SOUND EVENT DETECTION AND DIRECTION OF ARRIVAL ESTIMATION USING RESIDUAL NET AND RECURRENT NEURAL NETWORKS

Rishabh Ranjan[1], Sathish s/o Jayabalan[1], Thi Ngoc Tho Nguyen[1], Woon-Seng Gan[1]

[1] Nanyang Technological University, Singapore, 679798
{rishabh001, sathishj, nguyenth003, ewsgan}@ntu.edu.sg

ABSTRACT

This paper presents deep learning approach for sound events detection and localization, which is also a part of detection and classification of acoustic scenes and events (DCASE) challenge 2019 Task 3. Deep residual nets originally used for image classification are adapted and combined with recurrent neural networks (RNN) to estimate the onset-offset of sound events, sound events class, and their direction in a reverberant environment. Additionally, data augmentation and post processing techniques are applied to generalize and improve the system performance on unseen data. Using our best model on validation dataset, sound events detection achieves F1-score of 0.89 and error rate of 0.18, whereas sound source localization task achieves angular error of 8° and 90% frame recall.

Index Terms— Sound events detection, directional of arrival, residual net, recurrent neural networks

1. INTRODUCTION

Sound events localization and detection (SELD) system allows one to have automated annotation of a scene in spatial dimension and can assist stakeholders to make informed decisions. It is an important tool for various applications like identifying critical events like gunshots, accidents, noisy vehicles, mixed reality audio where spatial scene information enhanced the augmented listening, robots that listens just like humans and tracks the sound source of interest, smart homes and surveillance systems [1-5]. The three main objectives of SELD system are namely, (1) first, to detect presence of sound events, (2) second, to classify active sound events as textual labels, and (3) third, to estimate directions of active sound events.

The first key component of the SELD system is sound event detection (SED), which assigns pre-defined labels to the active sound events every frame [6]. In the past, many signal processing and machine learning methods have been extensively applied to the SED problem using supervised classification approach. The most popular methods include, dictionary learning [7], gaussian or and hidden markov model [8-9], non-negative matrix factorization (NMF) [10-11], principal component analysis [12], and deep learning methods like fully connected neural network (FCNN) [13], convolutional neural network (CNN) [14-15], recurrent neural networks (RNN) [16], residual network (ResNet) [17]. Most recently, combination of the CNN, RNN and FCNN networks were also proposed to improve the SED performance and present state-of-art results [18-20]. Furthermore, multi-channel audio inputs as well as ambisonics data has been employed in SED task to exploit the spatial nature of the data [20-21].

The second key component of SELD system is direction of arrival (DoA), which estimates the directions of active sound events in terms of azimuth and/or elevations angles. DoA problem is commonly dealt using various traditional signal processing based methods: time-difference [22], subspace methods such as multiple signal classification (MUSIC) [23], cross-correlation methods such as generalized cross-correlation with phase transform (GCC-PHAT) [24], steered response with phase transform (SRP-PHAT) [25], multichannel cross-correlation coefficient (MCCC) [26]. However, some of the common practical challenges with these methods is performance degradation in presence of noisy and reverberant environment as well as high computational cost. Recently, deep learning based methods is also being extensively employed to improve the DoA performance and outperforms the traditional methods in challenging environments [27-35]. DNN based approaches vary in terms of microphone array geometry- circular, linear, binaural, ambisonics. In addition, different input features like GCC [33], magnitude and phase transform [21] [31], eigen vectors [34], inter-aural cross-correlation features [32] and most recently raw temporal features [35] have been used to improve the DoA performance. Furthermore, most of these works have been shown to work on only azimuthal plane sources and/or single static sources except [31], which demonstrates working in both azimuth and elevation as well as for overlapping sound sources.

There are very few works jointly solving the SELD task using deep learning. Hirvonen [28] used spectral power of the multi-channel audio signals from circular array and used CNN based classifier to predict one of the 8 source directions on azimuthal plane for each sound event. In contrast, Adavanne [21] employed regression based continuous DoA output in both azimuth and elevation for 11 different type of overlapping sound classes. The authors employed a joint network using CRNN network with two branches each for SED and DoA to perform the combined SELD task.

