L3DAS21: Machine Learning for 3D Audio Signal Processing

IEEE MLSP Data Challenge 2021

Scope of the Challenge

3D audio is gaining increasing interest in the machine learning community in recent years. The field of application is incredibly wide and ranges from virtual and real conferencing to game development, music production, autonomous driving, surveillance and many more. In this context, Ambisonics prevails among other 3D audio formats for its simplicity, effectiveness and flexibility. Ambisonic recordings permit to obtain an impressive performance in many machine learning-based tasks, usually bringing out a significant improvement over the mono and stereo formats. Tasks like Sound Source Localization, Speech and Emotion Recognition, Sound Source Separation Separation, Speech Enhancement and Denoising, Acoustic Echo Cancellation, among others, benefit from tridimensional representations of sound field, thus leading to higher accuracy and perceived quality.

The L3DAS project (Learning 3D Audio Sources) aims at encouraging and fostering research on the afore-mentioned topics. In particular, the L3DAS21 Challenge focuses on 2 tasks: 3D Speech Enhancement and 3D Sound Event Localization and Detection, both relying on Ambisonics recordings. First, provide the training and development sets, alongside with a supporting python-based API to facilitate the data download and pre-processing. We also supply baseline results for both tasks, obtained using state-of-the art deep learning architectures. In a second step, we will release the test sets without truth labels.

Participants must submit the results obtained for the latter. In the end, the final ranking of the challenge will be presented at the IEEE Workshop on MLSP and released on the challenge webpage.

Dataset

The LEDAS21 dataset contains multiple-source and multiple-perspective B-format Ambisonics audio recordings. We sampled the acoustic field of a large office room, placing two first-order Ambisonics microphones in the center of the room and moving a speaker reproducing the analytic signal in 252 fixed spatial positions. Relying on the collected Ambisonics impulse responses (IRs), we augmented existing clean monophonic datasets to obtain synthetic tridimensional sound sources by convolving the original sounds with our IRs. Clean files have been extracted from the Librispeech [1], FSD50K [2] and Audioset [3] datasets.

We aimed at creating plausible and variegate 3D scenarios to reflect possible real-life situations in which sound and disparate types of background noises coexist in the same 3D reverberant environment. We provide normalized raw waveforms as predictors data and the target data varies according to the task, as specified in the next section.

Tasks

We propose 2 tasks: 3D Speech Enhancement and 3D Sound Source Localization and Detection. These tasks are aimed at fulfilling real-world needs related to real and virtual conferencing. Especially in multi-speaker scenarios it is 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 rather than assistive listening or entertainment, among others).

Each task presents 2 separate sub-tasks: 1-mic and 2-mic recordings, respectively containing the sounds acquired by one Ambisonics microphone and by an array of two Ambisonics microphones. To the best of our knowledge, this is the first time that Ambisonics recording performed with 2 microphones are considered for machine learning purposes. We expect higher accuracy/reconstruction quality when taking advantage of the dual spatial perspective of the two microphones. Both tasks rely on the same audio recordings, but with completely different targets, as described below.

- Task 1: 3D Speech Enhancement

  The objective of this task is the enhancement of speech signals immersed in a noisy 3D environment. Here the models are expected to extract the monophonic voice signal from the 3d mixture containing various background noises. Therefore, for this task we also provide the clean monophonic speech signal as target. The evaluation metric for this task is short-time objective intelligibility (STOI), which estimates the intelligibility of the output speech signal. Moreover, word error rate (WER) is also computed to assess the effects of the enhancement for speech recognition purposes.

- Task 2: 3D Sound Event Localization and Detection in Office Environment

  The aim of this task is to detect the temporal activities of a known set of sound event classes and to further locate them in the space, in particular, we focus on the sound event localization and detection (SELD). Sounds were chosen from the FSD50K sound dataset [2]. We selected 16 classes that are suitable with an office environment, paying attention to take only short sound events. We consider up to 3 simultaneously active sounds, 2 of which may be the same sound. Localization metric is based on the direction of arrival estimation error, while detection performance is evaluated on an F1-score metric.

References

[1] Panayotov, Vassil, et al. "Librispeech: an asr corpus based on public domain audio books." *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015.

[2] Fonseca, Eduardo, et al. "FSD50k: an open dataset of human-labeled sound events." *arXiv preprint arXiv:2010.00475* (2020).

[3] Gemmeke, Jort F., et al. "Audio set: An ontology and human-labeled dataset for audio events." *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

Timeline

- 27 Mar 2021 – Release of the datasets (training and development sets), supporting code and documentation
- 10 May 2021 – Release of the evaluation test set
- 20 May 2021 – Deadline for submitting results for both tasks
- 27 May 2021 – Notification of the results of participants
- 31 May 2021 – Deadline for 6-page paper submission
- 31 Jul 2021 – Notification of paper acceptance
- 02 Aug 2021 – Notification of challenge winners
- 31 Aug 2021 – Deadline for camera-ready papers
- 25 Oct 2021 – Opening of the IEEE Workshop of MLSP 2021

Challenge Website and Contacts

L3DAS21 Challenge Website: https://sites.google.com/uniroma1.it/l3das/mlsp2021

Email contact: l3das@uniroma1.it

Organizers

Danilo Comminiello, Associate Professor, Sapienza University of Rome, Italy

Eric Guizzo, Research Fellow, Sapienza University of Rome, Italy