Received 8 August 2023; revised 2 November 2023; accepted 3 November 2023. Date of publication 12 March 2024; date of current version 18 June 2024. The review of this article was arranged by Associate Editor R. Serizel.

Digital Object Identifier 10.1109/OJSP.2024.3376297

L3DAS23: Learning 3D Audio Sources for Audio-Visual Extended Reality

RICCARDO F. GRAMACCIONI [1] **(Graduate Student Member, IEEE), CHRISTIAN MARINONI** [1] **(Graduate Student Member, IEEE), CHANGAN CHEN**[2], **AURELIO UNCINI** [1] **(Senior Member, IEEE), AND DANILO COMMINIELLO** [1] **(Senior Member, IEEE)**

[1]Sapienza University of Rome, 00185 Roma, Italy
[2]UT Austin, Austin, TX 78712 USA

CORRESPONDING AUTHORS: RICCARDO F. GRAMACCIONI; CHRISTIAN MARINONI (e-mail: riccardofosco.gramaccioni@uniroma1.it, christian.marinoni@uniroma1.it).

The work of Riccardo F. Gramaccioni and Christian Marinoni was supported by the European Union through the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on National Centre for HP C, Big Data and Quantum Computin (CN00000013 - Spoke 6: Multiscale Modelling & Engineering Applications). The work of Aurelio Uncini was supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on Future Artificial Intelligence Research (P E0000013 - FAIR - Spoke 5: High Quality AI). The work of Danilo Comminiello was supported by Progetti di Ricerca Medi of Sapienza University of Rome for the project SAID: Solving Audio Inverse problems with diffusion models, under Grant RM123188F75F8072.

We support the dataset download and the use of the baseline models via extensive instructions provided on the official GitHub repository at https://github.com/l3das/L3DAS23. For more comprehensive information and in-depth details about the challenge, we invite the reader to visit the L3DAS Project website at http://www.l3das.com/icassp2023.

ABSTRACT The primary goal of the L3DAS (Learning 3D Audio Sources) project is to stimulate and support collaborative research studies concerning machine learning techniques applied to 3D audio signal processing. To this end, the L3DAS23 Challenge, presented at IEEE ICASSP 2023, focuses on two spatial audio tasks of paramount interest for practical uses: 3D speech enhancement (3DSE) and 3D sound event localization and detection (3DSELD). Both tasks are evaluated within augmented reality applications. The aim of this paper is to describe the main results obtained from this challenge. We provide the L3DAS23 dataset, which comprises a collection of first-order Ambisonics recordings in reverberant simulated environments. Indeed, we maintain some general characteristics of the previous L3DAS challenges, featuring a pair of first-order Ambisonics microphones to capture the audio signals and involving multiple-source and multiple-perspective Ambisonics recordings. However, in this new edition, we introduce audio-visual scenarios by including images that depict the frontal view of the environments as captured from the perspective of the microphones. This addition aims to enrich the challenge experience, giving participants tools for exploring a combination of audio and images for solving the 3DSE and 3DSELD tasks. In addition to a brand-new dataset, we provide updated baseline models designed to take advantage of audio-image pairs. To ensure accessibility and reproducibility, we also supply supporting API for an effortless replication of our results. Lastly, we present the results achieved by the participants of the L3DAS23 Challenge.

INDEX TERMS 3D audio, ambisonics, data challenge, sound event localization and detection, speech enhancement.

I. INTRODUCTION

Nowadays, 3D immersive audio is becoming a widespread reality thanks to new emerging technologies and commercial devices. The use of spatial audio can benefit a multitude of applications, including virtual and real conferencing, game development, music production, augmented reality and immersive technologies in virtual environments, speech communication, home assistants, multimedia services, audio surveillance in public spaces, and various other potential domains. The widespread adoption of 3D audio has not only brought practical benefits but has also fostered intriguing scientific advancements, particularly regarding deep learning methodologies for audio signal processing.

However, the development of efficient deep learning algorithms necessitates a substantial amount of data, which may not always be accessible for 3D audio applications. Recognizing this limitation, the L3DAS (Learning 3D Audio Sources) project aims to fill this gap by facilitating the availability of

3D audio datasets. Thereby, the primary goal of the project is to encourage the rise of novel deep learning techniques for spatial audio applications.

In the first edition of this project, L3DAS21 [1], we proposed a novel multi-channel audio configuration based on multiple-source, multiple-perspective (MSMP) Ambisonics recordings made with an array of two first-order Ambisonics microphones. As far as we know, that was the first time that a two-microphone Ambisonics configuration has been used for the tasks of 3D sound event localization and detection (3DSELD) and 3D speech enhancement (3DSE). The baselines used for the 2021 challenge were FaSNet [2] for 3DSE and SELDnet [3] for 3DSELD. Recordings were made in an office room with approximate dimensions of 6 × 5 × 3 m. We placed two first-order A-format Ambisonics microphones in the center of the room and we moved a speaker reproducing an analytic signal in 252 fixed spatial positions. The resulting L3DAS21 dataset contains approximately 65 hours of MSMP B-format Ambisonics recordings. The winning team for 3DSE of the L3DAS21 Challenge was 1024 k Team with the work presented in [4]. While the work with the best results [5] for 3DSELD was submitted by the EPUSPL Team. Detailed information can be found on the L3DAS project website for the 2021 edition.[1]