In this paper, we employ a ResNet architecture combined with RNN, referred as ResNet RNN, for the joint estimation of respective labels for SED and DoA for sound events in a reverberant scene with one or two active sound sources. In contrast to the baseline model [21], a classification-based output is employed for DoA and additional post-processing techniques are employed for both SED and DoA to further improve the overall SELD performance. The proposed model significantly outperforms the baseline model [21] using convolutional recurrent neural network (CRNN) specifically for the DoA task. In the next section, we give a detailed description of the proposed methodology and training set up.

https://doi.org/10.33682/93dp-f064

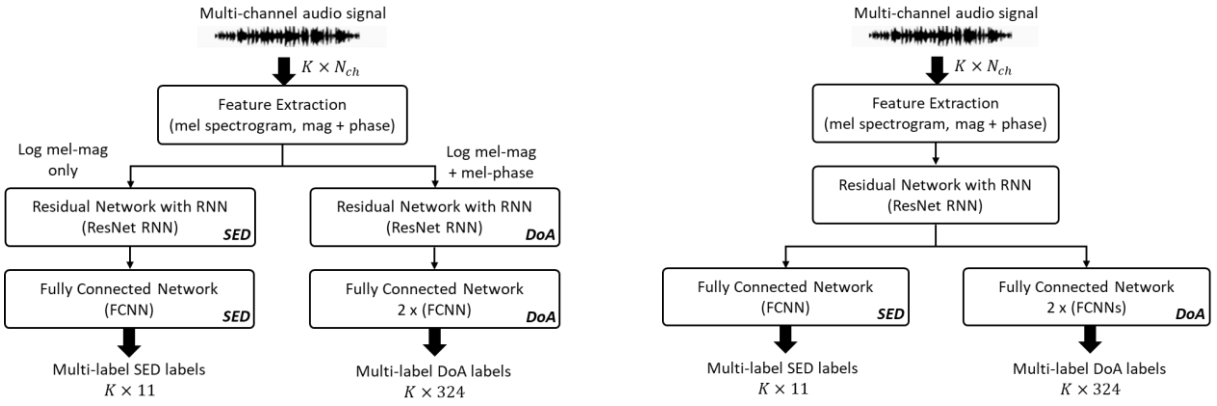

Figure 1: Proposed system overview (a) Individually trained models for SED and DoA (b) Jointly trained model

2. METHODOLOGY

For the SELD task, two different configurations using a modified version of ResNet architecture combined with RNN are employed. Figure 1 shows overview block diagrams of the two system configurations. First system is using individually trained models for SED and DoA, where input is log mel magnitude spectrogram for the SED task, while log mel magnitude and linear mel phase spectrogram for the DoA as shown in Figure 1(a). Second system is using a jointly trained model, where ResNet RNN architecture is common for both SED and DoA and subsequently, divided into two branches using FCNN layers as shown in Figure 1(b). One key advantage of joint model in this work is that they share the common resources of ResNet RNN and therefore, would need less computational resource when implementing on embedded devices. DoA branch for both the configurations is further divided into two parallel branches consisting of FCNN layers and the two network outputs are combined as post-processing step to enhance the DoA accuracy. For both the systems, SED and DoA is predicted as continuous output in range [0 1] as probabilities for 11 distinct sound events and 324 unique directions, respectively. In the next subsections, we explain the dataset, feature extraction, model architecture, training set up, data augmentations and post-processing techniques used.

2.1. Development Dataset

The development dataset is taken from detection and classification of acoustic scenes and events (DCASE) challenge 2019 task 3 for SELD task [36]. It consists of 4 splits and each split contains 100 audio files of length 60 sec and contains overlapping as well as non-overlapping sound events. Audio files is synthesized using 11 isolated sound labels taken from [37] and convolved with impulse responses (IR) measured from 5 different rooms at 504 unique combinations of azimuth-elevation-distance and finally, mixed with natural ambient noise collected at IR recording locations. In terms of unique target directions, there are 36 azimuths and 9 elevations resulting in total 324 directions. All the IRs were recorded using Eigenmike [38], a 32 microphone spherical array with only 4 of the microphones forming a tetrahedral shape were used for synthesis of DCASE 2019 task 3 dataset.

