, Male scream, Male speech, Ringing phone, Piano.

Performance of SELD task is evaluated using a newly proposed joint detection and localization metrics [28]. Two metrics are used for both SED and DOA estimation. For location-dependent detection, error rate (ER_{20°) and F-score (F_{20°) are computed, considering that a true positive is predicted only when the spatial error for the detected event is within the given threshold of 20° from the reference. The two metrics for class-dependent localization are localization error (LE_{CD}) expressing the average angular distance between predictions and references of the same class and localization recall (LR_{CD}) expressing the true positive rate of how many of these localization estimates are detected in a class out of the total class instances. All four metrics are computed based on one-second segments.

The audio clips with a length of 60 seconds are used for training. All DNN architectures are trained with Adam optimizer. The learning rate is set to 0.001 and is decreased by 50% if the SELD score does not improve in 80 consecutive epochs. For single DNN models, if sigmoid activation is used in the last FC layer of SED branch, we adopt a threshold of 0.5. Otherwise if softmax activation is used, we adopt a threshold of 0.33 to make sure overlapping sound events can be detected. For the ensemble model, we adopt a dynamic threshold.

3.2. Experimental results

By employing different architectures in the High-level Feature Representation module and Temporal Context Representation module, we trained several DNN models for the SELD task. Specifically, ResNet and Xception were used as High-level Feature Representation modules, whereas bidirectional GRU and TDNN-F were used as Temporal Context Representation modules, which resulted in four models with combinations of these two modules, namely ResNet-GRU, ResNet-TDNNF, Xception-GRU, and Xcetption-TDNNF when using sigmoid activation in the last FC layer of SED branch. For a comparison between activations for performing SED, we also trained a DNN model using softmax activation in the last FC layer of SED branch, denoted as ResNet-GRU-softmax. Table 1 shows the experimental results of the proposed approach for the development dataset.

Table 1: *Experimental results of the proposed approach for development dataset.*

	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
Baseline-FOA	0.72	37.4	22.8°	60.7
Baseline-MIC	0.78	31.4	27.3°	59.0
ResNet-GRU	0.29	76.4	9.4°	82.8
ResNet-GRU-Softmax	0.29	76.2	9.1°	81.6
ResNet-TDNNF	0.31	76.1	8.7°	81.3
Xception-GRU	0.33	73.5	10.0°	80.8
Xception-TDNNF	0.36	71.1	10.0°	78.8
Model Ensemble	0.26	79.9	6.8°	84.1
+ Post-processing	**0.26**	**80.9**	**6.8°**	**85.1**

The first two rows in Table 1 present the official baseline systems trained with FOA dataset and MIC dataset respectively. As shown in the table, all of the proposed individual DNN models outperform the baseline systems by a large margin for each metric. Among all the DNN models, using ResNet in the High-level Feature Representation module achieved better results than using Xception. By comparing the third and fourth row in the table, we found adopting softmax activation in the last FC layer of SED branch got similar results with using sigmoid activation. Although Xception architecture performed a little worse than ResNet, ensemble of the models was still effective, which demonstrated the complementarity between different DNN structures. With post-processing in the end, F_{20° and LR_{CD} could still be improved by 1 point. Compared to the Baseline-FOA system, the proposed approach showed 63.9% relative improvement on ER_{20° metric, 116.3% relative improvement on F_{20° metric, 80.2% relative improvement on LE_{CD} metric, and 38.6% relative improvement on LR_{CD} metric respectively.

Figure 4 shows an example of SED output of the proposed approach from test split, where the 14 sound event classes are listed in subsection 3.1. When using ResNet-GRU alone, one single sound event is recognized as two sound events as shown in the red dashed rectangular box in Figure 4(c). With the proposed approach, namely model ensemble and post-processing, the detection error has been corrected as shown in the red rectangular box in Figure 4(d), which demonstrates the effectiveness of the proposed approach.

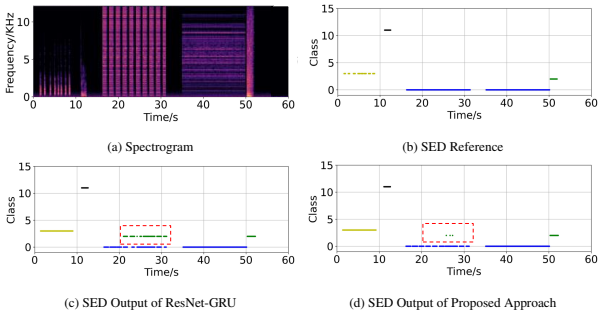


Figure 4: *Visualization of SED output of the proposed approach.*

4. Conclusions

In this paper, we proposed a model ensemble approach for SELD task which achieved the first place in DCASE 2020 challenge. We adopted one single model to solve SED and DOA estimation simultaneously with multitask learning. To exploit the complementarity between different model architectures, several DNN architectures were proposed, namely ResNet-GRU, ResNet-TDNNF, Xception-GRU, and Xception-TDNNF. Model ensemble and post-processing strategies were used to obtain more accurate SELD estimation. The experimental results evaluated on the development dataset showed that the proposed approach outperformed the baseline systems by a significant margin.

5. Acknowledgement

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Tencent.

6. References

[1] 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. IEEE, 1997, pp. 187–190.

[2] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[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, 2015.

[4] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *2010 18th European Signal Processing Conference*. IEEE, 2010, pp. 1267–1271.

[5] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[6] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE transactions on Neural Networks*, vol. 14, no. 1, pp. 209–215, 2003.

[7] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, 2015.

[8] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[9] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 559–563.

[10] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2742–2746.

[11] 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.

[12] F. Vesperini, L. Gabrielli, E. Principi, and S. Squartini, "Polyphonic sound event detection by using capsule neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 310–322, 2019.

[13] Y. Liu, J. Tang, Y. Song, and L. Dai, "A capsule based approach for polyphonic sound event detection," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1853–1857.

[14] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 771–775.

[15] 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.

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

[17] 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, 1989.

[18] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 1. IEEE, 2007, pp. I–121.

[19] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2386–2390.

[20] Z.-M. Liu, C. Zhang, and S. Y. Philip, "Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 12, pp. 7315–7327, 2018.

[21] 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)*. IEEE, 2018, pp. 1462–1466.

[22] M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto, "Sound event localization based on sound intensity vector refined by dnn-based denoising and source separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 651–655.

[23] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[24] 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, 2018.

[25] 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," *arXiv preprint arXiv:1905.00268*, 2019.

[26] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," *arXiv preprint arXiv:1908.00766*, 2019.

[27] http://dcase.community/challenge2020/task-sound-event-localization-and-detection.

[28] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 333–337.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[31] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[32] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv preprint arXiv:2006.01919*, 2020.