For the second edition of this project, L3DAS22 [6], we maintained a similar setting to that proposed in L3DAS21 but with some substantial improvements. Firstly, we generated a new dataset containing an augmented number of datapoints, increasing the total length of the dataset from 65 to more than 94 hours. Then, we modified the dataset synthesis pipeline in order to promote less resource-demanding training and facilitate both tasks. In addition, we updated the baseline for 3DSE, using a beamforming U-Net architecture [4], which provided the best metrics for the L3DAS21 Challenge on the 3DSE task. This network uses a convolutional U-Net to estimate B-format beamforming filters. The winning teams for the L3DAS22 edition of the Challenge were ESP-SE [7] for 3DSE and Lab9 DSP411 [8] for 3DSELD. Further information can be found on the L3DAS website.[2]

Our latest edition of the L3DAS project, presented as Signal Processing Grand Challenge at IEEE ICASSP 2023, is strongly inspired by the growing interest in augmented and virtual reality (AR & VR). In this context, enhancing speech and localizing sound events can be fundamental to ensure credible and safe experiences. The L3DAS23 Challenge, described in this paper, uses SoundSpaces 2.0 [9] to integrate 3D sound sources captured by first-order Ambisonics microphones and extended reality environments.

Moreover, it introduces two substantial evolutions from previous versions: a) the 3D audio recordings were not made in a physical location, but rather in 68 distinct simulated environments; b) an additional track was introduced taking into account the multimodal scenario, where information from RGB images of simulated acoustic environments can be added to the audio recordings. As a result, participants had the choice of either submitting results using only the information from the 3D audio recordings (named the audio-only track) or taking up the audio-visual track in which the audio recordings and images of the environments were made available to them. This choice was made because visual information proved to enhance the performance of deep learning models [10] and we believe it can also improve the results in the proposed 3DSE and 3DSELD tasks. The decision to participate in the audio-only track or the audio-visual one was left to the participants.

The use of a very large number of simulated acoustic environments allowed us to extend the total duration of the dataset to approximately 100 hours. We supply baseline models, 3D audio datasets for each task, and a Python-based API that facilitates the data download and preprocessing, the baseline models training and the results submission.

II. BACKGROUND

A. AMBISONICS MICROPHONES

Ambisonics is a multi-channel audio technology first introduced by M. A. Gerzon in [11]. This technology allows audio signals to be recorded, encoded and reproduced while fully preserving their spatial information. In fact, through such technology, the codification of the sound field also includes its directional characterization. Ambisonics is still one of the most complete microphone and sound reproduction systems available since it lets capture all spatial information of a sound source and permits multiple possible decodings of the signal based on the number of loudspeakers used during signal reproduction. Thus, reproduction is stereo compatible, being able to be performed with either 4 as well as 2 or 3 loudspeakers depending on specific needs. The microphones used in the L3DAS project are first-order Ambisonics microphones, which have four channels. Three of these channels correspond to three figure-of-eight capsules oriented according to the three orthogonal Cartesian axis X, Y and Z; a fourth channel (W) is associated with an omnidirectional microphone that assigns equal gain to all directions. The set of these four signals, *WXYZ*, recorded by the microphone, is called A-format. Once processed and mixed together they form the Ambisonics signal defined as B-format. The polar diagram of a first-order Ambisonics microphone is shown in Fig. 1.

Through such a structure, Ambisonics microphones allow sounds to be represented as spherical harmonics, enabling a spatially coherent representation. For this reason, this type of signal is extremely beneficial for tasks such as 3DSE and 3DSELD where the objective is to extract or recognize sound sources placed in noisy environments.

B. 3D SPEECH ENHANCEMENT

Let us consider a target signal of interest simultaneously reproduced together with other sound sources in the same environment. Its rendering will be probably unintelligible. The case in which the target signal immersed in a noisy environment is a speech signal is referred to as a cocktail party

[1] [Online]. Available: https://www.l3das.com/mlsp2021/results.html
[2] [Online]. Available: https://www.l3das.com/icassp2022/results

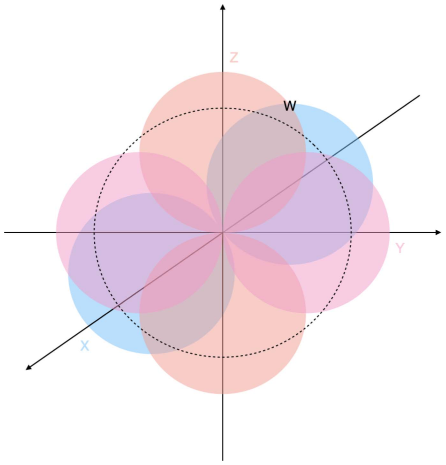

FIGURE 1. Polar diagram of the four microphones *W, X, Y, Z* of a first-order Ambisonics microphone in the 3D space.

