L3DAS22 Challenge: Learning 3D Audio Sources in a Real Office Environment

Eric Guizzo[*], Christian Marinoni[*], Marco Pennese[*], Xinlei Ren[†], Xiguang Zheng[†], Cheng Zhang[†], Bruno Masiero[‡], Aurelio Uncini[*], and Danilo Comminiello[*]

[*] *Dept. of Information Engineering, Electronics and Telecommunications (DIET), Sapienza University of Rome, Italy*

[†] *Kuaishou Technology Co., Beijing, China*

[‡] *Communications Dept. (DECOM), State University of Campinas, Brazil*

IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP 2022

Singapore, 7–13 / 22–27 May 2022

L3DAS: Learning 3D Audio Sources

- The spread of commercial devices that support 3D audio has opened new and interesting advances also from the scientific point of view.
- However, available amounts of native 3D data are *not always sufficient* for the development of new deep learning algorithms.
- The **L3DAS project** aims at filling this gap and fostering the proliferation of new deep learning methods for 3D audio.

L3DAS

Learning 3D Audio Sources

www.l3das.com/icassp2022

Introducing the L3DAS22 Challenge

- The **L3DAS22** Signal Processing and Grand Challenge is aimed at encouraging machine learning strategies for 3D audio applications in real reverberant environments.

- The challenge presents **2 tasks**:
  - 3D Speech Enhancement (SE)
  - 3D Sound Event Localization and Detection (SELD)

- 3D audio signals were collected by using 2 first-order A-format Soundfield Ambisonics microphones.

- Each task has 2 subtasks: 1-mic and 2-mic configurations.

New insights of this challenge edition

- This challenge improves and extends the tasks of the **L3DAS21** edition at IEEE MLSP 2021.
- We generated a new dataset with an extended number of data points (30 additional hours).
- We updated the baseline models, involving the architecture that ranked first in L3DAS21.
- We improved the dataset synthesis pipeline to promote less resource-demanding training.
- We wrote a new supporting API, improving its clarity and ease-of-use.
- We included prizes and features for participants (e.g., interactive Replicate demos).

Task 1: 3D Speech Enhancement

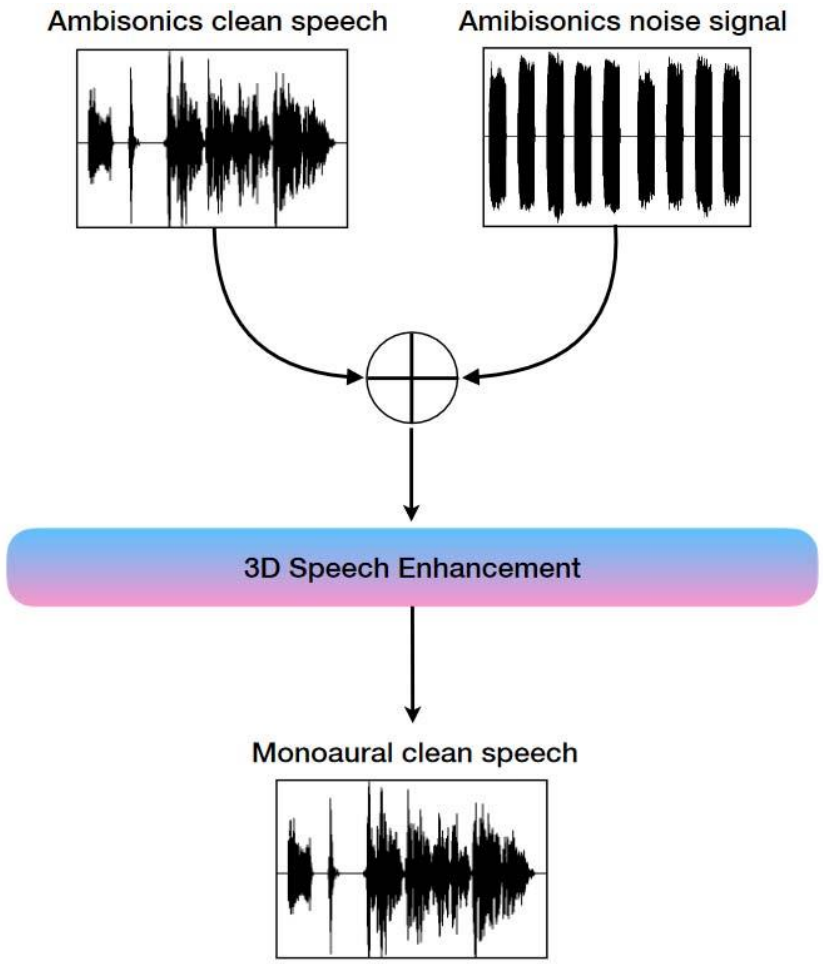
- Models are expected to extract the monophonic speech signal from a 3D noisy mixture.

