

L3DAS21 CHALLENGE: MACHINE LEARNING FOR 3D AUDIO SIGNAL PROCESSING

Eric Guizzo, Riccardo F. Gramaccioni, Saeid Jamili, Christian Marinoni, Edoardo Massaro, Claudia Medaglia, Giuseppe Nachira, Leonardo Nucciarelli, Ludovica Paglialunga, Marco Pennese, Sveva Pepe, Enrico Rocchi, Aurelio Uncini, and Danilo Comminiello

Department of Information Engineering, Electronics and Telecommunications
Sapienza University of Rome, Italy

ABSTRACT

The L3DAS21 Challenge¹ 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 L3DAS21 dataset, a 65 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: FaSNet 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 L3DAS21 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 UNet-based approaches provide competitive results in this context, both for monaural [6, 7] and multichannel SE tasks [8], at the expense of higher computational power demand. Other techniques to perform SE include recurrent neural networks (RNNs) [9], graph-based spectral subtraction [10], discriminative learning [11], dilated convolutions [12].

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 [13]. Also here, the state-of-the-art methods are based on deep learning strategies [14]. SELDnet [15] 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 [16], 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.

¹www.l3das.com/mlsp2021

Other novel solutions for this task include ensemble models [17], multi-stage training [18] and bespoke augmentation strategies [19].

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 L3DAS21 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 L3DAS21 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 submission².

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 LEDAS21 dataset contains approximately 65 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 microphones³ in the center of the room and we moved a speaker⁴ 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 16000 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 [20]. We finally converted the A-format signals into standard B-format IRs⁵.

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 variegated 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 [21] and FSD50K [22] 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-*

³Oktava MK-4012

⁴Event PS6

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

²<https://github.com/l3das/L3DAS21>

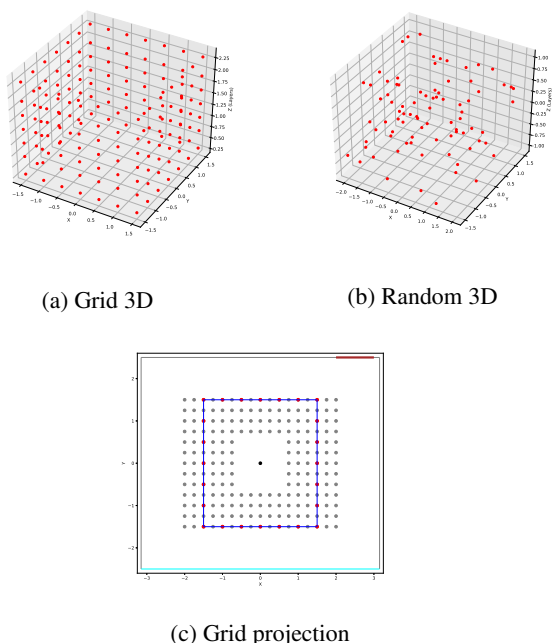


Fig. 1: (a) Tridimensional distribution of the randomly-selected speaker positions. (b) Tridimensional distribution of the fixed 3D grid. (c) Projection from above of the microphones position (center dot) and the speaker positions of the fixed 3D grid (red dots connected by the blue line).

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 30000 virtual 3D audio environments with a duration up to 10 seconds each, reaching a total duration of approximately 50 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 format⁶, 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 csv file).

2.2. 3D Sound Event Detection and Localization Dataset

This dataset section is targeted to task 2 and it is therefore optimized for SELD. Here we synthesized 900 1-minute-long data points, reaching a total length of 15 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.

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 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.

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

2.3. Dataset Splits

We split both dataset sections into a training set (44 hours for SE and 600 hours for SELD) and a test set (6 hours for SE and 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 sample) up to 10 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 L3DAS21 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 [23, 24]. 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 L3DAS21 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 [25] ar-

chitecture pre-trained on Librispeech 960h⁷ 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 L3DAS21 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 [26]. 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, which is half extension of the spatial axes.