2.2. Feature Extraction

Each of the audio file is sampled at 48kHz and short-time Fourier transform (STFT) is applied with hop size of 20 msec. Next, STFT

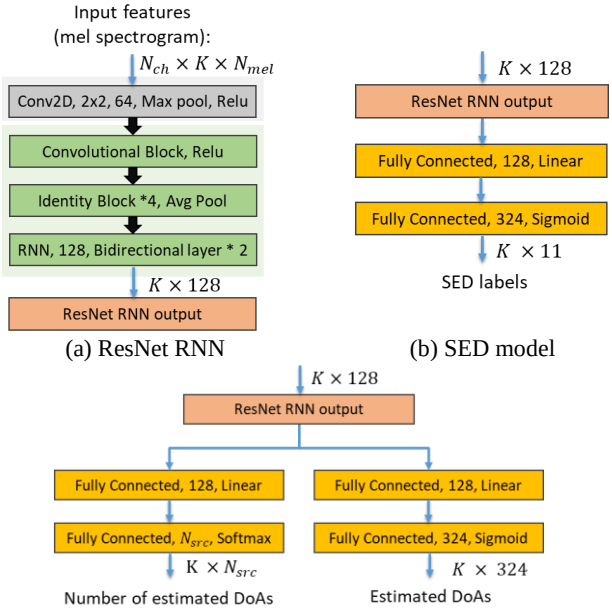

Figure 2: Model architectures (a) ResNet RNN (b) SED: FCNN (c) DoA: Two parallel FCNN branch

spectrogram is converted to log mel magnitude spectrogram from amplitude of STFT and linear mel phase spectrogram from phase component of STFT using dot product of STFT component and mel-filter banks. After converting into mel spectrogram features, low and high frequency components are removed and finally, resized to match the input shape of the neural network before training.

2.3. Model Architecture

Figure 2(a) shows the architecture of proposed modified ResNet combined with RNN. The ResNet model is adapted from residual net model originally designed for image recognition and described in [39]. As shown in the figure, output of the feature extraction is fed to the ResNet RNN model with feature dimension of $N_{ch} \times K \times N_{mel}$, where N_{ch} is the number of channels (= 4 when only magnitude is used and 8 when both magnitude and phase is used as input feature), K is the number of frames used as

sequence, and N_{mel} is the number of mel filter banks. The ResNet architecture consists of many 2D convolutional (conv) layers, however the distinct feature of ResNet architecture is the use of identity and convolutional block with skip connection to solve the vanishing gradient problem in deeper networks [39]. In this work, ResNet is a 2-stage architecture with first stage being a 2D conv layer with 64 filters, followed by batch normalization of outputs [40], 'ReLU' activation function and dimensionality reduction using max pooling of 2 along the mel frequency axis. Second stage consists of one convolutional block with three filters with output size as (64, 64, 256), and 4 identity blocks with three filters and same number of filters as in convolutional block. Finally, average pooling of size 16 is applied along the mel frequency axis. Subsequently, output from stage 2 is reshaped on the last two dimensions before feeding to two RNN layers to learn the contextual information from temporal sequence data of K frames. Each RNN layer consists of 128 nodes of either gated recurrent units (GRU) or long short-term memory (LSTM) with 'tanh' activation function. RNN block is followed by fully connected dense layers for both SED and DoA as shown in Figure 2(b) and (c). First FC layer in both the tasks consists of 128 nodes with linear activation function and dropout of 0.5 to improve the generalization ability of network. Final FC layer in SED consists of 11 nodes corresponding to 11 unique target sound classes with sigmoid activation function as shown in Figure 2(b). DoA, however, consists of two parallel branches of FC layers with one branch estimating number of active sources and other branch estimating actual direction estimates as probabilities. Final FC layer in first branch consists of N_{src} nodes corresponding to maximum number of active sources with 'softmax' activation function. For the second DoA branch, final FC layer consists of 324 nodes corresponding to 324 unique directions with 'sigmoid' activation function.

2.4. Model Training

For model development, 4 cross-fold sets from DCASE challenge 2019 task 3 dataset [4] is used with 3 of the splits used for training and one split for validation as shown in Table 1. During training, each processed audio feature file is split into sequence length of 128 frames and resized with fixed batch size of 96. For SED, binary cross-entropy loss function is used for model weights adaptation. For DoA second branch, weighted binary cross-entropy loss function is used to strongly penalize the false negatives because at most only two out of 324 DoA labels are true at any time frame in the ground truth. For both SED and DoA, adam optimizer is used with learning rate of 0.0005. Best model is saved using the combined SELD loss metric computed using the evaluation metrics provided by DCASE task 3 organizers and briefly explained in sub-section 2.7.

