L3DAS22 CHALLENGE: MACHINE LEARNING FOR 3D AUDIO SIGNAL PROCESSING

Eric Guizzo[†], *Christian Marinoni*[†], *Marco Pennese*[†], *Xinlei Ren*[*],
Xiguang Zheng[*], *Chen Zhang*[*], *Bruno Masiero*[⋆], *and Danilo Comminiello*[†]

[†]Sapienza University of Rome, Italy
[⋆] University of Campinas, Brazil
[*] Kuaishou Technology, China

ABSTRACT

The L3DAS22 Challenge[1] is aimed at encouraging and fostering collaborative research on machine learning for 3D audio signal processing, with particular focus on 3D speech enhancement (SE) and 3D sound localization and detection (SELD). Alongside with the challenge, we release the L3DAS22 dataset, a 90 hours 3D audio corpus, accompanied with a Python API that facilitates the data usage and results submission stage. Usually, machine learning approaches to 3D audio tasks are based on single-perspective Ambisonics recordings or on arrays of single-capsule microphones. We propose, instead, a novel multichannel audio configuration based multiple-source and multiple-perspective Ambisonics recordings, performed with an array of two first-order Ambisonics microphones. To the best of our knowledge, it is the first time that a dual-mic Ambisonics configuration is used for these tasks. We provide baseline models and results for both tasks, obtained with state-of-the-art architectures: a Beamforming U-Net for SE and SELDnet for SELD.

Index Terms— Data Challenge, 3D Audio, Ambisonics, Sound Source Localization, Sound Source Classification, Speech Enhancement

1. INTRODUCTION

3D audio is gaining increasing interest in the machine learning community in recent years. This field of application is incredibly wide and ranges from virtual and real conferencing to game development, music production, autonomous driving, surveillance and many more. Tasks like sound source localization, speech and emotion recognition, sound source separation, speech enhancement and denoising, and acoustic echo cancellation, among others, potentially benefit from tridimensional representations of sound field, as they carry additional spatial information [1, 2]. 3D audio formats permit to obtain an impressive performance in many machine learning-based tasks, usually bringing out a significant improvement over the single/dual-channel formats [2, 3]. In this context, Ambisonics prevails among other 3D audio formats for its simplicity, effectiveness and flexibility.

The L3DAS22 Challenge organized within the L3DAS (Learning 3D Audio Sources) project is designed to encourage and foster research on machine learning for 3D audio signal processing. In particular, we focus on two 3D audio tasks: 3D Speech Enhancement and 3D Sound Event Localization and Detection, both relying on multiple-source and multiple-perspective (MSMP) Ambisonics recordings.

3D SE aims at removing unwanted information from spurious spatial vocal recordings and further enhancing the speech intelligibility and clarity. A widespread strategy to perform SE is to use deep neural networks (DNNs) to estimate a time-frequency mask in the Fourier domain that extracts clean speech signals from noisy spectra [4]. Neural beamforming techniques as Filter and Sum Networks (FaSNet) [5] provide state-of-the art results for Ambisonics-based SE and are usually suitable for low-latency scenarios. Also U-Net-based approaches provide competitive results in this context, both for monaural [6, 7] and multichannel SE tasks [8, 9], at the expense of higher computational power demand. Other techniques to perform SE include recurrent neural networks (RNNs) [10], graph-based spectral subtraction [11], discriminative learning [12], dilated convolutions [13].

3D SELD, instead, aims at obtaining exhaustive spatiotemporal descriptions of 3D acoustic scenes, predicting which sound categories are present in the scene, and when and where each sound instance is active. SELD can be considered as a joining of the traditional sound event detection and sound source localization tasks, and it was presented for the first time in the DCASE2019 Challenge [14]. Also here, the state-of-the-art methods are based on deep learning strategies [15]. SELDnet [16] adopted a convolutional-recurrent design with two distinct branches for localization and detection and it was used as a baseline model in SELD tasks of the DCASE challenges. An improved SELDnet model was then introduced by [17], including temporal convolutions.

Corresponding author's email: danilo.comminiello@uniroma1.it. This work has been supported by "Progetti di Ricerca Grandi" of Sapienza University of Rome under grant number RG11916B88E1942F.

[1] www.l3das.com/icassp2022

Other novel solutions for this task include ensemble models [18], multi-stage training [19] and bespoke augmentation strategies [20].

