Improving audio event localization accuracy via derivative prediction

Ruchi Pandey
SPCRC Lab
IIIT Hyderabad, India
ruchi.pandey@research.iiit.ac.in

Shreyas Jaiswal
SPCRC Lab
IIIT Hyderabad, India
shreyas.jaiswal@research.iiit.ac.in

Huy Phan[◇]
Amazon Alexa
Cambridge, MA, USA
huypq@amazon.co.uk

Santosh Nannuru
SPCRC Lab
IIIT Hyderabad, India
santosh.nannuru@iiit.ac.in

***Abstract*—Accurate localization of sound sources is essential in many acoustic sensing and monitoring applications. In the absence of temporal continuity models, many methods produce unrealistic direction of arrival (DOA) estimates involving sudden changes. To address this, we propose an approach that trains a neural network to predict DOA derivatives in Cartesian coordinates (x', y', z'), which capture the rate of change in DOA (x, y, z) over time. By combining the predicted DOAs with the predicted derivatives, our method can suppress sudden DOA changes and generate smooth motion trajectories. We introduce an update rule that combines the predicted DOAs with the predicted derivatives to obtain the final DOAs. We validate our approach using the TAU-NIGENS Spatial Sound Events (TNSSE) 2021 dataset. Our results demonstrate that incorporating DOA derivatives improves the accuracy of DOA estimation, particularly in low signal-to-noise ratio scenarios.**

***Index Terms*—Deep learning, Microphone array, SALSA-Lite, Sound event localization and detection (SELD).**

I. Introduction

Sound source localization is an integral part of many modern applications such as video conferencing, hearing aids, and human-robot interactions [1], [2]. There are a plethora of localization methods existing in the literature [3]–[7], [9]. Recently data-driven methods have shown promising results for sound source localization in reverberant and low signal-to-noise ratio (SNR) scenarios [10]–[13]. The deep neural network (DNN) based methods were shown to outperform parametric methods in terms of high resolution, and low erroneous DOAs [13]–[18].

In the recent past, polyphonic sound event localization and detection (SELD) problems have garnered a lot of attention among researchers, which combine the detection and localization tasks and have many practical applications [12], [15], [19]. Since its introduction, various model architectures and features have been proposed [20]–[24]. In real-world scenarios, sudden large changes in the direction of arrival (DOA) are unexpected. To capture this, we train our network to learn and predict DOA derivatives to maintain temporal continuity. DOA derivatives provide information about the rate of change in the x, y, z positions. Thus, the combined DOA and DOA derivative prediction provides a mechanism to suppress sudden DOA changes (if predicted by the network) and estimate realistic, smooth motion trajectories. In this study, we focus on improving the localization accuracy of an existing model [24] by predicting both the DOAs and their derivatives, i.e., changes in the x, y, z positions over time. We compare the existing localization models (considering the detection ground truth to be known), which predicts only DOAs, with the model, which predicts both DOAs and their derivatives.

Our experiments reveal that localization can be a challenging task, even for immobile sound sources. This research aims to better understand deep learning-based models for localization tasks with a special focus on combining DOAs with their derivatives to improve the source trajectories. We hope the analysis will direct the research focus towards the localization aspect of the SELD task. Our proposed network builds on the CRNN network [23], [24] by introducing a derivative prediction arm. The focus of this paper is to show the effectiveness of incorporating the derivative information to obtain smoother trajectories.

The specific contributions of this paper are:
(a) we derive an update rule incorporating predicted DOAs and derivatives to improve localization accuracy;
(b) we conduct experimental validation using TAU-NIGENS Spatial Sound Events 2021 dataset [25];
(c) we perform SNR analysis by synthetically adding noise;
(d) we perform analysis of the proposed model when using the pre-trained model for initialization.