- **Evaluation metric:** combination of STOI and WER:

$$\frac{\left(\mathrm{STOI} + (1 - \mathrm{WER})\right)}{2}$$

This metric lies therefore in the 0-1 range, where the higher the better.

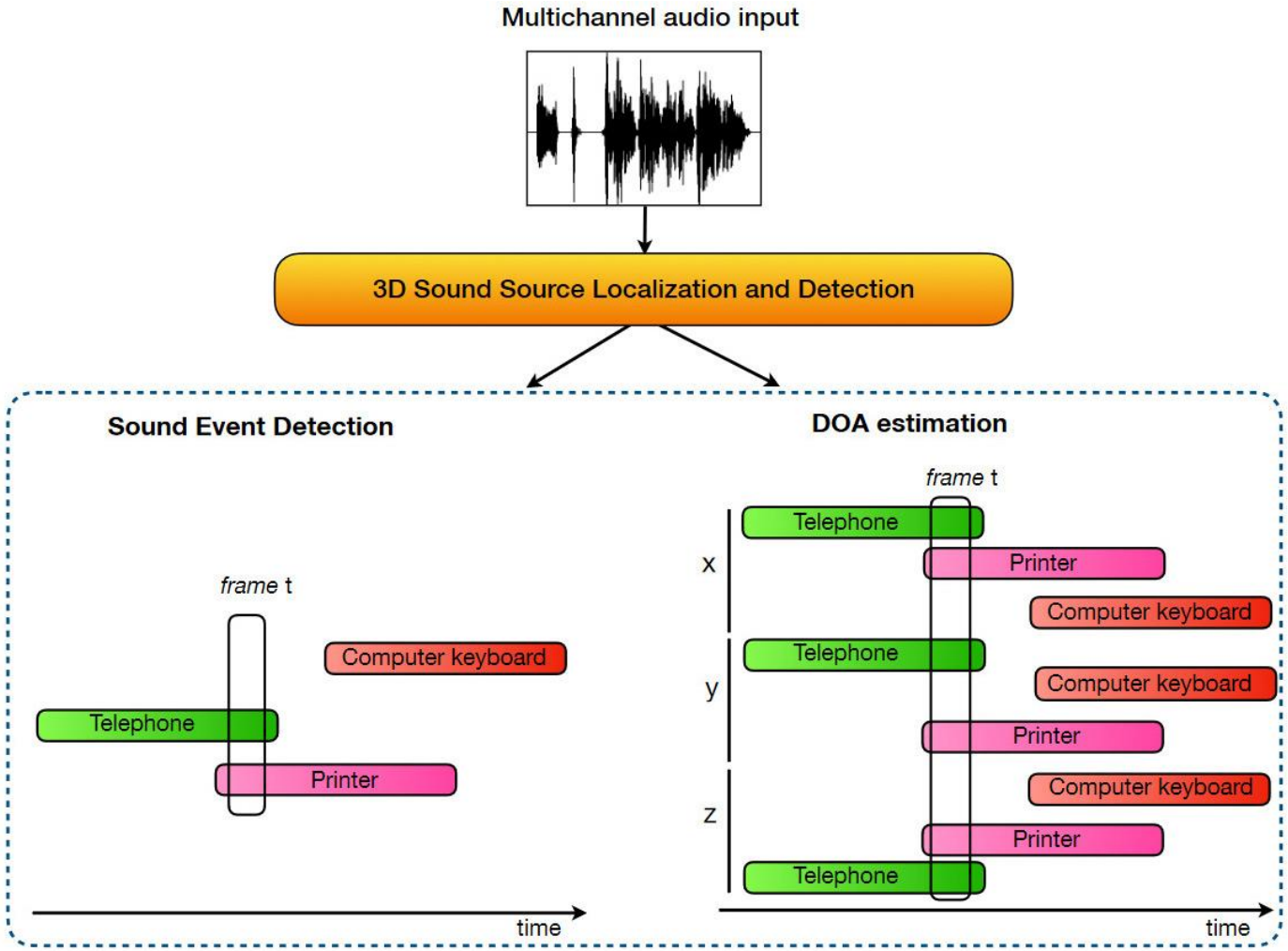
WER is obtained by using Wav2Vec.



