DOI: https://doi.org/10.1109/taslp.2023.3284525

Sound Events Localization and Detection Using Bio-Inspired Gammatone Filters and Temporal Convolutional Neural Networks

Karen Rosero, *Graduate Student Member, IEEE*, Felipe Grijalva, *Senior Member, IEEE*, and Bruno Masiero, *Member, IEEE*

***Abstract*—The auditory brain circuits are biologically constructed to recand localize sounds by encoding a combination of cues that help individuals interpret sounds. The development of computational methods inspired by human capacities has established opportunities for improving machine hearing. Recent studies based on deep learning show that using convolutional recurrent neural networks (CRNNs) is a promising approach for sound event detection and localization in spatial sound. Nevertheless, depending on the sound environment, the performance of these systems is still far from reaching perfect metrics. Therefore, this work intends to boost the performance of state-of-the-art (SOTA) systems by using bio-inspired gammatone auditory filters and intensity vectors (IVs) for the acoustic feature extraction stage, along with the implementation of a temporal convolutional network (TCN) block into a CRNN model, to capture long term dependencies. Three data augmentation techniques are applied to increase the small number of samples in spatial audio datasets. The mentioned stages constitute our proposed Gammatone-based Sound Events Localization and Detection (G-SELD) system, which exceeded the SOTA results on four spatial audio datasets with different levels of acoustical complexity and with up to three sound sources overlapping in time.**

***Index Terms*—Acoustical signal processing, acoustic scene analysis, sound event localization and detection, reverberation, spatial sound, deep learning, dilated convolutions.**

I. Introduction

THE faculty of detecting and localizing sounds in an environment imparts survival advantages and sensitive communication skills for natural human interaction with the surroundings [1]. The human auditory system processes sound arriving in our ears from sources distributed all over space. If we were to rely solely on our ears to recognize an unfamiliar environment, our auditory system would first recognize familiar sounds and then compare them with how those sounds were perceived in other familiar environments. This activity that seems so natural to us can be challenging for computers. Furthermore, considering that our natural listening is three-dimensional, why is it that most of the audio signals we usually listen to do not maintain the spatial information of the sound field? Based on this premise, the goal of spatial audio is to recreate the listener's perception in the real world, maintaining all the characteristics that allow our auditory system to process the content and direction of sound sources. In that sense, this area of machine hearing considers the use of spatial audio recordings, along with systems inspired by human hearing, to enhance the detection and localization of sounds. Several applications relate to machine hearing, such as intelligent meeting rooms [2], helping deaf people to know the sounds of their environment [3], [4], and acoustic monitoring of urban environments or wildlife [5], [6].

The sound events localization and detection (SELD) task implies multi-class sound events detection (SED) and sound source localization (SSL) of multiple directions of arrival (DOAs) with respect to the microphone. Regarding DOA estimation techniques, we found systems based on the time difference of arrival (TDOA) [7], the steered-response power (SRP) [8], the generalized side lobe canceller [9], and beamforming techniques such as compressive beamforming [10], and the minimum variance distortionless response (MVDR) beamforming [11]. These methods vary in algorithmic complexity, compatibility with microphone arrangements, and assumptions regarding the acoustic scenario. To overcome these complications and estimate the number of active sources directly from the input features, authors in [12] studied the use of deep neural networks (DNNs) for direction of arrival (DOA) estimation.

Recent studies have accomplished SELD with a multi-task perspective. In [13], the spectrogram is used as an intermediate representation of audio, which is processed by four convolutional layers and three fully-connected (FC) layers. The SELDnet system [14] also uses the spectrogram as input, but it extracts the phase and magnitude components as separate features. The SELDnet architecture comprises a convolutional neural network (CNN) with three convolutional blocks for feature extraction

Manuscript received 17 September 2022; revised 7 April 2023; accepted 1 June 2023. Date of publication 9 June 2023; date of current version 21 June 2023. This work was supported in part by the São Paulo Research Foundation (FAPESP) under Grants 2017/08120-6 and 2019/22945-3 and in part by the Universidad San Francisco de Quito through the Poli-Grants Program under Grant 17993. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Romain Serizel. *(Corresponding author: Karen Rosero.)*

Karen Rosero is with the School of Electrical and Computer Engineering, University of Campinas, Campinas 13083-852, Brazil, and also with the Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75080-3021 USA (e-mail: karenrosero14@ieee.org, kgr220000@utdallas.edu).