These tasks are complementary each other and are aimed at fulfilling real-world needs related to real and virtual conferencing. Especially in multi-speaker scenarios, it is in fact very important to properly understand the nature of a sound event and its position within the environment, what is the content of the sound signal and how to leverage it at best for a specific application (e.g., teleconferencing and assistive listening or entertainment, among others).

Alongside with the challenge, we present the L3DAS22 datasets, aimed at solving SE and SELD tasks making use of MSMP Ambisonics files, obtained performing 3D audio recordings with an array of two Ambisonics microphones, as further discussed in the next Section. For the first time, to our best knowledge MSMP dual-mic Ambisonics recordings are considered for machine-learning purposes, giving us the possibility to test the effectiveness of this particular 3D audio format. Furthermore, the SELD task of the L3DAS22 for the first time proposes a scenario where multiple sounds of the same class may be active at the same time. This is a well-established scenario in vision-related object detection tasks and, to our knowledge is an important real-life-oriented study case also in the audio domain. We supply baseline models and results for both tasks, obtained using state-of-the-art deep learning architectures. Datasets and models are supported by a Python-based API aimed at facilitating the data download and preprocessing, the baseline models training and the results submission[2].

The specific contributions of this work are the following:

- We present a novel multiple-source and multiple-perspective Ambisonics dataset aimed at SELD and SE tasks.
- We propose a machine learning challenge based on this dataset that comprehends two distinct tasks, with two separate tracks each.
- We present baseline models for both tasks, obtained with state-of the art methods.
- We provide a python-based supporting API that can be used for any purpose beyond the challenge. This includes a PyTorch re-implementation of the SELDNet architecture.

2. DATASET DESCRIPTION

The L3DAS22 dataset contains approximately 90 hours of MSMP B-format audio recordings. We sampled the acoustic field of a large office room with the approximate dimensions of 6 m (length) by 5 m (width) by 3 m (height). The room has typical office furniture: desks, chairs and a wardrobe. The floor is made of wood parquet, while the walls and the ceiling are made of painted concrete.

We placed two first-order A-format Ambisonics microphones[3] in the center of the room and we moved a speaker[4] reproducing an analytic signal in 252 fixed spatial positions. One microphone (mic A) lies in the exact center of the room, and the other (mic B) is 20 cm distant towards the width dimension. Both are shown as a unique grey dot in the center of Fig. 1c. Both microphones are positioned at the same height of 1.3 m, which is the average ear height ear of a seated person. The capsules of both mics have the same orientation.

The speaker placement is performed according to two different criteria: a fixed 3D grid (168 positions) and a 3D uniform random distribution (84 positions). Figure 1c shows a 2D projection of the grid from above. For the first criterion, we placed the speaker in a 3D grid with a 50 cm step in the length-width dimensions and a 30 cm step in the height dimension, as represented in Fig. 1c with blue dots. There are 7 position layers in the height dimension at 0.3 m, 0.7 m, 1 m, 1.3 m, 1.6 m, 1.9 m, 2.3 m from the floor, as shown in Figure 1a. The random positions, instead, respect a uniform distribution and are depicted in Fig. 1b. All random-selected positions are quantized in a virtual 3D grid with a 25 cm step. For all measurements we directed the speaker's tweeter towards mic A.

The analytic signal is a 24-bit exponential sinusoidal sweep that glides from 50 Hz to 19000 Hz in 20 seconds, reproduced at 90 dB SPL on average. The IR estimation is then obtained by performing a circular convolution between the recorded sound and the time-inverted analytic signal, as introduced by [21]. We finally converted the A-format signals into standard B-format IRs[5].

Relying on the collected Ambisonics impulse responses, we augmented existing clean monophonic datasets to obtain synthetic tridimensional sound sources by convolving the original sounds with our IRs. The result of this convolution operation is the virtual placement of a sound source in the spatial position occupied by the speaker, perceived from the position of the 2 microphones. We aimed at creating plausible and variegate 3D scenarios to reflect office-like situations, in which disparate types of sound sources and background noises coexist in the same 3D reverberant environment.