Task 2: 3D Sound Event Localization and Detection

- Models are expected to predict a list of the active sound events and their respective location at regular intervals of 100 milliseconds.

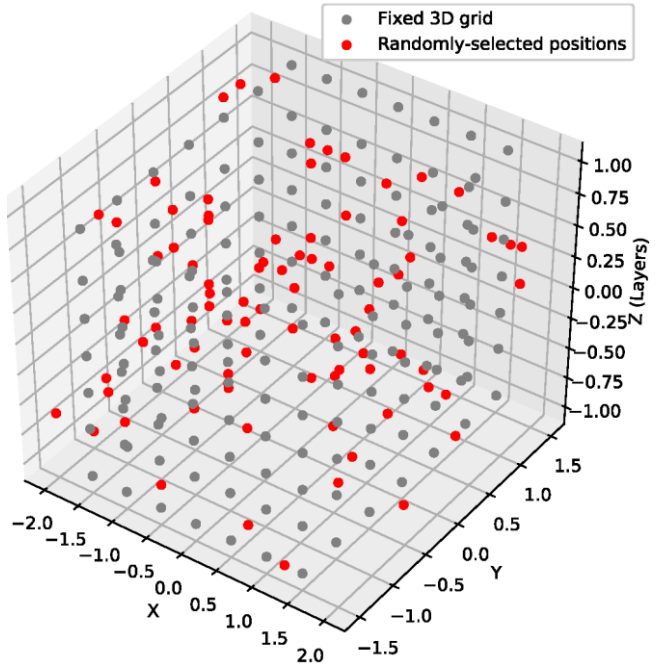
- **Evaluation metric:** we use a location-sensitive detection error, based on the Cartesian distance between the predicted and true events with the same label.