2.5. Data Augmentation

To improve model generalization capability on unseen test data, data augmentation using frame shifting is applied to each of the processed audio file. Each audio feature set is shifted in negative time by 32, 64 and 96 frames across temporal dimension before splitting into sequence of 128 frames. In this way, we create 3 shifted copies of audio segments, which helps in generalizing the model performance. Therefore, total data after augmentation is 4 times larger than the original dataset size and each audio feature file including shifted copies are selected randomly for training in each epoch.

Table 1: Cross-fold configuration for model evaluation

Fold	Training sets	Validation sets
1	Split 2, 3, 4	Split 1
2	Split 3, 4, 1	Split 2
3	Split 1, 2, 4	Split 3
4	Split 1, 2, 3	Split 4

2.6. Output Post-processing

First post-processing technique applied to both SED and DoA outputs is by predicting on frame shifted audio feature sequences and then, taking geometric mean of the shifted probability estimates:

$$\mathbf{p}_{avg} = \sqrt[3]{\mathbf{p}(t_0) \cdot \mathbf{p}(t_1) \cdot \mathbf{p}(t_2)}\,. \quad (1)$$

where $\mathbf{p}(t_i)$ is the probabilities predicted using the final trained model weights for each audio feature file $\mathbf{X}(t)$ shifted by t_i frames and padding zeros in front and excluding first t_i frames from the predicted probabilities as final estimates:

$$\mathbf{p}(t_i) = predict([\mathbf{0}(t_i) \quad \mathbf{X}(t - t_i)])\mathbf{.} \quad (2)$$

Above averaging method helps in averaging out the spurious outliers in the final prediction. It is also found that geometric mean gives slightly better results than arithmetic mean and thus, were used to compute SED and DoA output probabilities.

Final SED labels were obtained frame wise by comparing the output probabilities for each label with a given threshold. Those labels with probabilities more than the threshold are selected as active sound events and in the case of none of the labels' probabilities more than threshold no activity, i.e., ambience is assigned. For DoA estimations, we merge the outputs of two branches as explained in following sub subsection.

2.6.1. DoA Post-processing

As explained earlier in sub-section 2.3 and Figure 2(c), there are two outputs from DoA model as number of active sources and 324 direction labels probabilities. To obtain final estimated directions per frame, we take the following steps:

1. Convert the DoA output 324 probabilities estimate into 2D array with size 36 azimuths × 9 elevations
2. Find the local peaks in the 2D array above a given threshold and a minimum neighboring distance between two peaks
3. Compute n_{src} as number of active sources by selecting label with maximum probability in the first DoA branch.
4. Select n_{src} peaks from the output of second step as final DoA estimate.

By using above post-processing steps of peak finding with minimum neighboring constraint, we filter out the redundant DoA peaks which are close by and also improve the DoA frame recall by capping the number estimated DoAs based on first branch output. Finally, both SED and DoA outputs are combined together frame wise based on the presence of active sound events or directions in any of the SED or DoA outputs. In the case of multiple sources, to match the DoA and SED outputs, we take into account the precedence of single source SED and DoA outputs in previous time frames and use this prior information to match the second source outputs in current time frame.

Table 2: Proposed ResNet RNN model Vs Baseline CRNN model performance fold-wise for training on 3-splits

Fold	Model	ER	F-Score	DoA Error (°)	FR (%)
1	**Proposed-I**	**0.1640**	**89.83**	**8.72**	**91.36**
	Proposed-J	0.2479	85.10	12.08	89.77
	Baseline	0.3055	82.96	30.13	85.42
2	**Proposed-I**	**0.1903**	**89.13**	**8.55**	**90.0**
	Proposed-J	0.2716	84.32	12.58	86.57
	Baseline	0.3273	82.17	31.32	82.44
3	**Proposed-I**	**0.1611**	**90.71**	**7.76**	**91.28**
	Proposed-J	0.2543	85.33	13.35	89.62
	Baseline	0.2676	84.93	31.75	86.12
4	**Proposed-I**	**0.2143**	**86.77**	**7.72**	**90.7**
	Proposed-J	0.3084	82.03	12.09	88.47
	Baseline	0.2937	82.64	31.05	87.23
over-all	**Proposed-I**	**0.1824**	**89.10**	**8.19**	**90.84**
	Proposed-J	0.2706	84.18	12.53	88.61
	Baseline	0.2986	83.16	31.06	85.30