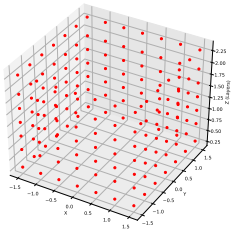
For this purpose we used the Librispeech [22] and FSD50K [23] datasets. We selected a total of 1440 noise sound files from FSD50K, divided into 14 transient noise classes: *computer keyboard, drawer open/close, cupboard open/close, finger snapping, keys jangling, knock, laughter, scissors, telephone, writing, chink and clink, printer, female speech, male speech*, and 4 continuous noise classes: *alarm, crackle, me-*

[2] https://github.com/l3das/L3DAS22

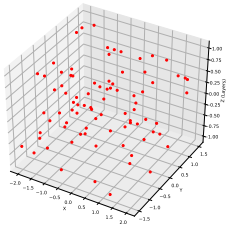
[3] Oktava MK-4012

[4] Event PS6

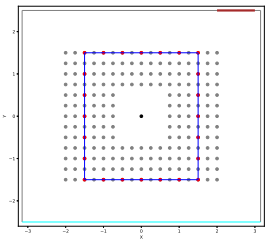
[5] http://pcfarina.eng.unipr.it/Public/B-format/A2B-conversion/A2B.htm



(a) Grid 3D



(b) Random 3D



(c) Grid projection

Fig. 1: (a) Tridimensional distribution of the fixed 3D grid. (b) Tridimensional distribution of the randomly-selected speaker positions. (c) Projection from above of the microphones position (center dot) and the speaker positions of the fixed 3D grid (red dots connected by the blue line).

chanical fan, microwave oven. We collected 80 sounds for each noise class (both for transient and continuous noises). Furthermore, we extracted clean speech signals (without background noise) from Librispeech, taking only sound files up to 10 seconds.

The dataset is divided in two main sections, respectively dedicated to the challenge tasks. We provide normalized raw waveforms of all Ambisonics channels (8 signals in total) as predictors data for both sections, whereas the target data varies significantly. Moreover, we created different types of acoustic scenarios, optimized for each specific task.

We include a first-order Ambisonics decoder (supporting decoding to mono, stereo and binaural formats) as part of the supporting API, in order to facilitate the use of our dataset with formats that are easier to handle and can be used also for different applications.

2.1. 3D Speech Enhancement Dataset

This dataset section is dedicated to task 1 and, thus, is optimized for SE. Here we created more than 40000 virtual 3D audio environments with a duration up to 12 seconds each, reaching a total duration of more than 80 hours. In each data point a speech signal is always present, mixed with various types of background noise. We extracted all sounds from the *clean* subset of Librispeech (approximately 53% male and 47% female speech). We add up to 3 non-speech background noises of the above-mentioned categories, extracting them from FSD50K. With a 25% chance, one of the background noises is a continuous noise. The signal-to-noise ratio ranges from 6 to 16 dBFS (referring to the signals' RMS amplitude), where the voice is always the prominent signal. We randomly place all sound sources in the 3D environment, paying attention to obtain a uniform distribution of locations within this dataset section.

The predictors data of this section are released as 8-channel 16 kHz 16 bit wav files, consisting of 2 sets of first-order Ambisonics recordings. The channels order follows the Ambisonic Channel Number (ACN) system of the AmbiX format [6], thus having [WA, YA, ZA, XA, WB, YB, ZB, XB], where the letters A,B refer to the used microphone and W,Y,Z,X, refers to the B-format Ambisonics channels. The target data provided for this section contains the clean monophonic recordings of the only speech signals (16 kHz 16 bit mono wav files), as well as the words uttered in each data point (in a txt file).

In this section we provide also an informative csv file for each subset, where we annotated the coordinates and spatial distance of the IR convolved with the target voice signals for each datapoint. This may be useful to estimate the delay caused by the virtual time-of-flight of the target voice signal and to perform a sample-level alignment of the input and ground truth signals.

2.2. 3D Sound Event Detection and Localizaion Dataset

This dataset section is targeted to task 2 and it is therefore optimized for SELD. Here we synthesized 900 30-seconds-long data points, reaching a total length of 7.5 hours of audio. Each data point contains a simulated 3D office audio environment in which up to 3 simultaneous acoustic events may be active at the same time. Moreover, when multiple sounds are active at the same time, with an approximate probability of 12% at least 2 sounds may belong to the same class. Although, when this happens, we forced the simultaneous sounds of the same class to be virtually positioned at least 1 meter distant one another.