problem. The objective of speech enhancement methods is precisely to extract the target speech signal from a sound mixture composed of ambient sounds and speech sounds of other speakers and make it intelligible. In the case of 3D speech enhancement (3DSE), the aim is to use the additional spatial information captured by the Ambisonics microphone in order to perform a more precise extraction of the target signal from the sound mixture. More formally, the 3DSE task can be explained as follows: let us consider the corrupted Ambisonics signals $x_p(n) = \sum_{i=0}^{M-1} h_{p,i}(n)s(n-i) + \epsilon(n)$, resulting from a clean speech signal $s(n)$ affected by an acoustic impulse response $\mathbf{h}_p(n) \in \mathbb{R}^M$ of M coefficients, where $p = \{W, X, Y, Z\}$ represents the four channel of the Ambisonics microphone, and additive noise $\epsilon(n)$. We seek a mapping function $f(\cdot)$ that is able to estimate the speech target signal $s(n)$ from $x_p(n)$, i.e., $\hat{s}(n) = f(x_W(n), x_X(n), x_Y(n), x_Z(n))$. In L3DAS23, the additional information, provided by an RGB image of the environment in which the sound mixture is reproduced, can be used as input to the deep learning models developed by the challenge participants. Such a situation is schematized in Fig. 2. One commonly employed strategy for conducting speech enhancement involves utilizing deep neural networks (DNNs) to estimate a time-frequency mask in the Fourier domain. This mask is designed to isolate clean speech signals from noisy spectra [12]. Cutting-edge results in Ambisonics-based 3DSE can be achieved through neural beamforming techniques such as Filter and Sum Networks (FaSNet), which are particularly well suited for low-latency scenarios. Additionally, U-Net-based approaches demonstrate competitive outcomes in both monaural [13], [14] and multichannel SE tasks [15], albeit with increased computational requirements. Alternative techniques for the SE task include recurrent neural networks (RNNs) [16], graph-based spectral subtraction [17], discriminative learning [18] and dilated convolutions [19], [20]. For the 3DSE task, we use a beamforming U-Net architecture, which provided the best metrics for L3DAS21 on the 3DSE task. In L3DAS23, we consider monaural speech signal as output.

C. 3D SOUND EVENT LOCALIZATION AND DETECTION

For the 3D Sound Event Localization and Detection (3DSELD) task, the setting is very similar to that of 3DSE: some target sounds - in this case, not necessarily speech signals - are played in a noisy environment in which other sound sources may be active simultaneously with the target sources. The goal here is to recognize the target source in the sound mixture and to be able to detect when and where it is active. In other words, in addition to correctly labeling the target sound, a model for 3DSELD must be able to provide temporal and spatial information about its specific source.

Modern deep learning methods have proven to solve this task efficiently [21]. The SELDNet [3] used as a baseline for L3DAS21 and taken up in later editions of the project is based on a convolutional-recurrent design with two distinct branches for localization and detection. An alternative version based on time convolutions has been proposed in [22]. Other solutions for this task include ensemble models [23], multi-stage training [24] and bespoke augmentation strategies [25], [26]. As a baseline for the current version of the L3DAS project, we used a variant of the SELDnet architecture, with small changes. We ported to PyTorch the original Keras implementation and we modified its structure to make it compatible with the L3DAS23 dataset. This situation is illustrated in Fig. 3. To achieve more consistent spatio-temporal descriptions of the 3D acoustic scene, we modified this network so that it could accept as additional input an RGB image of the virtual environment in which the sounds are reproduced, similar to what was done for 3DSE.

III. L3DAS23 DATASET

A. GENERAL DESCRIPTION

Each of the two tasks is supported by an appropriate dataset. The L3DAS23 datasets contain multiple-source and multiple-perspective B-format Ambisonics audio recordings. We sampled the acoustic field of multiple simulated environments, placing two first-order Ambisonics microphones in random points of the rooms and capturing up to 737 room impulse responses in each one. The datasets also contain multiple RGB pictures showing the frontal view from the main microphone. 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.

The datasets of both Task 1 (3DSE) and Task 2 (3DSELD) share a common basis: the techniques adopted for generating it. Indeed, we used Soundspaces 2.0 [9] to generate Ambisonics Room Impulse Responses (ARIRs) and images in a selection of simulated 3D houses from the Habitat - Matterport 3D Research Dataset [27]. Each simulated environment has a different size and shape, and includes multiple objects and surfaces characterized by specific acoustic properties (i.e., absorption, scattering, transmission, damping).

For the 3DSE task, the computed ARIRs are convolved with clean sound samples belonging to distinct sound classes

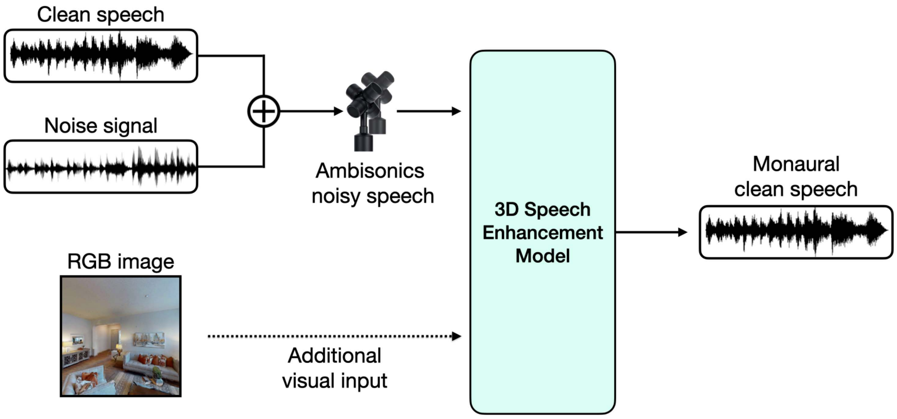

FIGURE 2. Schematic overview of the 3D speech enhancement task. Ambisonics microphones record the target speech signal along with other noisy sources in the environment. The 3DSE model recovers this target speech signal from the noisy mixture and produces a clean monaural speech signal.