2.7. Evaluation Metrics

Model performance is evaluated using 4 metrics, 2 each for SED and DoA. SED is evaluated using error rate (ER) and F-score. ER is the total error based on total number of insertions (I), deletions and substitutions [41]. F-score is calculated as harmonic mean of precision and recall [41]. DoA is evaluated using average angular error and frame recall (FR). DoA error is defined as average angular error in degrees between estimated and ground truth directions and computed using Hungarian algorithm [42] to account for the assignment problem of matching the individual estimated direction with respective reference direction. DoA FR is defined as percentage of frames where number of estimated and reference directions are equal out of total frames. In addition, combined SED, DoA and SELD metrics were computed using mean of respective error metrics and used for evaluating models.

3. RESULTS

Table 2 shows the performance of two proposed models: individually trained models (Proposed-I) and jointly trained models (Proposed-J). Clearly, the Proposed-I models outperforms the baseline model in terms of all the 4 metrics and for all validation splits. Specifically, there is significant overall improvement in terms of DoA angular error from 31° for baseline to 8.2° for the individually trained models. However, jointly trained models do not perform as good as the Proposed-I models but yet provides noticeable improvement over baseline, especially for DoA. Poor performance for Proposed-J model can be explained by the fact that by using shared ResNet layers' trained weights may not be optimal for either SED and DoA because of joint training. On the other hand, for individual models, respective weights for both SED and DoA ResNet layers are optimally trained and thus, giving better performance. Additionally, joint model incurred around 1.4 million parameters against 3 million parameters for combined individual SED and DoA model. Clearly as mentioned earlier, joint models require less computational resource and therefore, would ensure faster prediction time as compared to individual models for an SELD system running in real-time on an embedded device.

Table 3: Proposed ResNet RNN model Vs Baseline CRNN model performance for Ov1 and Ov2

Fold	Model	ER	F-Score	DoA Error (°)	FR (%)
Ov1	Proposed-I	**0.1571**	**91.26**	**3.90**	**97.07**
	Proposed-J	0.2077	88.86	6.6	93.96
	Baseline	0.2834	85.32	26.41	93.23
Ov2	Proposed-I	**0.1954**	**87.93**	**10.49**	**84.60**
	Proposed-J	0.3029	81.64	15.59	83.25
	Baseline	0.3064	82.00	33.70	77.38

Table 4: Proposed-I model performance for 5 RIRs

IR	ER	F-Score	DoA Error (°)	FR (%)
IR1	0.1728	89.49	8.31	90.46
IR2	0.1940	88.57	7.85	92.06
IR3	0.1737	89.83	7.69	90.22
IR4	0.1788	89.33	8.36	91.49
IR5	0.1937	88.24	8.73	89.95

Table 3 shows the proposed models performance for single source (Ov1) and two overlapping sources (Ov2). Proposed models performs much better for single source scenario as compared to two sources, especially with DoA error as low as 3.9° and FR as high as 97 %. Proposed model performance for 5 different room impulse responses is also summarized in Table 4. Except for the IR5 and IR3 in terms of SED ER, proposed models perform similar across all the IRs.

4. CONCLUSION

In this paper, a 2-stage ResNet architecture combined with RNN is used for both sound events classification and localization task. With data augmentation and post-processing techniques, the proposed model performance is significantly improved, especially for the DoA task with error as low as 8º and frame recall of 90 %. The proposed work is also demonstrated in DCASE challenge 2019 Task 3 and showed superior performance over baseline on evaluation dataset. Jointly trained model is useful for edge implementations because of lower complexity but at the cost of sub-optimal performance. This needs to be further investigated and has been identified as future work to further improve the performance of joint model.

5. ACKNOWLEDGMENT

This research was conducted in collaboration with Singapore Telecommunications Limited and supported by the Singapore Government through the Industry Alignment Fund - Industry Collaboration Projects Grant.

The Authors are also thankful to Maggie Leong at amazon web services Singapore to generously provide the resources for model development on the cloud.