Felipe Grijalva is with the Colegio de Ciencias e Ingenierías "El Politécnico", Universidad San Francisco de Quito, Quito 170901, Ecuador (e-mail: fgrijalva@usfq.edu.ec).

Bruno Masiero is with the School of Electrical and Computer Engineering, University of Campinas, Campinas 13083-852, Brazil (e-mail: masiero@unicamp.br).

Our Python code is available in this repository.

Digital Object Identifier 10.1109/TASLP.2023.3284525

and dimensionality reduction. It also established the use of a recurrent neural network (RNN) based on gated recurrent units (GRUs) to learn temporal context information from the output of the convolutional blocks. Then, separate branches containing FC layers perform the classification and localization tasks. Based on SELDnet, an improved framework was presented as a baseline for the Task 3 of the DCASE2021 Challenge [15], which objective was the localization and detection of sound events in multichannel audio. This SELDnet-DCASE2021 version receives log-Mel spectrograms and intensity vectors (IVs) as intermediate audio representations. Instead of using separate branches for each task, it adopted the Activity-Coupled Cartesian Direction of Arrival (ACCDOA) representation [16] to unify both classification and localization losses.

In contrast, concerned about an efficient implementation of these types of systems on embedded hardware, the SELD-TCN system [17] proposed to substitute the recurrent blocks of SELDnet with temporal convolutional network (TCN) blocks containing dilated convolutions that capture long-term dependencies of data. TCNs also avoid the sequential computing of the input by processing the whole sequence in parallel via convolutions. The SELD-TCN framework maintains the original SELDnet characteristics regarding the intermediate audio representations and the separate output branches.

A modified version of SELDnet was implemented in PyTorch as a baseline for Task 2 of the L3DAS21 Challenge [18], which also aims to achieve the SELD problem. The phase and magnitude components of the spectrogram were used as features. As in the SELDnet system, the phase is expected to contain information on location and the magnitude of detection and classification. Regarding the architecture, one additional convolutional block and one more recurrent block were added to augment the network's capacity, while the two branches' output structure of SELDnet was maintained. The ability to detect multiple sound sources of the same class that overlap in time was also implemented through an augmented output matrix.

Regarding intermediate representations of audio, the Mel auditory model has been used in Automatic Speech Recognition (ASR) [19] and SELD [15] tasks. However, it still presents limitations in the attempt to model the human ear. By contrast, gammatone filter impulse responses were obtained from measures on the basilar membrane of small mammals. Moreover, applying a gammatone filter bank to the spectrogram has shown to be more robust against ambient noise in acoustic event monitoring compared with Mel-scale filter bank representations [20], [21]. The gammatone filter bank has also shown good performance in automatic audio captioning systems [22] and active noise control systems [23]. For this reason, a gammatone filter bank is explored in this work to obtain a log-gammatone spectrogram that will be used as the intermediate audio representation, along with the IVs.

We also propose a novel deep learning architecture for the SELD task, which joins the independent improvements proposed by the state-of-the-art (SOTA) systems. First, we question the plain inclusion of new convolutional and recurrent blocks aiming to improve the performance of the SELD system. Instead, we propose to include in the middle of the CNN and RNN blocks a TCN block that captures long-term dependencies and, at the same time, continues with the identification of core features by using dilated convolutions. Additionally, we adopt the single branch ACCDOA representation and modify it to detect multiple sound sources of the same class overlapping in time. The mentioned stages constitute our proposed Gammatone based - Sound Events Localization and Detection system, which will be referred to as the G-SELD system.