II. Model architecture

A. Features

The SALSA-Lite was introduced as an efficient computational version of the Spatial Cue-Augmented Log-Spectrogram (SALSA) feature for MIC (audio format) data [23], [24]. For M-channel audio recording, SALSA-Lite is a $(2M-1)$ channel feature consisting M log-power spectrogram with $(M-1)$ frequency-normalized interchannel phase differences (NIPDs). The NIPD (Λ) approximating the relative distance of arrival (RDOA) can be written as

$$\Lambda(t,f) \approx -\frac{c}{2\pi f} \arg\, |\boldsymbol{H}_1^*(t,f)\boldsymbol{H}_{2:M}(t,f)|, \quad (1)$$

$$\approx [d_{12}(t) \dots d_{1M}(t)], \quad (2)$$

where $\boldsymbol{H}_m(t,f) = e^{\frac{j2\pi f d_{1m}(t)}{c}}$ is the array response for any arbitrary array structure under the far-field assumption

[◇]The work was done when H. Phan was at Centre for Digital Music, Queen Mary University of London, UK and prior to joining Amazon.

and $d_{1m}(t)$ is the RDOA between the first (reference) and m^{th} mic. The SALSA-Lite provides the exact time-frequency positioning between the spectrogram and the NIPD resulting in the model being able to localize multiple overlapping sources.

B. Architecture

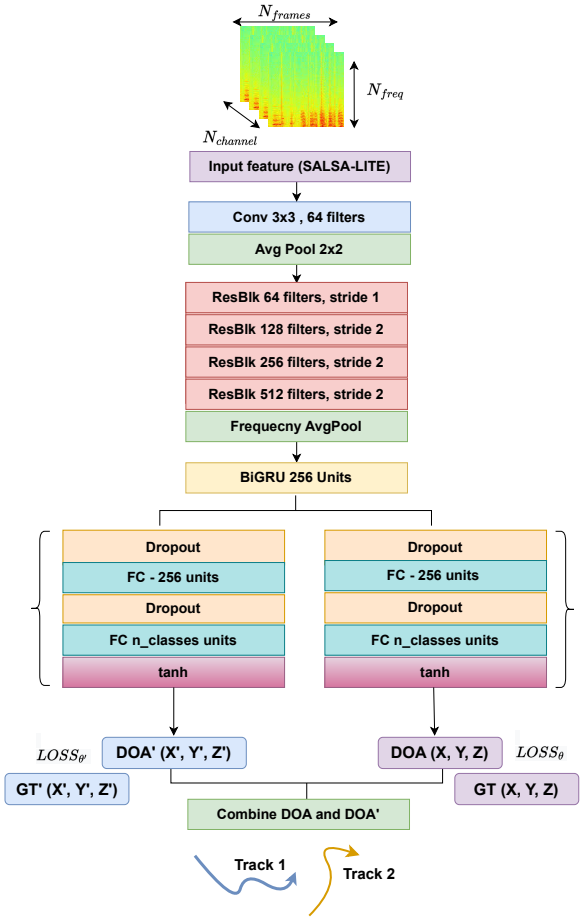

Fig. 1: Model architecture predicting both DOAs and DOA derivatives

Figure 1 shows the neural network architecture designed to predict the DOAs and their derivatives simultaneously. Here derivative represents the change in DOAs between two time frames. The SALSA-lite is fed to the CRNN network, which consists of one convolutional layer, one average pooling layer followed by four ResNet22 blocks [26] in the network body [23], [24].

The output of ResNet block is fed into a two-layer bidirectional Gated Recurrent Unit (GRU) followed by two distinct regression heads for predicting DOAs and their derivatives in Cartesian coordinates ($x, y,$ and z), respectively. Unlike the SELD network in [24], we focus only on the localization task and replace the detection head with the regression head, which predicts the derivatives of DOAs (x', y', z') at different timestamps as shown in Fig. 1. Along with predicting the intermediate DOAs, the additional derivative information helps to obtain better overall DOA estimates. Once the network predicts the DOAs and their derivatives, the final DOAs are obtained using the following update equation

$$\hat{x}_n^{\text{final}} = \frac{\hat{x}_n + (\hat{x}_{n-1} + \hat{x}'_n)}{2}, \quad n = 0, 1, \ldots, N-1 \quad (3)$$