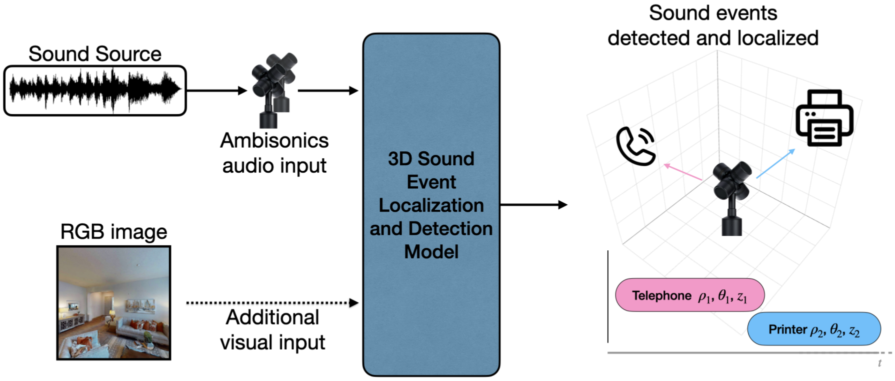
Sound Source
Ambisonics audio input
RGB image
Additional visual input
3D Sound Event Localization and Detection Model
Sound events detected and localized
Telephone ρ_1, θ_1, z_1
Printer ρ_2, θ_2, z_2
t

FIGURE 3. Schematic overview of the 3D sound event localization and detection task. Ambisonics microphones record the sound mixture of the acoustic environment and the 3DSELD model must be able to estimate the labels of the active sound sources in each time interval for the detection and their DOA to localize them: In this example, there are two active sound sources and they are identified by the labels ***telephone*** and ***printer***.

to generate the spatial sound scenes. The noise sound event dataset we used for Task 1 is the well-known FSD50 K dataset [28]. In particular, we have selected 12 transient classes, representative of the noise sounds that can be heard in an office: *computer keyboard, drawer open/close, cupboard open/close, finger-snapping, keys jangling, knock, laughter, scissors, telephone, writing, chink and clink, printer*, and 4 continuous noise classes: *alarm, crackle, mechanical fan and microwave oven*. Furthermore, we extracted clean speech signals (without background noise) from Librispeech [29], selecting only sound files up to 12 seconds.

For the 3DSELD task, the measured ARIRs are convolved with clean sound samples belonging to distinct sound classes. Sound events for Task 2 are taken again from the FSD50 K dataset [28]. We have selected 14 classes, most representative of the sounds that can be heard in an office: the 12 classes already used for 3DSE, plus *female speech* and *male speech*.

B. RECORDING PROCEDURE

We placed two Ambisonics microphones in 443 random positions of 68 houses and generated B-Format ACN/SN3D impulse responses of the rooms by placing the sound sources in random locations of a cylindrical grid defining all possible positions. Microphone and sound positions have been selected according to specific criteria, such as minimum distance from walls and objects, and minimum distance between mic positions in the same environment (RT60 between 0.3 and 0.8 s). One microphone (mic A) lies in the exact selected position, and the other (mic B) is 20 cm distant towards the x dimension from mic A. Both are shown as blue and orange dots in the topdown map in Fig. 4. The two microphones are positioned at the same height of 1.6 m, which can be considered as the average ear height of a standing person. The capsules of both mics have the same orientation.

In every room, the speaker placement is performed according to five concentric cylinders centered in mic A, where the single positions are defined following a grid that guarantees a minimum Euclidean distance of 50 cm between two sound sources placed at the same height. The radius of the cylinders ranges from 1 m to 3 m with a 50 cm step and all have 6 position layers in the height dimension at 0.4 m, 0.8 m, 1.2 m, 1.6 m, 2 m, 2.4 m from the floor, as shown in Fig. 5.

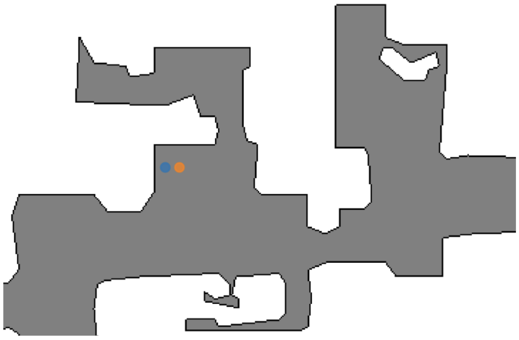

FIGURE 4. **Topdown map showing mic A (blue dot) and mic B (orange dot). Microphones can only be placed in the gray area (i.e., the area where no obstacles are located, namely the navigable area). On the contrary, sounds can be placed also outside the gray area, as long as they do not collide with objects and remain within the perimeter of the environment.**

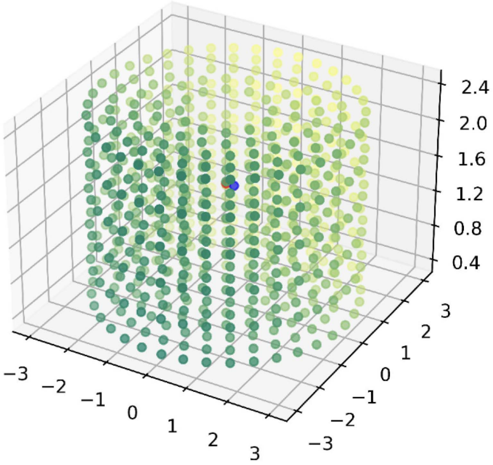

FIGURE 5. **All the concentric cylinders. The partially visible red and blu dots represent mic A and B.**