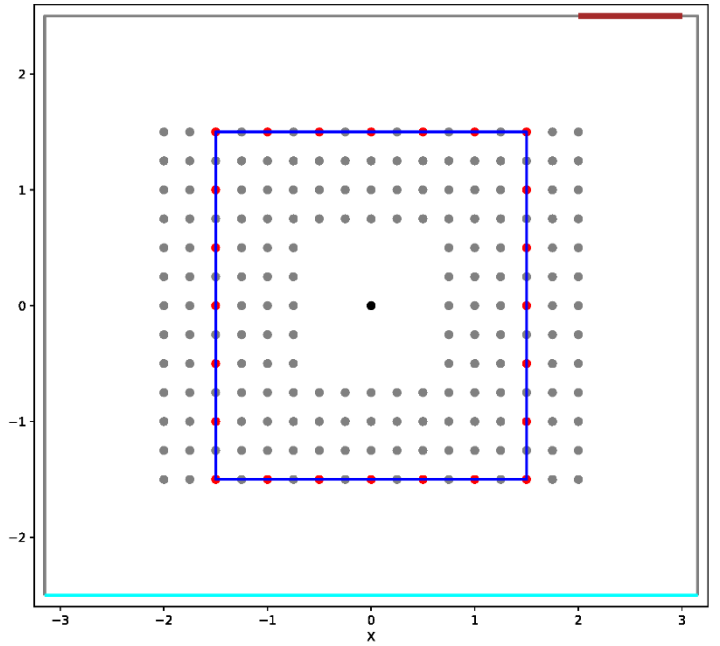


A 3D audio dataset from a real reverberant office environment

- The L3DAS22 dataset contains approximatively 98 hours of multiple-source multiple-perspective (MSMP) Ambisonics recordings.
- We sampled the acoustic field of a real office room (6x5x3 m).
- We used 2 first-order Ambisonics, 20 cm apart from each other.
- Analytic signal: 24-bit exponential sinusoidal sweep (50-to-16000 Hz).

3D room acoustic sampling





- 252 spatial positions: 168 from a fixed 3D grid (minimum distance 50 cm) and 84 from a 3D uniform random distribution (minimum distance 25 cm).

- Datasets were achieved by convolving IRs with sound sources from Librispeech (voice) and FSD50K (background noises).