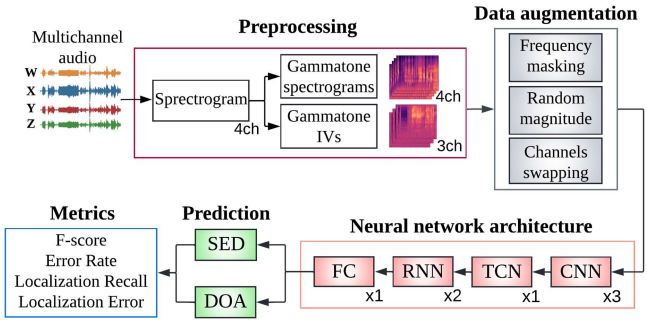

Fig. 1. Flowchart of the G-SELD system. In the preprocessing stage, the four channels spectrogram is processed into four log-gammatone spectrograms and three IVs. Regarding the neural network architecture, the number of blocks are depicted.

Since creating labeled datasets of spatial audio for the SELD task is a demanding and maybe imprecise process, the datasets usually contain less than a thousand samples. That restriction hinders the generalized learning of supervised deep learning approaches that require as many data samples as possible to be trained. Therefore, three suitable methods of data augmentation for spatial audio are also explored in this work: frequency masking, channel swapping, and random magnitude.

This article is organized as follows: Section II explains each stage of the proposed methodology. In Section III, we present the results of our G-SELD system by evaluating it on different polyphony levels and under different sound scene conditions. We also present an ablation study for each stage of the G-SELD system. Finally, Section IV presents the conclusions of this work.

II. Methodology

A methodology overview of the proposed G-SELD system is shown in Fig. 1. First, each audio input channel is processed into a spectrogram, from which the log-gammatone spectrogram and the IVs are obtained. The metadata is preprocessed to support the detection of up to three simultaneous sound events of the same class. Later, in the data augmentation stage, two new feature samples are generated using three techniques: frequency masking, random magnitude, and channel swapping. Subsequently, the original and the synthetically augmented samples input the deep learning G-SELD architecture, formed by a single branch containing CNN, TCN, RNN and FC blocks. The predicted classes and locations are obtained by processing the network output vector. Finally, we adopt four metrics used in analogous works to evaluate the system's performance. Further details of each stage will be explained in the following sections.

A. Datasets

The spatial audio datasets used in this work are provided in the Ambisonics format, which relies on the spatial decomposition of the sound field in the orthogonal basis of spherical harmonic functions [24]. The First Order Ambisonics (FOA) B-format consists of four signals that encode the overall sound in terms of pressure and particle velocity components. This format contains a signal **W** that represents an omnidirectional pattern and three orthogonal signals **X, Y, Z** aligned with the Cartesian coordinate axes.

The FOA spatial audio datasets used to evaluate the G-SELD system were selected due to their inherent diverse acoustic characteristics. These include anechoic and reverberant audio scenes, synthetic and measured impulse responses (IRs), background noise, and interference sound. Our objective is to evaluate the performance of the G-SELD system across various levels of difficulty that correspond to the level of effort required by humans to detect and localize sounds. For instance, we expect that the performance of G-SELD will be better in an environment without background noise than in a noisy scenario. All datasets contain FOA B-format audio files in which the sound events are spatially positioned, accompanied by a set of accurate metadata that includes time-boundaries, DOA, and sound type. Moreover, all datasets contain at least one subset in which up to three sound events may overlap in time. In order to provide a better understanding of why each dataset plays a significant role in the robust evaluation of the G-SELD system, we will briefly describe the acoustical and technical characteristics of each dataset.