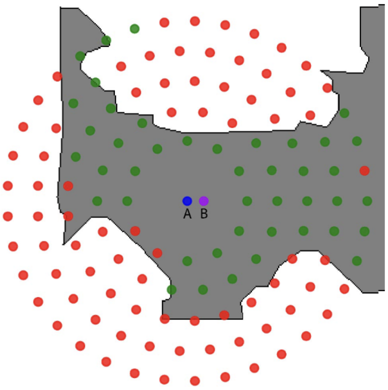

FIGURE 6. **Accepted source positions in green, discarded positions in red for one height level.**

A sound can therefore be reproduced in a room in any of the 700+ available positions (300 k+ total positions in the selected environments), to which should be subtracted all those positions that collide with objects or exceed the room space. Fig. 6 shows an example of source positioning at 0.4 m above the floor: the green dots represent accepted position (in this case, all positions within the room that do not collide with a sofa and two armchairs), while the red dots show discarded positions. No constraint is placed on the need to have the sound source in the microphone's view (and thus a direct sound). A sound could then be placed behind an obstacle (such as a column in the center of a room). SoundSpaces natively supports all these scenarios as it propagates sounds according to a bidirectional path tracing algorithm. Therefore, sound sources in SoundSpaces 2.0 are to be considered omnidirectional, meaning that sound can propagate in all directions.

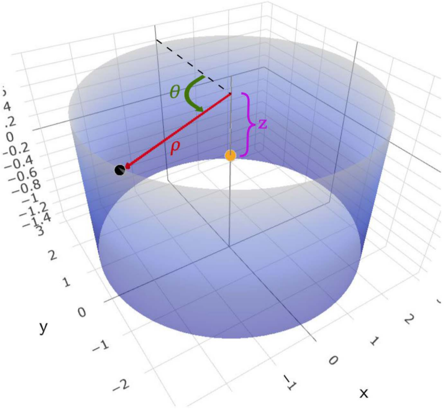

FIGURE 7. **(ρ, θ, z) for one speaker position (black dot). Mic A is represented as an orange dot.**

Each speaker position is identified in cylindrical coordinates w.r.t. microphone A by a tuple (ρ, θ, z), where ρ is in the range $[1.0, 3.0]$ (with a 0.5 step) and z in $[-1.2, +0.8]$ (with a 0.4 step). θ is in the range $[0^\circ, 360^\circ)$, with a step that depends on the value of ρ and is chosen so as to satisfy the minimum Euclidean distance between sound sources ($\theta = 0^\circ$ for frontal sounds). All labels are consistent with this notation; elevation and azimuth or Euclidean coordinates are however easily obtainable.

Fig. 7 visually represents the tuple (ρ, θ, z). The orange dot in the picture is mic A and the black dot is a speaker placed on one of the concentric cylinders. ρ represents the distance of a sound source from mic A, θ is the angle from the y-axis, and z is the height relative to mic A. Mic B is on the x-axis and thus in position $(0.2, 0, 0)$ of a local coordinate system. Being frontal to the hearer, sounds placed on the y-axis have $\theta = 0$ in the dataset. θ is therefore calculated with respect to the y-axis to comply with this principle. This has a direct impact on the way in which it is possible to switch from one notation to another.

The dataset is divided into 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, the target data varies significantly. For 3DSE, the corresponding dataset is composed of 16 kHz 16-bit AmbiX wav files. For the 3DSELD, the audio files of the dataset are 32 kHz 16-bit AmbiX wav files.

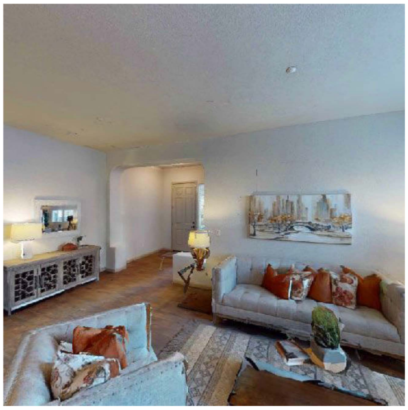

FIGURE 8. Example of a simulated view of the environment in front of the microphone.

Moreover, we created different types of acoustic scenarios, optimized for each specific task.

We split both dataset sections into: a training set (80 hours for 3DSE and 5 hours for 3DSELD) and a test set (7 hours for 3DSE and 2.5 hours for 3DSELD), paying attention to creating similar distributions. The train set of the 3DSE section is divided into two partitions: train360 and train100, and contains speech samples extracted from the correspondent partitions of Librispeech [29] (only the sample up to 12 seconds). All sets of the 3DSELD section are divided into: OV1, OV2, OV3. These partitions refer to the maximum amount of possible overlapping sounds, which are 1, 2, or 3, respectively.

C. AUDIO-VISUAL TRACKS

In addition to Ambisonics recordings, the dataset provides, for each microphone position in the rooms, an image of size 512 × 512, representing the environment in front of the main microphone (mic A). We derived these images by virtually placing a RGB sensor at the same height and orientation as mic A, and with a 90-degree field of view. An example is shown in Fig. 8. Since the microphone is placed in multiple different environments, the models will have to perform the tasks by adapting to different reverberation conditions.