6. REFERENCES

[1] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," in *Journal of Robotics and Mechatronics*, vol. 29, no. 1, 2017

[2] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Worksh. on Apps. of Signal*

Processing to Audio and Acoustics (WASPAA'05), New Paltz, NY, USA, Oct. 2005, pp. 158–161.

[3] C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of low-cost urban acoustic monitoring devices," *Applied Acoustics*, vol. In Press, 2016.

[4] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *International Conference on Robotics and Automation (ICRA)*, 2018.

[5] C. Grobler, C. Kruger, B. Silva, and G. Hancke, "Sound based localization and identification in industrial environments," in *IEEE Industrial Electronics Society (IECON)*, 2017.

[6] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *J. Acoust. Soc. America*, vol. 122, no. 2, pp. 881–891, 2007.

[7] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 171–175.

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

[9] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[10] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classification," M.S. thesis, ATIAM, Paris Tech, Paris, France, Aug. 2011

[11] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, et al., "An exemplar-based nmf approach to audio event detection," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[12] E. Benetos, "Automatic transcription of polyphonic music exploiting temporal evolution," Ph.D. dissertation, School of Electron. Eng. And Comput. Sci., Queen Mary University of London, London, U.K., Dec. 2012.

[13] E. Cakır, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi-label deep neural networks," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2015.

[14] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," in *INTERSPEECH*, 2016.

[15] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[16] 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.

[17] S. Hershey, C. Sourish, E. P. W. Daniel, G. F. Jort, J. Aren, M. R. Channing, P. Manoj. "CNN architectures for large-scale audio classification." In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131-135. IEEE, 2017.

[18] S. Adavanne, A. Politis, and T. Virtanen, "Multichannel sound event detection using 3D convolutional neural networks for learning inte channel features," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2018.

[19] E. C¸ akır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, 2017.

[20] S. Adavanne, P. Pertila, and T. Virtanen, "Sound event detection using ¨spatial features and convolutional recurrent neural network," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

[21] S. Adavanne, P. Pertila, N. Joonas and T. Virtanen. "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks." *IEEE Journal of Selected Topics in Signal Processing* (2018).

[22] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," in *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, 2001.

[23] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, 1986.

[24] C. Knapp, and C. Glifford. "The generalized correlation method for estimation of time delay." *IEEE transactions on acoustics, speech, and signal processing* 24, no. 4 (1976): 320-327.

[25] DiBiase, Joseph Hector. *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. PhD thesis, Brown University, 2000.

[26] J. Benesty, J.D. Chen, and Y.T.Huang, "Time delay estimation via linear interpolation and cross correlation," *IEEE Transactions on speech and audio processing*, vol. 12, no. 5, pp. 509–519, 2004.

[27] Q. Li,Z. Xueliang, and L. Hao. "Online Direction of Arrival Estimation Based on Deep Learning." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2616-2620. IEEE, 2018.

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

[29] C. Pang, L. Hong, and L. Xiaofei. "Multitask Learning of Time-Frequency CNN for Sound Source Localization." *IEEE Access* 7 (2019): 40725-40737.

[30] S. Chakrabarty, and H. AP. Emanuël. "Broadband DOA estimation using convolutional neural networks trained with noise signals." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 136-140. IEEE, 2017.

[31] 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)*, pp. 1462-1466. IEEE, 2018.

[32] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," in International Journal of Applied Engineering Research, vol. 12, no. 22, 2017.

[33] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[34] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[35] J. Vera-Diaz, P. Daniel, and M. G. Javier. "Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates." *Sensors* 18, no. 10 (2018): 3418.

[36] DCASE Challenge Task 3: http://dcase.community/challenge2019/task-sound-event-localization-and-detection

[37] TUT audio dataset: http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-synthetic-audio#audio-dataset

[38] Eigenmike: https://mhacoustics.com/products

[39] K. He, Z. Xiangyu, R. Shaoqing, and S. Jian. "*Deep residual learning for image recognition.*" In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning*, 2015.

[41] A. Mesaros, T. Heittola, and T. Virtanen. "*Metrics for polyphonic sound event detection.*" Applied Sciences 6, no. 6 (2016): 162.

[42] H. W. Kuhn, "The hungarian method for the assignment problem," *in Naval Research Logistics Quarterly*, no. 2, 1955, p. 8397.