1) ANSYN: The *TUT Sound Events 2018-Ambisonic, Anechoic and Synthetic Impulse Response (ANSYN)* dataset contains static point sources associated with a spatial coordinate described in terms of azimuth, elevation, and distance. The anechoic environment was synthesized using artificial IRs, and the individual sounds were extracted from Task 2 of the DCASE 2016 Challenge [25], which objective was the detection of sound events in synthetic audio. The sounds were recorded in residential areas and home scenes, from which these 11 classes of sounds were selected: speech, laughter, cough, clear throat, door slam, page-turning, phone ringing, keyboard sounds, keys dropping, door knock, and drawing sound. This dataset is divided into three subsets: OV1, which consists of audio samples with no sound events overlapping in time, and the OV2 and OV3 subsets, in which up to two and three sound events overlap in time can be found, respectively. Each subset contains three cross-validation splits, all with 240 development samples and 60 evaluation samples, summing a total of 900 audio files sampled at 44.1 kHz for 30 s. The whole dataset containing OV1, OV2, and OV3 subsets consists of 2700 audio samples with their corresponding metadata files.

2) REAL: The *TUT Sound Events 2018-Ambisonic, Reverberant and Real-life Impulse Response (REAL)* dataset contains static points sources positioned in a reverberant three-dimensional scene. The IRs were collected from a university corridor with classrooms. The isolated real-life sound events were extracted from the Urban-Sound8k dataset [26], from which eight classes of urban environment sounds are used: car horn, dog barking, drilling, engine idling, gunshot, jackhammer, siren, and street music. Air conditioner and children playing sounds are used as background noises. The subsets' distribution is the same as ANSYN dataset, as well as the sampling frequency of 44.1 kHz and the duration of 30 s.

3) L3DAS21: We use data related to the 3D SELD task of the L3DAS21 dataset [18], which IRs were recorded with two FOA microphones positioned in a small reverberant office environment equipped with typical office furniture. In this project, only the samples related to the microphone placed exactly in the center of the room are used. Fourteen clean types of sounds typical of an office environment were extracted from Librispeech [27] and FSD50K [28] datasets (computer keyboard, drawer open/close, cupboard open/close, finger-snapping, keys jangling, knock, laughter, scissors, telephone, writing, chink and clink, printer, female speech, and male speech). Four background noises (alarm, crackle, mechanical fan, and microwave oven) were selected from FSD50 K. The dataset contains four training splits, summing 600 audio samples and one evaluation split with 150 audio samples. The duration of each audio is 60 s, and the sampling frequency is 32 kHz. Each subset contains the same amount of files associated with one, two, and three sound events overlapping in time.

4) DCASE2021: The DCASE2021 dataset [15] was provided for the Task 3 of the DCASE2021 Challenge. Besides containing time-overlapping sound events, this dataset includes directional interference events, moving sound sources, and an additional layer of background noise in all samples. The IRs were collected in 13 rooms with different reverberant conditions, in which circular and linear trajectories were recorded, changing the fonts' heights, distances, and elevations. The ambient noise of each room was recorded during 30 min, and later, 1 min-duration segments were added to every spatial audio file with varying signal-to-noise ratios (SNR) ranging from 30 dB to 6 dB. Twelve classes of isolated sound events (alarm, crying baby, crash, barking dog, female scream, female speech, footsteps, knocking on the door, male scream, male speech, phone, piano) were extracted from the NIGENS general sound events' database [29], from which two additional classes (running engine and burning fire) were used as interference events, out of the target classes. The available development dataset consists of six folds, four for training, one for validation, and one for testing. Each split contains 100 one-minute-long audio samples with a sampling rate of 24 kHz.

B. Preprocessing

Each channel of the audio wave files is scaled from a 16-bit pulse-code modulation (PCM) to a float vector with values ranging from -1.0 to 1.0. Then, a spectrogram is computed for each Ambisonic B-format channel with a 40 ms Hanning window, 20 ms hop length, and a 1024-point fast-Fourier transform (FFT) with 512 frequency bins. Two intermediate representations are extracted from the multichannel spectrogram: four log-gammatone spectrograms that yield frequency information at different time instances, and three acoustic IVs that express net acoustic energy flux (Fig. 2).