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 Filter and Sum Network architecture (FaSNet) [5], adapted from this public PyTorch implementation⁸. This network is a state-of-the-art neural beamformer that operates in the time domain and, therefore, work on both the magnitude and the phase information of the signal. The objective of this network is to estimate spatial filters that are individually applied to each input signal channel. In the end all channels are summed up to obtain a mono-aural output channel. The network operates in two consecutive stages: first, the spatial filter for a single randomly-selected reference channel is computed, and then such information is used to infer the filters for all remaining channels. Both stages are based on temporal convolutional networks [27]. Please, refer to the original paper [5] for further details on this architecture.

⁷<https://huggingface.co/facebook/wav2vec2-base-960h>

⁸<https://github.com/ylyuo42/TAC>

For task 2, instead, we developed a variant of the SELDNet architecture [15]. We ported to the PyTorch language the original Keras implementation⁹ and we modified its structure in order to make it compatible with the L3DAS21 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.

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

4.2. Results

Tables 1a and 1b show the baseline results we obtained for task 1 and task 2, respectively. Even though we used state-of-the-art architectures for each task, the results show that there is significant room for improvement. For task 1, the network clearly exhibits underfitting. This may be partly due to the relatively narrow volume difference between the target voice and the background noises (randomly selected between 6 and 16 dB). This difference is based on the RMS amplitude of the signals and therefore it happens frequently that transients and small portions of noise sounds may be louder than the target voice. Moreover, we used a neural beamforming technique, which, by its very nature, is not able to suppress unwanted sounds that are located near the source to be isolated. Therefore, to address this limitation, different networks may be considered for this task, as, for instance, u-net-like architectures. For task 2, instead, both underfitting and overfitting are evident from the results. We also tested a vanilla implementation of SELDNet, using the same configuration as described in the original paper [15] (with an augmented output dimension, in order to be compatible with the L3DAS21 dataset), obtaining a test F-score of 0.14. The performance

Metric	Train	Val	Test
Joint	0.67	0.62	0.65
STOI	0.74	0.72	0.73
WER	0.41	0.46	0.44

(a) Task1: Speech Enhancement.

Metric	Train	Val	Test
F-Score	0.66	0.45	0.39
Precision	0.75	0.52	0.45
Recall	0.6	0.4	0.34

(b) Task 2: Sound Event Localization and Detection.

Table 1: Baseline results for the two challenge tasks. Task 1 results are obtained with a FasNet architecture, while task 2 results are obtained with an augmented SELDNet architecture.

gain provided by the augmented SELDNet (test F-score of 0.39) indicates that the standard capacity of the vanilla SELDNet, which provides state-of-the-art results for many SELD datasets, is not sufficient for the L3DAS21 dataset. We hypothesize that the increased difficulty depends largely on the fact that in the L3DAS21 dataset multiple sounds of the same class may be active at the same time. This forces the models to predict larger matrices (3 times the standard SELDNet output, in our case). Nevertheless, this behavior well reflects real-life scenarios, where there is no limit on the number of similar simultaneous sounds.

5. CONCLUSIONS

This paper presented the details of the L3DAS21 Challenge: the L3DAS21 dataset, the challenge tasks and the baseline models. The audio dataset has been generated in a novel multichannel configuration based on an array of two first-order Ambisonics microphones. The baseline results, obtained using the only signals of one microphone, show that both tasks are considerably challenging even for state-of-the-art models (FaSNet and SELDNet, respectively). We obtained a test metric of 0.65 for task 1 (a combination of WER and STOI) and a test F-Score of 0.39 for task 2 (based on location-sensitive detection). We hypothesize that these relatively low metrics could be due primarily to the narrow SNR ratio for task 1 and to the presence of overlapping sounds of the same class for task 2. Therefore the challenging dataset setup and the novel recording format leave room for experimentation in multichannel SE and SELD.

As future work we intend to analyze and discuss the results submitted by the challenge participants in a new manuscript. Moreover, we are working on an improved version of the L3DAS21 dataset that contains additional modeled room IRs and a broader selection of sound classes.

⁹<https://github.com/sharathadavanne/seld-net>

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] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] Ali Aroudi and Sebastian Braun, “DBNET: DOA-driven beamforming network for end-to-end farfield sound source separation,” *arXiv preprint: arXiv:2010.11566v1*, 2020.
- [25] 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.
- [26] 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.
- [27] Yi Luo and Nima Mesgarani, “Tasnet: Surpassing ideal time-frequency masking for speech separation,” *CoRR*, vol. abs/1809.07454, 2018.