where $\hat{x}'_n$ is prediction of the DOA derivative in x position (i.e., $x_n - x_{n-1}$) at n^{th} time and $\hat{x}'_0 = 0$ is the first derivative assumed to be zero. Similarly, the update rule for $\hat{y}_n^{\text{final}}$ and $\hat{z}_n^{\text{final}}$ can be obtained. By additionally incorporating derivative predictions, we expect the DOA of moving targets to be estimated more accurately. The number of active sources is assumed to be known for both the regression heads (predicting DOAs and derivatives). The ground truth is used to compute the losses for both prediction heads. The mean squared error (MSE) loss is minimized while training both the network heads and can be written as

$$\text{LOSS}_x = w_1 \sum_{n=1}^{N} (x_n - \hat{x_n})^2 + w_2 \sum_{n=1}^{N} (x'_n - \hat{x_n}')^2 \quad (4)$$

$$\text{LOSS}_{\text{total}} = \text{LOSS}_x + \text{LOSS}_y + \text{LOSS}_z \quad (5)$$

where x_n and x'_n are ground truth DOAs and their derivatives at n^{th} timestamps, and $\hat{x_n}$ and $\hat{x_n}'$ represent the predicted DOAs and the predicted derivatives at n^{th} timestamps respectively. In (4), LOSS_x represents loss computed for x positions and similarly LOSS_y and LOSS_z can also be computed. The total loss minimized by the network is given in (5). Note that using appropriate weights for DOA loss and its derivatives loss is important; we use equal weights for both DOA and derivatives loss ($w_1 = w_2 = 0.5$). However the loss weights (w_1 and w_2) can be automatically optimized as described in [27] and can be incorporated in (4).

III. Simulations and Results

A. Dataset and training

The TAU-NIGENS Spatial Sound Events 2021 dataset has been employed in this research paper to analyze the performance of proposed models. The dataset comprises 600 recordings, each one minute long and with four channels, and is in the MIC data format with a sampling frequency of 24 kHz. The dataset includes a diverse range of sound events featuring both stationary and mobile sources from 12 distinct classes. For this study, 400 recordings are used for training, while 100 recordings are allocated to both validation and testing. The angular range for azimuth and elevation angles are $[-180,\ 180)^\circ$ and $[-45,\ 45]^\circ$, respectively.

For feature extraction, we followed the same setup as in [24], and frequencies from 50Hz to 2kHz are processed to avoid the aliasing. While training, Adam optimizer [28] is used, with the initial learning rate as $3\text{x}10^{-4}$ which linearly decreases to 10^{-4} over last 15 epochs. A total of 70 epochs with 32 batch size is processed. The validation set is used for model selection, whereas the test data is used for the performance analysis.

The DCASE data provides ground truth DOA labels at every 0.1 second. While generating the ground truth for DOA

derivatives, the derivatives are considered as the difference between the previous and current DOA. As a significant gap in DOA update time would result in unreliable derivative estimate, the derivative is consider as zero if the source appears for the first time or a missing source reappears after more than 20 frames ($\approx$ 2 sec). We believe this is a reasonable choice since human motion patterns do not change abruptly within short time spans.

B. Accuracy metric

Following the framework of our baseline method [24], we use the detection ground truth to compute the error only for the frames where the sources are present/active. In this study, DOA/spatial error $\Delta\sigma$ is computed as the angular distance between the predicted and true positions, [15], [29].

$$\Delta\sigma = \arccos\left(\mathbf{n}_{\text{true}}.\mathbf{n}_{\text{pred}}\right).\frac{180}{\pi}, \tag{6}$$

where $\mathbf{n}_{\text{true}}$ and $\mathbf{n}_{\text{pred}}$ are unit norm vectors corresponding to the true and predicted positions, respectively. In applications requiring indoor audio source localization, knowing the angular distance of a source is more important than its Euclidean distance [32], [33].