Both the audio-only track and the audio-visual track were composed of two subtracks, namely 1-mic configuration and 2-mic configuration. In fact, participants could choose to use recordings from only one microphone or both of them.

IV. BASELINES

A. METRICS

For 3DSE, we adopted a metric $M_{3\text{DSE}}$ that is the combination of two distinct metrics. This evaluation metric is the combination of the short-time objective intelligibility (STOI), which estimates the intelligibility of the output speech signal, and word error rate (WER), which indicates the ratio of error in a speech-to-text transposition, computed to assess the effects of the enhancement for speech recognition purposes. We use a Wav2Vec [30] architecture pre-trained on Librispeech 960 h to compute the WER. The final metric for this task is given by:

$$M_{3\text{DSE}} = \frac{\text{STOI} + (1 - \text{WER})}{2}, \tag{1}$$

which lies in the [0,1] range, where the higher the better.

For 3DSELD, we use a joint metric for localization and detection: F-score based on the location-sensitive detection [31]. The F-score allows combining precision and recall of a model, where the precision is the number of true positives predicted by the model divided by the number of false positives plus true positives and the recall is the number of true positives divided by the number of true positives plus false negatives. The F-score is given by:

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * TP}{2 * TP + FP + FN}, \tag{2}$$

where TP is the number of true positives classified by the model, FP and FN are respectively the numbers of false positives and false negatives classified by the model. This metric considers a true positive only if a sound class is correctly predicted in a temporal frame and if its predicted location lies within a Cartesian distance from the true position of at most 1.75 m.

B. 3DSE BASELINE

For Task 1 (3DSE), we use a beamforming U-Net architecture [4], which provided the best metrics for the L3DAS21 Challenge on the 3DSE task. This network uses a convolutional U-Net [32] to estimate B-format beamforming filters. It is composed of three main modules: 1) an encoder path for extracting high-level features gradually, 2) the corresponding decoder for the reconstruction of the original size of input features from the output of the encoder, and 3) skip connections for concatenating each layer in the encoder with its corresponding layer in the decoder. The input of the model is the B-format audio signals, of dimension $\mathbb{R}^{C\times(T\times S)}$, where C is the channel number and is equal to 4 in the case of 1-mic configuration and 8 in the case of 2-mic configuration, T is the duration in seconds of the audio signal and S is the sample rate. These signals are first transported into the time-frequency domain via an STFT, resulting in a representation of dimension $\mathbb{C}^{C\times(L\times F)}$, where $L = 600$ is the number of frames and $F = 256$ is the first 256 frequency bins of the complex spectrogram. The enhancement process is performed as that of the traditional signal beamforming: we multiply the complex spectrogram of B-format noisy signal with the filters estimated by U-Net, $\mathbf{W} \in \mathbb{C}^{C\times(L\times F)}$, through element-wise multiplication, and then sum the result over the channel axis to estimate a single-channel enhanced complex spectrogram, $\hat{\mathbf{S}} \in \mathbb{C}^{L\times F}$. In the end, the iSTFT is performed to obtain the enhanced time-domain signal.

With this baseline model, we obtained a baseline test metric for Task 1 of 0.557, with a WER of 0.57 and an STOI of 0.68. We adapted this model to the audiovisual task by using a CNN-based extension part whose output features are concatenated along the filter dimension with those generated by the encoder part of the U-Net. The visual features allow a sensible decrease in the number of epochs required to achieve results comparable to those of the audio-only track.

C. 3DSELD BASELINE

For Task 2 (3DSELD), instead, we used a variant of the SELDnet architecture [3], with small changes with respect to the one used in the L3DAS22 Challenge. We ported to the PyTorch language the original Keras implementation and we modified its structure in order to make it compatible with the L3DAS23 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 into 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.

This network obtains a baseline test F-score of 0.147, with a precision of 0.176 and a recall of 0.126. We adapted this model to the audiovisual task by using a CNN-based extension part whose output features are concatenated to the ones of our augmented 3DSELD just before passing them to the two separate branches. This simple change resulted in a 7% improvement in the F-score (0.158), with a precision of 0.182 and a recall of 0.140.

V. CHALLENGE RESULTS

Among the challenge participants, those who presented models capable of beating the proposed baselines were: SEU Speech, JLESS, CCA Speech for the 3DSE and JLESS and NERCSLIP-USTC for the 3DSELD. The fifth best-performing team, although below the baseline, is Speech-Lab410 with a model for 3DSE. The main contributions of these teams are briefly summarised below:

1) SEU Speech proposed a dual-path convolutional recurrent network with group attention for 3DSE [33]. The model is structured as a convolutional encoder-decoder with frequency-time blocks based on group attention introduced in the middle. The encoder extracts the local representation from the spectrogram, the correlation between frequency and time axes are captured through groups of time-frequency processing modules, and the key information in the feature flow is extracted by the group attention.
2) JLESS team proposed a two-stage system based on DPRNN and U-Net for the 3DSE task and a Conformer-based system for the 3DSELD task [34]. This is the only team to have participated in both tasks of the challenge and also to have developed a model for the audio-visual track as part of the 3DSELD. In the two-stage U-Net for the audio-only 3DSE, the amplitude of the STFT is fed into the network for estimating the mask, and the phase of the mixed signal is used for speech reconstruction. They add 4 DPRNN modules between the encoder and decoder of U-Net for transient modeling and extraction of dynamic voice information. The STFTs of the multi-channel speech signals are first fed into the U-Net with DPRNN, and the estimated STFT is formed using beamforming. Then, the estimated STFT is sent into the second U-Net without DPRNN for the estimating finer mask. Sigmoid is used to activate the mask of the output layer; after that, the masked estimation results of the first level are connected with the masked estimation results of the second level by residual. Regarding the Conformer-based SELD system, the log-Mel and intensity vectors are calculated for both mic-A and mic-B audio signals. Then, the time difference of arrival (TDOA) of 2 mics is computed using kernel density estimator (KDE) theory [35]. For the visual signal, images are resized into 224 × 224 px and normalized for fine-tuning the pretrained model. A Res-Conformer-based SELD model is adapted in the audio-visual scene. Audio features are fed into four residual convolution blocks following two Conformer encoder blocks. Images are fed into Resnet18 with pre-trained weights. The embedding of images is then concatenated with audio features before the last output layer. The authors applied some data augmentation methods, such as cutout, frequency shift, time shift, mixing, brightness, hue, saturation, and contrast jitter.
3) CCA Speech team developed a stream attention-based U-Net to remove background noise and reverberation for 3DSE [36]. Their model consists of three parts, encoder, decoder, and channel fusion module. They proposed stream attention to fuse various channels in order to fully use the information between channels and this is done also in the encoder stage. Key, query, and value are generated by three convolutional networks. A softmax function is applied to the last dimension of the product of the key and query. The decoder part is composed of only convolutional blocks, while an LSTM block is used in the encoder part.
4) NERCSLIP-USTC proposed a method based on the combinations of ResNet and Conformer architectures to model both local and global patterns [37]. ResNet blocks are used to extract high-dimension feature

TABLE 1 Results of Task 1 Participants

Rank	Team Name	WER ↓	STOI ↑	T1 Metric ↑
1	SEU Speech	0.101	0.902	0.901
2	JLESS	0.174	0.836	0.831
3	CCA Speech	0.240	0.831	0.796
-	Baseline	0.567	0.673	0.553
4	SpeechLab410	0.643	0.608	0.483

TABLE 2 Results of Task 2 Participants

Rank	Team Name	Precision ↑	Recall ↑	T2 Metric ↑
1	JLESS	0.288	0.204	0.239
2	NERCSLIP-USTC	0.275	0.216	0.242
-	Baseline	0.182	0.140	0.158

representation from the input features, while Conformer blocks are effective to extract local fine-grained features and long-range global information, respectively. The authors also adopted several data augmentation techniques (SpecAugment, Mixup and ACS) to expand the official dataset.

5) SpeechLab410 proposed a refine-beamfomer system to enhance 3D speech signals. The beamforming network consists of two U-Net beamforming networks. In the first stage, they employed a neural beamforming network to initially enhance the 3D speech signal. Then the generation characteristics of a diffusion model were utilized to further enhance the speech signal. The two stages of this enhancement model were trained separately.

Tables 1 and 2 show the results obtained by the participants on the test set. For Task 1 (3DSE) the models had to predict monaural sound waveforms, containing the enhanced speech signals extracted from the multichannel noisy mixtures, with a sampling rate of 16 kHz. For Task 2 (3DSELD) the models were expected to predict the spatial coordinates and class of the sound events active in a multichannel audio mixture. Such information had to be generated for each frame in a discrete temporal grid with 100-millisecond non-overlapping frames. Each submitted file for this task was a csv table listing, for every time frame, the class and spatial coordinates of each predicted sound event. All participants worked with the 2-mic configuration, so the results shown in Tables 1 and 2 refer to the 2-mic configuration.

Participants had to submit the only results obtained for the blind test. The submission had to contain up to two zip archives, one for the audio-only track and one for the audio-visual track, enclosing two separate folders for the challenge tasks, named task1 and task2. From these results, we derived an overall ranking of the participants as reported in Table 3. All these teams were allowed to submit their work as a 2-page paper to ICASSP 2023.

TABLE 3 Final Rankings of the L3DAS23 Challenge

Ranking	Team Name	Task
1st	SEU Speech	3DSE
2nd	JLESS	3DSE + 3DSELD
3rd	CCA Speech	3DSE
4th	NERCSLIP-USTC	3DSELD
5th	SpeechLab410	3DSE

VI. CONCLUSION

This paper presented the details of the L3DAS23 Signal Processing Grand Challenge at ICASSP 2023, including: the L3DAS23 dataset, the challenge tasks, the baseline models and the results obtained by the winning participants. The current version of the L3DAS project introduces the use of visual information for 3DSE and 3DSELD tasks, given the growing and stimulating interest in AR & VR. The introduction of visual input extracted from analyzed acoustic environments, whether simulated or not, can drastically benefit the research in the field of 3D audio signal processing. For this reason, future work of the L3DAS team will primarily involve the study of new methods to improve the interaction of visual information with Ambisonics audio signals, in order to further improve the results obtained with this challenge. Then, we plan to incorporate new 3D acoustic scenarios, diverse microphone configurations, and novel tasks that could be of great relevance in the context of augmented and virtual reality applications. Moreover, different tasks than 3DSE and 3DSELD will be definitely taken into account, together with the collection of real-recorded data.

ACKNOWLEDGMENT

The authors support the dataset download and the use of the baseline models via extensive instructions provided on the official GitHub repository at https://github.com/l3das/L3DAS23. For more comprehensive information and in-depth details about the challenge, we invite the reader to visit the L3DAS Project website at http://www.l3das.com/icassp2023.