The data points of this section contain an average of 26 acoustic events, with a standard deviation of 11. The sound events belong to the aforementioned 14 transient noise classes and are therefore 1120 in total. As opposed to the SE dataset, here the data points are not forced to contain speech signals, although they may contain voice sounds. The volume difference between the different sounds ranges from 0 to 20 dBFS (referring to the signal's RMS amplitude). Also here, we randomly place all sound sources in the 3D environment, paying

[6] http://pcfarina.eng.unipr.it/aurora/B-Format_to_UHJ.htm

attention to obtain a uniform distribution of locations.

The predictors data of this section have the same form of the SE section, except for the sampling frequency, which here is 32 kHz. As target data, we provide a csv file containing the onset and offset time stamps, the typology class and the spatial coordinates of each individual sound event present in a data point.

2.3. Dataset Splits

We split both dataset sections into a training set (approximately 80 hours for SE and 5 hours for SELD) and a test set (approximately 6 hours for SE and 2.5 hours for SELD), paying attention to create similar distributions. The train set of the SE section is divided in two partitions: train360 and train100, and contain speech samples extracted from the correspondent partitions of Librispeech (only the samples up to 12 seconds). All sets of the SELD section are divided in: OV1, OV2, OV3. These partitions refer to the maximum amount of possible overlapping sounds, which are 1, 2 or 3, respectively.

The test set of both dataset sections is further split into two equally-long subsets that present a similar distribution: one development and one blind test set. The first one is part of the initial release of the dataset, and it is aimed, as usual, at the model's hyperparameters fine-tuning. The latter, instead, is aimed at the submissions' evaluation and was initially released with the only predictors data, without target labels/signals.

3. CHALLENGE TASKS

We propose 2 different tasks, both based on our L3DAS22 dataset: *3D Speech Enhancement in Office Reverberant Environment* and *3D Sound Event Localization and Detection in Office Reverberant Environment*. Each one is divided in 2 sub-tasks: one-mic and dual-mic recordings, respectively relying on the sounds acquired by one or both Ambisonics microphones, as described in Section 2.

In this context, the information predicted for one task may be beneficial for the other one. For instance, the sound localization parameters may be re-used to improve the performance of 3D speech enhancement networks, as in [24, 25]. Therefore, participants are encouraged to develop a strategy to bootstrap the resources and exploit the output of one model to enhance the performance of the other one (although this is not mandatory).

3.1. Task 1: 3D Speech Enhancement in Office Reverberant Environment

The objective of this task is the separation and enhancement of speech signals immersed in a noisy 3D environment, basing on the SE section of the L3DAS22 dataset. Here the models are expected to extract the monophonic voice signal from the 3D mixture that contains various background noises. The evaluation metric for this task is a combination of the short-time objective intelligibility (STOI), which estimates the intelligibility of the output speech signal, and word error rate (WER), computed to assess the effects of the enhancement for speech recognition purposes. We use a Wav2Vec [26] architecture pre-trained on Librispeech 960h[7] to compute the WER. The final metric for this task is a combination of these two measures given by $(STOI+(1-WER))/2$. This metric lies therefore in the 0-1 range and higher values are better.

3.2. Task 2: 3D Sound Event Localization and Detection in Office Reverberant Environment

The aim of this task is to detect the temporal activity, spatial position and typology of a known set of sound events immersed in a synthetic 3D acoustic environment. This task is performed on the SELD section of the L3DAS22 dataset. Here the models are expected to predict a list of the active sound events and their respective location at regular intervals of 100 milliseconds.

We use a joint metric for localization and detection: Location-sensitive detection error, as defined in [27]. This metric is computed on each time frame and consists of measuring the cartesian distance between the predicted and true events with the same label, and counting a true positive only when its label is correct and its location is within a threshold from its reference location. After this operation, we compute the regular F score. In this challenge, we fixed the spatial error threshold to 1 meter.

4. BASELINE METHODS

As baseline methods we propose state-of-the-art architectures, specifically adapted for each task. For both tasks, we used the only signals coming from one Ambisonics microphone (mic A), leaving room for experimentation with the dual-mic configuration.

4.1. Models

For task 1 (SE), we use a beamforming U-Net architecture [9], which provided the best metrics for the L3DAS21 Challenge on the SE task. This network uses a convolutional U-Net to estimate B-format beamforming filters and contains three main modules: encoder for extracting high-level features gradually, decoder for reconstructing the size of input features from the output of the encoder, and skip connections for concatenating each layer in the encoder with its corresponding layer in the decoder. The enhancement process is performed as that of the traditional signal beamforming. We

[7] https://huggingface.co/facebook/wav2vec2-base-960h

multiply the complex spectrogram of B-format noisy signal with the filters estimated by U-Net through element-wise multiplication, and then sum the result over the channel axis to estimate a single-channel enhanced complex spectrogram. In the end the ISTFT is performed to obtain the enhanced time-domain signal. With this model we obtained a baseline test metric for task 1 of 0.82, with a word error rate of 0.23 and a STOI of 0.88.