Following the DCASE challenge convention [30], a source is consider as localized only when the DOA error (averaged across the active frames within a block) is less than 20°; we call these cases as true positive (TP).The false negative (FN) counts the number of incidences when the averaged spatial distance is more than 20°. The evaluation criteria and threshold of 20° is used from the original DCASE challenge TP criteria [12], [15], [29], [30]. We report the probability of detection (P_{d} in %) as the fraction of frames where the distance between the predicted and true positions is less than 20°. Since the network is designed to provide only one prediction per class, the error was not calculated for the cases where multiple sources were present from the same class in the frame.

C. Effect of combining derivative

This subsection demonstrates the effect of combining the predicted derivatives with predicted DOAs. Fig. 2 shows the predicted source trajectories from both Model1 (with derivative estimation) and Model2 (without derivative estimation) along with class-wise mean absolute error (MAE) computed for the correctly detected sources for one of the recordings. The cross $(\times)$ in MAE plot denotes the cases when the source has not been detected by the models. It can be seen that Model1 detects more sources, hence resulting in higher MAE for some classes compared to Model2. The final DOA exhibits a smoother trajectory since the outliers are eliminated by combining the derivatives via the proposed update rule (3). The update rule is helpful even for static sources. We observed that Model2 gives more erroneous DOAs for static sources than Model1. Overall, Model1's estimates are closer to true trajectories resulting in higher P_{d}. For this recording, the average P_{d} for static and moving sources for Model1 and Model2 are, $P_{\text{ds}} = 64\%$, $P_{\text{dm}} = 78\%$, $P_{\text{ds}} = 53\%$, and $P_{\text{dm}} = 52.4\%$, respectively. The total P_{d} averaged over 100 recordings from the test data is reported in Table I.

SNR	Model	TP_{s}	TP_{m}	FN_{s}	FN_{m}	P_{ds}	P_{dm}
Clean	Model1	28843	21523	13056	9892	68.8	68.5
	Model2	27637	21389	14262	10026	65.9	68
-2dB	Model1	17169	13097	24730	18318	40.9	41.7
	Model2	17199	12487	24700	18928	41	39.7
-5dB	Model1	15812	10919	26087	20496	37.7	34.7
	Model2	14136	10522	27763	20893	33.7	33.4

TABLE I: Performance of Model1 and Model2 at different SNR (averaged over test data).

SNR	Model	TP_{s}	TP_{m}	FN_{s}	FN_{m}	P_{ds}	P_{dm}
Clean	Model1	33033	24687	8866	6728	78.8	78.5
	Model2	33431	24531	8468	6884	79.7	78
-2dB	Model1	18349	12249	23550	19166	43.7	39
	Model2	17777	11906	24122	19509	42.4	37.8
-5dB	Model1	15365	10060	26534	21355	36.7	32
	Model2	15404	9816	26495	21599	36.7	31.2

TABLE II: Performance of Model1 and Model2 using the pretrained CRNN SELD model at different SNR (averaged over test data).

From Table I, it is evident that both Model1 and Model2 show similar performance for the clean dataset. We observed that Model1's performance degrades when the network predicts the erroneous DOAs; hence combining them with equal weights leads to incorrect estimates. As a correction step, the current DOA prediction with derivative and the past prediction can be weighted depending on the threshold. A choice must be made depending on confidence in the present and past predictions.

D. Effect of transfer learning

To speed up the training process and reduce the risk of overfitting, we repeat the experiments using the pre-trained CRNN weights from an existing SELD model using SALSA-Lite, where the best model is obtained at the 47^{th} epoch [24]. The dataset, architecture, and framework for the pre-trained model detailed in [24] are same as the CRNN body used in this paper. Keeping the CRNN body's weights fixed using the pre-trained SELD model increases the overall P_{d} is increased by 10 % for both Model1 and Model2, as shown in Table II. From Fig. 3, it can be seen that Model1 outperforms Model2 with higher P_{d} and lower DOA error.

E. Effect of low SNR levels

In order to assess the robustness of the models, we introduced synthetic additive white Gaussian noise to the recordings from TAU-NIGENS Spatial Sound Events 2021 dataset despite the presence of unknown interference and noise in the dataset. The results from Table I indicate a significant degradation in the P_{d} of both models as the SNR level decreases. Nevertheless, Model1 exhibits superior performance in noisy scenarios due to the improved final DOAs resulting from estimated derivatives. The impact of SNR level on the source trajectories obtained from both models is presented in Fig. 4. Our analysis suggests that Model1 provides more reliable estimates than Model2.