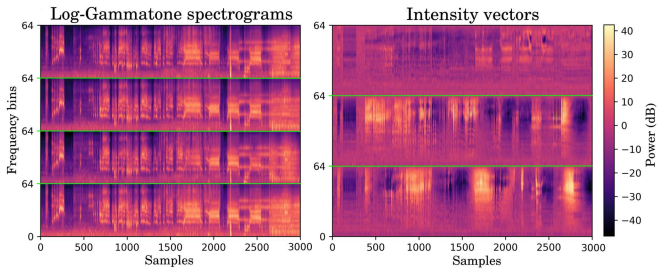

Fig. 2. Intermediate representations of multichannel audio. The log-gammatone spectrograms of the FOA channels (left) and the three IVs (right) are shown. The vertical concatenation of the channels is employed just for illustration.

The frequency axis of both intermediate representations or *features* are wrapped into the 64 bands of the gammatone filter bank. The lowest and highest frequencies of the filter bank were set to 0 Hz and half of the Nyquist frequency, respectively. Finally, the feature map has a dimension of $7 \times T \times 64$, where 7 is the number of input channels, T represents the number of time frames for each dataset, and 64 is the number of frequency bins.

The metadata files contain information about every sound source in the recording, such as onset and offset times in seconds, class, and localization, which can be expressed in spherical or Cartesian coordinates. The preprocessing stage of SOTA systems such as SELDnet and SELD-TCN is restricted when more than one sound event of the same class is overlapping in time. In contrast, inspired by the L3DAS21 framework, in the proposed G-SELD system, we overcome the location overwriting for the second or third sound event of the same class, bringing the possibility of, for example, localizing up to three people simultaneously speaking. We process one annotation every 100 ms, which also allows us to track moving sound sources.

C. Data Augmentation

Considering the reduced number of samples in spatial audio datasets, we use three data augmentation techniques in the spectral domain features: frequency masking, FOA channels swapping, and random magnitude.

Frequency masking was proposed in [30] to be applied in one-channel Mel spectrograms for ASR. In this project, we adapt it to mask a maximum of F consecutive frequency bins of the log-gammatone spectrograms and the IVs every 100 ms, maintaining the same instantaneous mask for the seven channels of the feature map. We compute two augmented outputs: one with $F = 16$ and just one mask per time frame, and a second with two masks per time frame and $F = 8$ aiming to position two mask blocks in different frequency bins of the same time frame. The initial frequency bin, as well as the number of masked bins, are randomly selected. Fig. 3 shows an example of frequency masking with one mask per time frame. Note the different sizes of masks in different time frames. In this technique, the annotations did not need to be modified.

The FOA channels swapping strategy was initially proposed in [31] for increasing the number of DOAs of the sound events

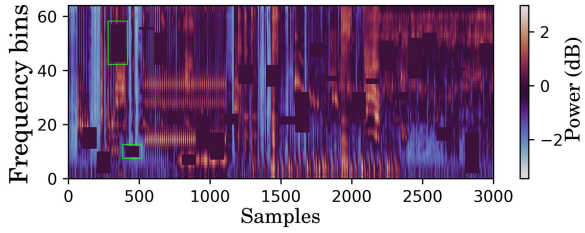

Fig. 3. Example of frequency masking applied on one channel of a log-gammatone spectrogram. One mask is allowed per time frame, and $F = 16$.