Dataset section for Task 1: 3D Speech Enhancement

- We synthesized more than 40000 virtual 3D audio signals for a total length of 90 hours.
- Each data frame may contain speech and up to 3 simultaneous background noise sources.
- The signal-to-noise ratio ranges from 6 to 16 dBFS; speech is always the prominent signal.
- Target data contain the clean speech signals and the words uttered in each data frame.
- The training set contains approximatively 80 hours of audio (divided in 2 partitions).
- The test set was split into a development test and a blind test sets.

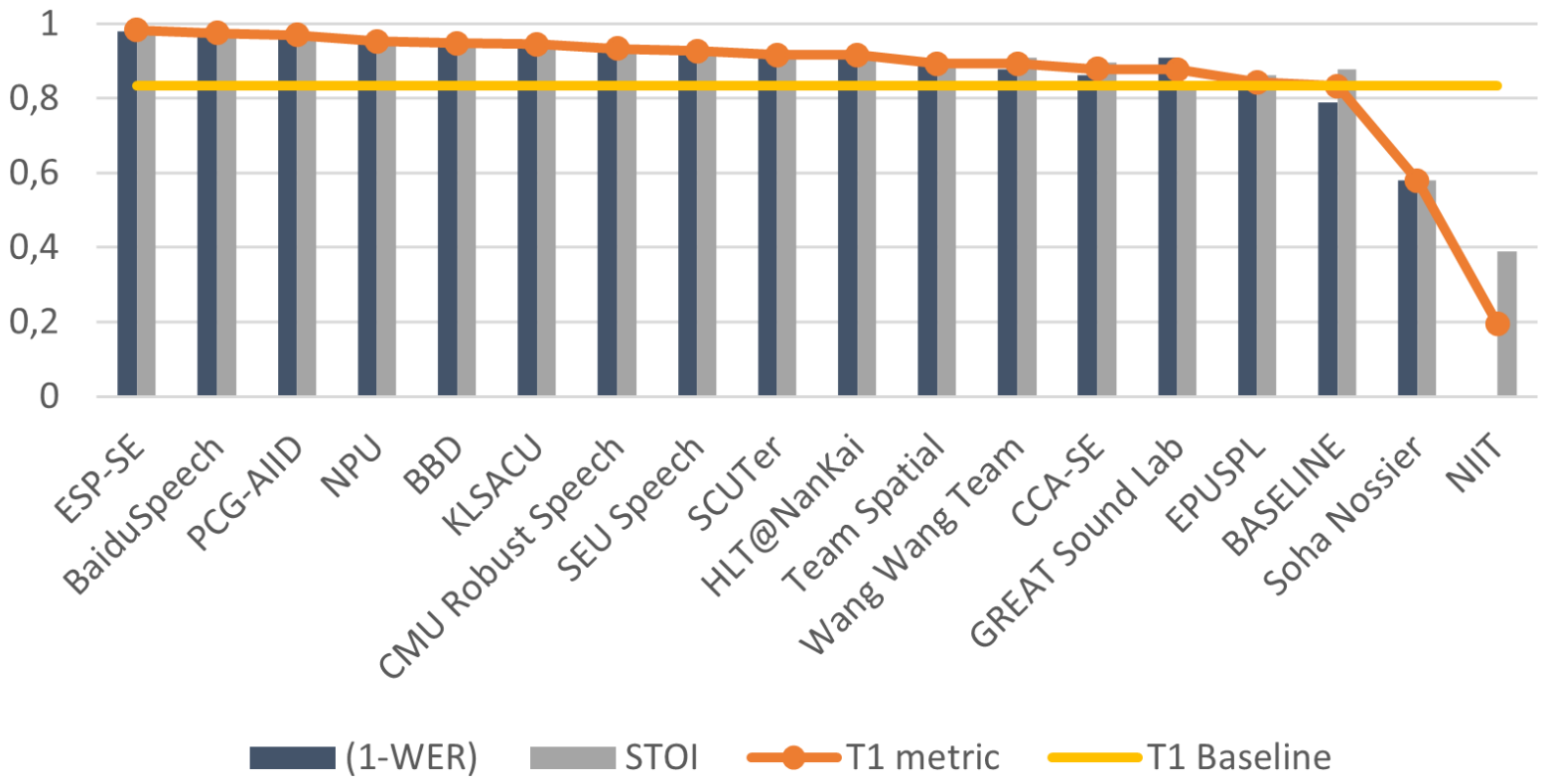
Dataset section for Task 2: 3D Sound Event Localization and Detection