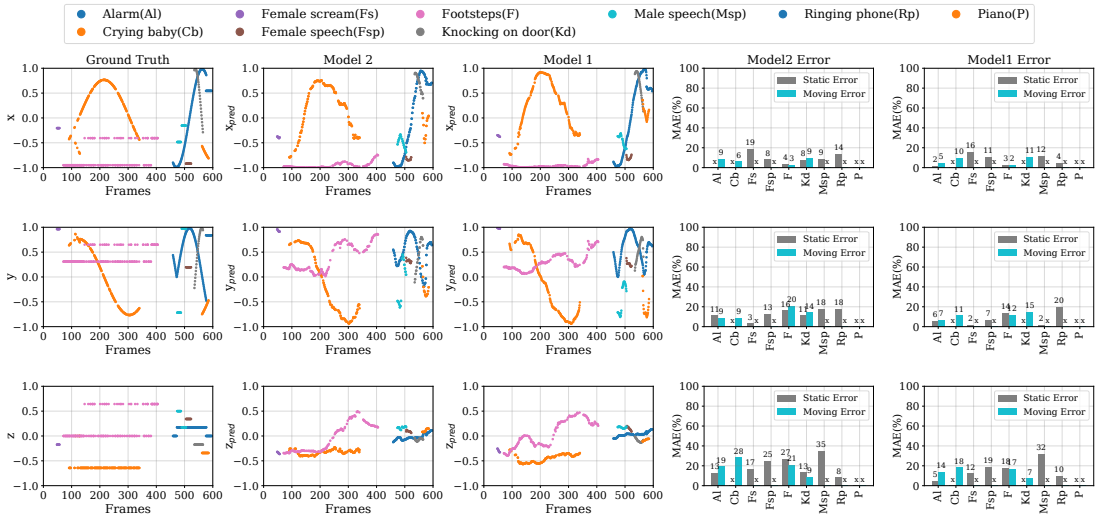

Fig. 2: Effect of derivatives: true and predicted trajectories from Model1 and Model2 with classwise MAE.

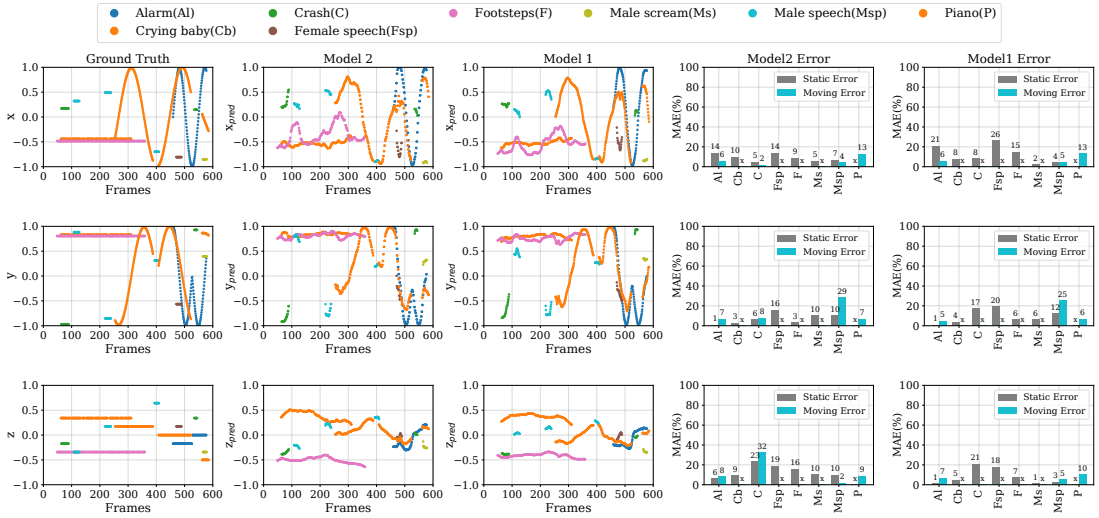

Fig. 3: Effect of transfer learning: true and predicted trajectories from Model1 and Model2 with classwise MAE.