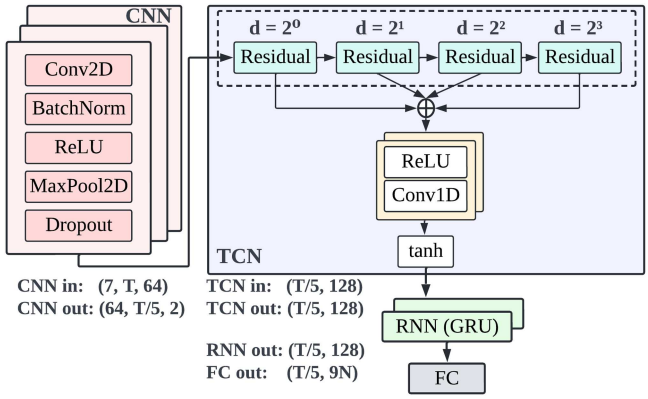

Fig. 4. G-SELD neural network architecture. T represents the time frames, and N is the number of classes for each dataset.

contained in the dataset. As proposed in [32], the input feature channels that correspond to the **X**, **Y**, **Z** FOA signals can be randomly swapped, and their signs randomly reversed in order to change the direction of the sound events. Due to its omnidirectional nature, the **W** channel is not modified with this technique. Considering the correlation between the log-gammatone spectrograms and the IVs, we equally transformed both. Two modified feature samples are computed for each original sample, and the original annotations were transformed.

The third data augmentation technique is inspired by the random magnitude technique proposed in [32], which modifies the overall volume of an audio sample by adding a random scalar value to the log-Mel spectrograms. We modify the magnitude of the log-gammatone spectrogram by adding random variables sampled from a normal distribution with a mean equal to 0 and a standard deviation of 0.02. For this technique, the intensity vectors and the annotations are not modified.

D. G-SELD Model

As depicted in Fig. 4, the G-SELD model receives a feature map with dimension $7 \times T \times 64$, which passes by three CNN blocks responsible for identifying translation invariant patterns and reducing the dimension of the input data by maintaining the most important features. The CNNs are expected to learn inter-channel features from the four gammatone spectrogram channels and the three channels of IVs. Each convolutional block contains 2D convolutions with 64 filters, which kernel size is 3×3 and their stride size is 1×1. The 2D max-pooling values

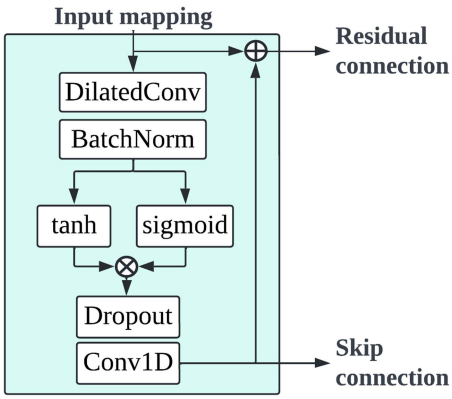

Fig. 5. Residual block used in the TCN. Modified from [17].

for the three convolutional blocks are 5×4, 1×4, and 1×2, respectively, and the dropout is 0.05. The first and second axis of the CNN blocks' output are permuted, aiming to let the time dimension T in the first position since the TCN block processes a sequence.

Originally proposed in [33] and later adapted for audio signals in [17], the use of dilated convolutions embedded in a residual block flexibly expands the feature map. As shown in Fig. 5, the output of the stacked layers is added to the input mapping using a shortcut connection, which then passes to the next residual block. By using this residual learning framework, the layers stacked in the residual block are optimized to learn the residual mapping instead of an individual mapping after each layer [34], [35]. Each residual block has a dilated convolutional layer with 256 filters of kernel equal to 3, followed by a batch normalization layer and the sigmoid and $\tanh$ activation functions. Regarding the purpose of the activation functions, the sigmoid function controls the flow of information through the input mapping, behaving like a gate, with its output ranging between 0 and 1. In contrast, the $\tanh$ function regulates the network values, preventing excessively large or small values that could hamper the network's learning process. Therefore, the $\tanh$ function ensures that the values range between -1 to 1. The activation function outputs pass by an element-wise multiplication, and later, the dropout rate is set to 0.5. Lastly, aiming to ensure the same dimension before adding the residual connection, a 1D convolutional layer is used to guarantee the exact shape of the input vector and the skip connection vector.