- We synthesized 900 30-second 3D audio signals for a total length of 7,5 hours.
- Each data frame may contain up to 3 simultaneous 3D sound events.
- In case of 3 overlaps, two events may belong to the same class (minimum distance of 1 m).
- The volume difference between the different sounds ranges from 0 to 20 dBFS.
- The training set contains approximatively 5 hours of audio.
- The test set was split into a development test and a blind test sets.

Baseline methods

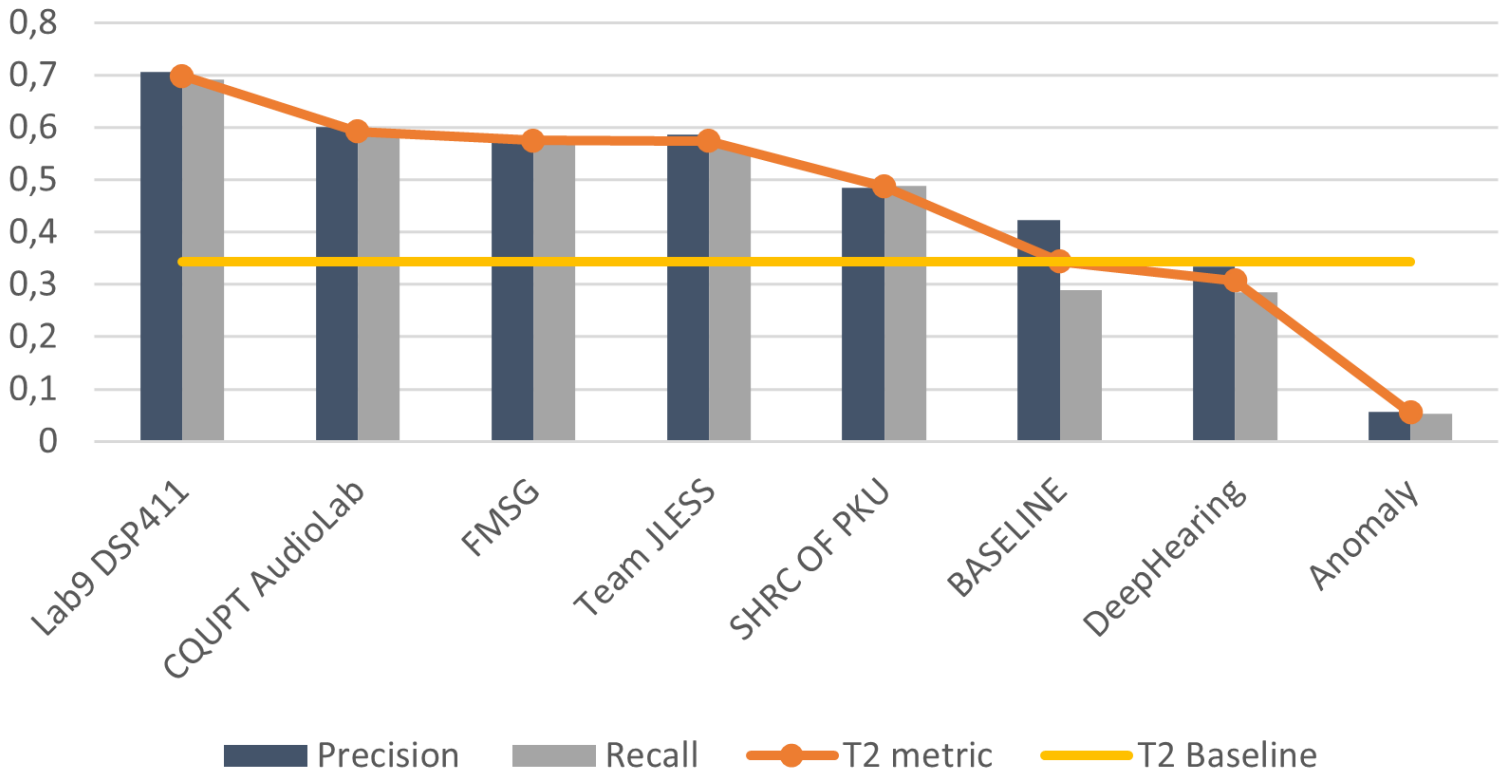
- For Task 1 (3D SE), we use a beamforming U-Net architecture, which provided the best metrics for the L3DAS21 Challenge on the SE task.
- This model yields a baseline test metric of **0.83**, with a WER of 0.21 and a STOI of 0.88.
- For Task 2 (3D SELD), we developed a variant of the SELDnet architecture, with an augmented network capacity and the ability to predict multiple sources of the same class.
- This network obtains a baseline test score of **0.34**, with a precision of 0.42 and a recall of 0.29.

L3DAS22 challenge results for Task 1



- 17 teams submitted their results for the Task 1: 3D speech enhancement.
- The winner team for Task 1, **ESP-SE**, has obtained a metric score of **0.984**, with a WER of 0.019 and a STOI of 0.987.

L3DAS22 challenge results for Task 2



- 7 teams submitted their results for the Task 2: 3D SELD.
- The winner team for Task 2, **Lab9 DSP411**, has obtained a metric score of **0.699**, with a precision of 0.706 and a recall of 0.691.

Conclusion

- The L3DAS22 Challenge has received **46 registrations** and **24 result submissions**!
- We released 2 datasets, for 3D SE and 3D SELD, freely available on Kaggle. A repository is also available on GitHub.
- Kuaishou Technology supported L3DAS22 Challenge with prizes for winners.
- Replicate provided a free account for interactive demos to challenge participants.
- The L3DAS22 Challenges has been endorsed by the International Speech Communication Association (ISCA), which accepted (the rest of the) papers on its archive providing a DOI.

Future L3DAS challenges

- Future challenges by the L3DAS Team will involve:
  - ✓ more challenging 3D SE and SELD scenarios and new 3D audio tasks,
  - ✓ different 3D microphone configurations,
  - ✓ benefits and interactive features for participants,
  - ✓ extended supporting API, e.g., including TensorFlow and MATLAB codes,
  - ✓ new partnerships.

- Acknowledgements

SAPIENZA
UNIVERSITÀ DI ROMA

KUAISHOU

Replicate

UNICAMP

IEEE
Signal
Processing
Society

isca

22-27 MAY 2022

ICASSP 2022

SINGAPORE

IEEE Signal Processing Society

IEEE

Thank You for Your Attention

Danilo Comminiello

danilo.comminiello@uniroma1.it

L3DAS

Learning 3D Audio Sources

www.l3das.com/icassp2022