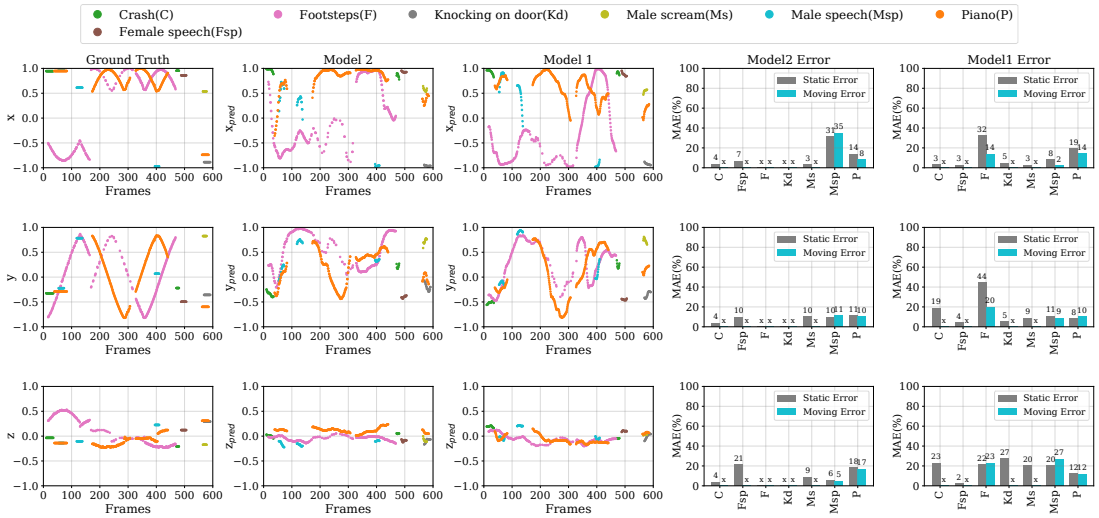

Fig. 4: Effect of low SNR levels: true and predicted trajectories from Model1 and Model2 with classwise MAE.

IV. CONCLUSION AND FUTURE WORK

This study highlights the significance of predicting DOA derivatives in conjunction with DOAs (Model1) for enhancing the overall localization performance compared to solely predicting DOAs (Model2). Furthermore, we demonstrate that Model1 is resilient to noise and performs better than Model2 under low SNR conditions.

Given the broad range of applications of SELD tasks, our analysis reveals that estimating DOAs and their derivatives cumulatively improve the source trajectories and overall performance. In the future, it would be intriguing to investigate the potential impact of incorporating higher-order derivatives in SELD tasks where detection and localization are simultaneously performed.

REFERENCES

[1] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Maximum likelihood approach to "informed" sound source localization for hearing aid applications," in IEEE Inter. Conf. Acous., Spe., Sig. Proces. IEEE, 2015, pp. 16–20.

[2] L. Wan, G. Han, L. Shu, S. Chan, and T. Zhu, "The application of DOA estimation approach in patient tracking systems with high patient density," in IEEE Trans. Ind. Electron., vol. 12, no. 6, pp. 2353–2364, 2016.

[3] H. L. Van Trees, "Optimum Array Processing (Detection, Estimation, and Modulation Theory, Part IV)," John Wiley & Sons, 2002.

[4] R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propagation, vol. 34, no. 3, pp. 276–280, 1986.

[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech Sig. Process., vol. 24, no. 4, pp. 320–327, 1976.

[6] A. Xenaki, P. Gerstoft and K. Mosegaard, "Compressive beamforming," J. Acoust. Soc. Am., vol. 136, no. 1, pp. 260–271, 2014.

[7] P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki and S. Nannuru, "Multi-snapshot sparse Bayesian learning for DOA,"IEEE Sig. Process. Lett., vol. 23, no. 10, pp. 1469–1473, 2016.