As shown in the TCN block of Fig. 4, we use four residual blocks with dilation factors that change in the range of $[2^0, 2^1, 2^2, 2^3]$. Then, the TCN output passes by two recurrent blocks with 128 GRUs, as used in SELDnet. Finally, a fully-connected layer reduces the dimension to a suitable prediction vector.

E. Prediction

The ACCDOA algorithm unifies the SED and SSL losses into a single weighted regression loss, avoiding the use of separate branches of dense layers for each subtask [16]. We adapted this algorithm to deal with three time-coincident sound events of the same class with different DOAs. The prediction vector contains up to three estimated locations in Cartesian coordinates for each possible class. However, it does not directly contain the probability of occurrence for each class. Therefore, class prediction is obtained from the vector norm or magnitude of each location estimator as $\sqrt{x^2 + y^2 + z^2}$, from which every magnitude greater than 0.5 is considered an active sound event of each class. Summing up, the predictor direction indicates the DOA, and its length constitutes the probability of occurrence of the corresponding sound class. Finally, the estimated DOAs are transformed from Cartesian into spherical coordinates to be consistent with the metrics computation.

F. Experimental Setup

We assessed three aspects of the G-SELD system, which include: 1) its ability to perform at varying levels of polyphony, 2) its ability to perform in different sound environments with varying levels of complexity, and 3) an ablation study to evaluate the individual contribution of each proposed improvement.

The evaluation of our G-SELD system followed the same experimental setup in all tested cases. The Adam optimizer [36] was used with an initial learning rate of 0.001, the batch size was set to 64, and the maximum number of training epochs was 100. The G-SELD system was developed using the TensorFlow framework and a computer with a $9th$ generation Intel Core $i7$ processor equipped with an NVIDIA Titan V GPU.

G. Evaluation Metrics

We adopted the metrics proposed in [37], which were used in the DCASE2021 Challenge [15]. It formulates *location-sensitive detection metrics* that evaluate sound event detection with specific spatial error allowance, and *class-sensitive localization metrics* that measure the spatial error between sound events with the same classification. The spatial error is calculated as the angular distance between reference and predicted DOAs, for which a threshold of 20° is allowed. The Error Rate (ER) and the F_{score} (F1) are location-sensitive detection metrics, whereas the Localization Recall (LR) and the Localization Error (LE) are class-sensitive localization metrics. A combination of the metrics ($\text{SELD}_{\text{score}}$) is used as the early stopping parameter and is defined as:

$$\text{SELD}_{\text{score}} = \frac{\text{ER} + \left(1 - \frac{F_{\text{score}}}{100}\right) + \frac{\text{LE}}{180} + \left(1 - \frac{\text{LR}}{100}\right)}{4} \quad (1)$$

The ideal metrics are: ER$= 0$, $F_{score} = 100\%$, LR$= 100\%$, LE$= 0^{\cdot}$, and $\text{SELD}_{\text{score}} = 0$. Finally, the early stopping process monitors the $\text{SELD}_{\text{score}}$ with a patience of 30 epochs. The values presented in Tables I to V represent each metric average obtained from the cross-validation scheme.

III. Results and Discussion

In this section, we provide a detailed account of the results obtained from the polyphony evaluation and the assessment of the G-SELD system's performance in sound environments with increasing complexity. Additionally, we present the findings of an ablation study conducted on the proposed improvements of the G-SELD system.

TABLE I
G-SELD AND SELDNET PERFORMANCES UNDER DIFFERENT POLYPHONY LEVELS IN A FREE FIELD CONDITION*

	SELDnet [14]			G-SELD		
Metric	**OV1**	**OV2**	**OV3**	**OV1**	**OV2**	**OV3**
F1 ↑	97.70	89.00	85.60	98.60	95.00	86.40
ER ↓	0.04	0.16	0.19	0.02	0.09	0.19
LR ↑	99.40	85.60	70.20	99.60	95.70	90.50
LE ↓	3.40	13.80	17.30	2.40	7.30	12.90
SELD ↓	0.022	0.12	0.18	**0.013**	**0.06**	**0.12**

* Note that the arrows show if the metric improves by increasing or decreasing its value.