For task 2, instead, we developed a variant of the SELDnet architecture [16]. We ported to the PyTorch language the original Keras implementation [8] and we modified its structure in order to make it compatible with the L3DAS22 dataset. The objective of this network is to output a continuous estimation (within a fixed temporal grid) of the sounds present in the environment and their respective location. The original SELDNet architecture is conceived for processing sound spectrograms (including both magnitudes and phase information) and uses a convolutional-recurrent feature extractor based on 3 convolution layers followed by a bidirectional GRU layer. In the end, the network is split in two separate branches that predict the detection (which classes are active) and location (where the sounds are) information for each target time step. We augmented the capacity of the network by increasing the number of channels and layers, while maintaining the original data flow. Moreover, we discard the phase information and we perform max-pooling on both the time and the frequency dimensions, as opposed to the original implementation, where only frequency-wise max-pooling is performed. In addition, we added the ability to detect multiple sound sources of the same class that may be active at the same time (3 at maximum in our case).To obtain this behavior we tripled the size of the network's output matrix, in order to predict separate location and detection information for all possible simultaneous sounds of the same class. We also apply a simple post-processing stage: we rescale the location prediction by a fixed value (0.5) to match at best the input distribution.

For further implementation details on our baseline models, please refer to the L3DAS official GitHub repository (link above).

5. CHALLENGE RULES AND EVALUATION CRITERIA

5.1. Rules and Requirements

The goal of the challenge is to foster research on machine learning for 3D audio. All participants should adhere to the following rules to be eligible for the challenge:

- All participants must submit the obtained results for at least one of the 2 tasks. The results should be accompanied by a paper describing the proposed method.
- Each individual participant cannot be included in multiple participating teams. Therefore, a participant is allowed to submit only one set of results.
- Winners will be selected according to the best performance for each single task, separately. Therefore, one winner for each task will be selected.
- There are no restrictions on the proposed methodologies. However, in case of a tie, the Challenge Committee will take into account the novelty and originality of the proposed approach. Also, a method that can be used for both the 1-mic and 2-mic configurations will be positively evaluated.
- Participants are not restricted to use the L3DAS22 dataset only. It is in fact allowed to augment this dataset and/or to integrate additional data to train/pre-train the models.
- The Challenge can have up to 5 papers from the top ranked teams. The format should be consistent with ICASSP regular paper, and should be submitted before the camera-ready deadline.
- Accepted papers will be presented at a Signal Processing Grand Challenge special session of the IEEE ICASSP 2022 conference. Authors, who are not interested in participating the challenge but want to make a contribution to the topic, are encouraged to submit a paper to this track, even without specifically use the proposed datasets.

5.2. Paper and Results Submissions

- Results and paper must be submitted within the deadlines shown in Subsection 5.4.
- Results should be prepared and formatted according to the guidelines detailed in the challenge submission page[9].
- The accompanying paper must describe the proposed method and must contain all the details to ensure reproducibility. The paper must also include information about the computational complexity of the model (e.g., in terms of number of parameters or execution time on a specific device).
- Papers should be prepared according to the guidelines of IEEE ICASSP 2022 [10].
- Submitted papers will undergo the standard peer-review process of IEEE ICASSP 2022.

[8] https://github.com/sharathadavanne/seld-net

[9] www.l3das.com/icassp2022/submission

[10] https://2022.ieeeicassp.org/papers/paper_kit.php

Optionally, you can inform us at l3das@uniroma1.it about your submission.

5.3. Support

The challenge participants are encouraged to contact our team at l3das@uniroma1.it for any issue of clarification about the challenge or the dataset.