REFERENCES

[1] E. Guizzo et al., “L3DAS21 challenge: Machine learning for 3D audio signal processing,” in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process.*, 2021, pp. 1–6.

[2] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S. -C. Liu, “FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing,” in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 260–267.

[3] 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. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Mar. 2019.

[4] X. Ren et al., “A neural beamforming network for b-format 3D speech enhancement and recognition,” in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process.*, 2021, pp. 1–6.

[5] H. R. Guimarães, W. Beccaro, and M. A. Ramírez, “Optimizing time domain fully convolutional networks for 3D speech enhancement in a reverberant environment using perceptual losses,” in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process.*, 2021, pp. 1–6.

[6] E. Guizzo et al., “L3DAS22 challenge: Learning 3D audio sources in a real office environment,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 9186–9190.

[7] Y.-J. Lu et al., “Towards low-distortion multi-channel speech enhancement: The ESPNET-SE submission to the L3DAS22 challenge,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 9201–9205.

[8] J. Hu et al., “A track-wise ensemble event independent network for polyphonic sound event localization and detection,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 9196–9200.

[9] C. Chenet al., “SoundSpaces 2.0: A simulation platform for visual-acoustic learning,” *Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2022.

[10] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, “Deep audio-visual learning: A survey,” *Int. J. Automat. Comput.*, vol. 18, pp. 351–376, 2020.

[11] M. A. Gerzon, “The design of precisely coincident microphone arrays for stereo and surround sound,” *J. Audio Eng. Soc. Conv.*, 1975.

[12] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[13] H. R. Guimarães, H. Nagano, and D. W. Silva, “Monaural speech enhancement through deep Wave-U-Net,” *Expert Syst. Appl.*, vol. 158, 2020, Art. no. 113582.

[14] C. Macartney and T. Weyde, “Improved speech enhancement with the Wave-U-Net,” 2018, *arXiv:1811.11307*.

[15] A. Bosca, A. Gu’erin, L. Perotin, and S. Kitic, “Dilated U-net based approach for multichannel speech enhancement from first-order ambisonics recordings,” in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 216–220.

[16] P.-S. Huang, M. Kim, M. A. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1562–1566.

[17] X. Yan, Z. Yang, T. Wang, and H. Guo, “An iterative graph spectral subtraction method for speech enhancement,” *Speech Commun.*, vol. 123, pp. 35–42, 2020.

[18] C. Fan, B. Liu, J. Tao, J. Yi, and Z. Wen, “Discriminative learning for monaural speech separation using deep embedding features,” in *Proc. Interspeech*, 2019.

[19] Y. Luo and N. 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, Aug. 2019.

[20] O. Yazdanbakhsh and S. Dick, “Multivariate time series classification using dilated convolutional neural network,” 2019, *arXiv:1905.01697*.

[21] A. Politis, S. Adavanne, and T. Virtanen, “A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection,” in *Proc. Workshop Detection Classification Acoustic Scenes Events*, 2020

[22] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif, and B. Yang, “SELD-TCN: Sound event localization & detection via temporal convolutional networks,” in *Proc. 28th Eur. Signal Process. Conf.*, 2020, pp. 16–20.

[23] S. P. Chytas and G. Potamianos, “Hierarchical detection of sound events and their localization using convolutional neural networks with adaptive thresholds,” in *Proc. Workshop Detection Classification Acoustic Scenes Events*, 2019.

[24] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, “Polyphonic sound event detection and localization using a two-stage strategy,” in *Proc. Workshop Detection Classification Acoustic Scenes Events*, 2019.

[25] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, “First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation,” in *Proc. Workshop Detection Classification Acoustic Scenes Events*, 2019.

[26] P. Pratik, W. J. Jee, S. Nagisetty, R. Mars, and C. S. Lim, “Sound event localization and detection using CRNN architecture with mixup for model generalization,” in *Proc. Workshop Detection Classification Acoustic Scenes Events*, 2019.

[27] S. K. Ramakrishnan et al., “Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI,” *Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021.

[28] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, “FSD50K: An open dataset of human-labeled sound events,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2022.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.

[30] A. Baevski, H. Zhou, A. R. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020..

[31] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, “Joint measurement of localization and detection of sound events,” in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 333–337.

[32] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput. Assist.-Intervention*, 2015.

[33] J. Cheng, C. Pang, R. Liang, J. Fan, and L. Zhao, “Dual-path dilated convolutional recurrent network with group attention for multi-channel speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[34] J. Bai, S. W. Huang, H. Yin, Y. Jia, M. Wang, and J. Chen, “3D audio signal processing systems for speech enhancement and sound localization and detection,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[35] V. V. Reddy, A. W. H. Khong, and B. P. Ng, “Unambiguous speech DOA estimation under spatial aliasing conditions,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2133–2145, Dec. 2014.

[36] H. Wang, Y. Fu, J. Li, M. Ge, L. Wang, and X. Qian, “Stream attention based U-Net for L3DAS23 challenge,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[37] H. Yan, H. Xu, Q. Wang, and J. Zhang, “The NERCSLIP-USTC system for the L3DAS23 challenge Task2: 3D sound event localization and detection (SELD),” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

Open Access funding provided by ‘Università degli Studi di Roma “La Sapienza” 2’ within the CRUI CARE Agreement