[8] D. Malioutov, M. Çetin and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," IEEE Trans. Sig. process., vol. 53, no. 8, pp. 3010–3022, 2005.

[9] R. Pandey, S. Nannuru and A. Siripuram, "Sparse Bayesian Learning for Acoustic Source Localization," IEEE Inter. Conf. Acous., Spe., Sig. Proces., pp. 4670–4674, 2021.

[10] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," IEEE/ACM Trans. Aud., Spe., Lang. Proces., vol. 29, pp. 300–311, 2020

[11] R. Opochinsky, G. Chechik, and S. Gannot, "Deep ranking-based DOA tracking algorithm," in 29th European Sig. Process. Conf. (EUSIPCO), pp. 1020–1024, 2021.

[12] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," IEEE J. Sel. Top. Sig. Proces., vol. 13, no. 1, pp. 34–48, 2018.

[13] P. A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," J. Acoust. Soc. Am., vol. 152, no. 1, pp. 107–151, 2022.

[14] S. Chakrabarty and E. A. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in Workshop on App. of Sig. Proces. to Aud. and Acous (WASPAA), pp. 136-140, 2017.

[15] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," arXiv preprint arXiv:1905.08546, 2019.

[16] 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 Inter. Conf. Acous., Spe., Sig. Proces., pp. 2814–2818, 2015.

[17] D. Suvorov, G. Dong, and R. Zhukov, "Deep residual network for sound source localization in the time domain," arXiv preprint arXiv:1808.06429, 2018.

[18] Q. Nguyen, L. Girin, G. Bailly, F. Elisei, and D. C. Nguyen, "Autonomous sensorimotor learning for sound source localization by a humanoid robot, " in Workshop on Crossmodal Learning for Intel. Robot. in conj. with IEEE/RSJ IROS, 2018.

[19] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network, " in Euro. Sig. Proces. Conf. (EUSIPCO), pp. 1462–1466, 2018.

[20] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in IEEE Inter. Conf. Acous., Spe., Sig. Proces., pp. 915–919, 2021.

[21] H. Phan, L. Pham, P. Koch, N. Q. Duong, I. McLoughlin, and A. Mertins, "On multitask loss function for audio event detection and localization," arXiv preprint arXiv:2009.05527, 2020.

[22] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and D. M. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in IEEE Inter. Conf. Acous., Spe., Sig. Proces., pp. 885–889, 2021.

[23] T.N.T. Nguyen, K.N. Watcharasupat, N.K. Nguyen, D.L. Jones, and W.S. Gan, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," IEEE/ACM Trans. Aud., Spe., Lang. Proces., vol. 30, pp. 1749–1762, 2022.

[24] T.N.T. Nguyen, D.L. Jones, K.N. Watcharasupat, H. Phan, and W.S. Gan, "SALSA-Lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in IEEE Inter. Conf. Acous., Spe., Sig. Proces., pp. 716–720, 2022.

[25] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," arXiv preprint arXiv:2106.06999, 2021.

[26] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Trans. on Aud., Spe., and Lang. Proces., vol. 28, pp. 2880–2894, 2020.

[27] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "XSleepNet: Multi-view sequential model for automatic sleep staging," in IEEE Trans. on Patt. Ana. and Mach. Inte., vol. 44, no. 9, pp. 5903–5915, 2021.

[28] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[29] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in IEEE Work. on Appli. of Sig. Proces. to Aud. and Acous. (WASPAA), pp. 333–337, 2019.

[30] https://dcase.community/challenge2020/task-sound-event-localization-and-detection#metrics

[31] 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.

[32] H. W. Löllmann, et al., "The LOCATA challenge data corpus for acoustic source localization and tracking," IEEE Sensor Array Multichannel Sig. Process. Workshop, pp. 410–414, 2018.

[33] R. Pandey, S. Nannuru, and P. Gerstoft, "Experimental Validation of Wideband SBL Models for DOA Estimation," in Euro. Sig. Proces. Conf. (EUSIPCO), pp. 219-–223, 2022.