5.4. Timeline

- **Nov 22, 2021** – Release of the Training and Development Sets, Code, Baseline Methods and Documentation.
- **Dec 15, 2021** – Release of the Evaluation Test Set
- **Dec 22, 2021** – Deadline for Submitting Results for Both Tasks
- **Jan 5, 2022** – Notification of Top Ranked Teams
- **Jan 20, 2022** – Deadline for Paper Submission (Top Ranked 5 Only)
- **Feb 10, 2022** – Grand Challenge Paper Acceptance Notification
- **Feb 16, 2022** – Camera-Ready Grand Challenge Papers Deadline
- **Mar 22, 2022** – Opening of the IEEE ICASSP 2022 and Winner Announcement

6. REFERENCES

[1] Jakob Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, 2020.

[2] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features," in *2018 IEEE International Joint Conference on Neural Networks, (IJCNN)*, Rio de Janeiro, Brazil, Jul. 2018, pp. 1–7.

[3] Sharath Adavanne, Pasi Pertilä, and Tuomas Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 771–775.

[4] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, pp. 1702–1726, 2018.

[5] Yi Luo, Cong Han, Nima Mesgarani, Enea Ceolini, and Shih-Chii Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, Dec. 2019, pp. 260–267.

[6] Heitor R. Guimarães, Hitoshi Nagano, and Diego W. Silva, "Monaural speech enhancement through deep Wave-U-Net," *Expert Syst. Appl.*, vol. 158, pp. 1–10, 2020.

[7] Craig Macartney and Tillman Weyde, "Improved speech enhancement with the Wave-U-Net," *arXiv preprint: arXiv:1811.11307v1*, 2018.

[8] Amélie Bosca, Alexandre Guérin, Lauréline Perotin, and Srdan Kitic, "Dilated U-Net based approach for multichannel speech enhancement from first-order Ambisonics recordings," in *28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, Netherlands, Jan. 2021, pp. 216–220.

[9] Xinlei Ren, Lianwu Chen, Xiguang Zheng, Chenglin Xu, Xu Zhang, Chen Zhang, Liang Guo, and Bing Yu, "A neural beamforming network for b-format 3d speech enhancement and recognition," *2021 IEEE International Workshops on Machine Learning for Signal Processing, L3DAS21 Challenge track*, Oct. 25–28 2021.

[10] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 1562–1566.

[11] Xue Yan, Zhen Yang, Tingting Wang, and Haiyan Guo, "An iterative graph spectral subtraction method for speech enhancement," *Speech Commun.*, vol. 123, pp. 35–42, 2020.

[12] Cunhang Fan, Bin Liu, Jianhua Tao, Jiangyan Yi, and Zhengqi Wen, "Discriminative learning for monaural speech separation using deep embedding features," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Gernot Kubin and Zdravko Kacic, Eds., Graz, Austria, Sep. 2019, pp. 4599–4603.

[13] Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[14] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 684–698, 2021.

[15] Archontis Politis, Sharath Adavanne, and Tuomas Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv preprint: arXiv:2006.01919v2*, 2020.

[16] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 1, pp. 34–48, 2019.

[17] Karim Guirguis, Christoph Schorn, Andre Guntoro, Sherif Abdulatif, and Bin Yang, "SELD-TCN: Sound event localization & detection via temporal convolutional networks," in *28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, Netherlands, Jan. 2021, pp. 16–20.

[18] Sotirios Panagiotis Chytas and Gerasimos Potamianos, "Hierarchical detection of sound events and their localization using convolutional neural networks with adaptive thresholds," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE)*, 2019, pp. 50–54.

[19] Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, and Mark D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," *arXiv preprint: arXiv:1905.00268v4*, 2019.

[20] Luca Mazzon, Yuma Koizumi, Masahiro Yasuda, and Noboru Harada, "First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," *arXiv preprint: arXiv:1910.04388v1*, 2019.

[21] Angelo Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *108th Convention of the Audio Engineering Society*, Paris, France, Feb. 2000.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 5206–5210.

[23] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "FSD50K: an open dataset of human-labeled sound events," *arXiv preprint: arXiv:2010.00475v1*, 2020.

[24] Shlomo E. Chazan, Hodaya Hammer, Gershon Hazan, Jacob Goldberger, and Sharon Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *27th European Signal Processing Conference, (EUSIPCO)*, A Coruña, Spain, Sep. 2019, pp. 1–5.

[25] Ali Aroudi and Sebastian Braun, "DBNET: DOA-driven beamforming network for end-to-end farfield sound source separation," *arXiv preprint: arXiv:2010.11566v1*, 2020.

[26] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2020.

[27] Annamaria Mesaros, Sharath Adavanne, Archontis Politis, Toni Heittola, and Tuomas Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2019, pp. 333–337.