TABLE II
G-SELD AND SELDNET PERFORMANCES UNDER DIFFERENT POLYPHONY LEVELS IN A REVERBERANT ENVIRONMENT*

	SELDnet			G-SELD		
Metric	**OV1**	**OV2**	**OV3**	**OV1**	**OV2**	**OV3**
F1 ↑	60.30	53.10	51.10	85.00	78.50	57.50
ER ↓	0.40	0.49	0.53	0.19	0.31	0.51
LR ↑	64.90	41.50	24.60	82.30	82.00	55.90
LE ↓	26.60	33.70	36.10	6.00	13.20	21.30
SELD ↓	0.32	0.43	0.49	**0.14**	**0.19**	**0.37**

* The arrows show if the metric improves by increasing or decreasing its value.

A. Polyphony Evaluation

The performance of the G-SELD system was evaluated under different polyphonic levels in audio scenes, starting with a scene containing sources that do not overlap in time (OV1), followed by scenes with higher polyphonic levels (OV2, OV3). This experiment was performed on two datasets that simulate a free field condition and a reverberant environment.

1) Free Field Condition: For this evaluation, we used the ANSYN dataset that represents an ideal free field condition with no reflections. We compare our results with SELDnet [14], which was trained for a maximum of 1000 epochs, against our G-SELD system trained for 100 epochs, saving the last best model. Table I shows the results of this experiment, in which, as expected, the best performance of the G-SELD system was achieved with no overlapping sound events, followed by the performances related to two and three overlapping sounds of the same dataset. The $\text{SELD}_{\text{score}}$ gives us a general idea of all metrics, simplifying the overall performance comparison. Note that the arrows show if the metric improves by increasing or decreasing its value. We show that the G-SELD polyphony evaluation metrics for a free field condition dataset surpass the equivalent SELDnet metrics.

This experiment can be partially compared with [38], in which the human listening ability to identify and localize the total number of simultaneous sound sources spatially distributed was studied. The estimation depends on the audio signal type (speech or tone stimuli) and the overlapped sounds. The percentages achieved by the listeners were in the range of $68-93\%$ for a single sound source, $42-84\%$ for two sounds overlapping in time, and $34-70\%$ for three sounds overlapping in time. The metric that could be considered analogous is the LR, as it evaluates the number of sound events that were correctly located. Comparing the results obtained with human listeners with the performance of the G-SELD system, we noted that our machine hearing system surpasses by 6.6%, 11.7%, and 20.5% the best localization performances of the human auditory system in the aforementioned scenarios.

2) Reverberant Environment: The G-SELD system was also evaluated in the REAL dataset, which emulates a reverberant environment. Table II shows the results for the evaluation split using SELDnet and G-SELD systems. We identified the exact behavior of the G-SELD system on the ANSYN dataset, where the metrics worsen as the number of overlapping sound events increases. We also evinced that the $\text{SELD}_{\text{score}}$ and all metrics of each subset OV1, OV2, and OV3 were improved with the G-SELD system, compared with SELDnet results. Also, G-SELD metrics improvement exceeds the boost obtained on the ANSYN dataset.

B. Sound Environment Evaluation

In this section, we present the results for the G-SELD system evaluated on four spatial audio datasets which represent different sound scene conditions. The selected datasets, previously presented in Section II-A, include ANSYN, REAL, L3DAS21, and DCASE2021 datasets. The system does not need to know beforehand the number of sound events in each sample, and we are not focusing on the polyphony level anymore but rather on the sound environmental conditions. However, polyphony is limited to three sound sources, which is the maximum number of target sound sources present in all the considered datasets. We applied the k-fold cross-validation technique to the training samples of each dataset, whereas the testing splits were always maintained. Note that training, testing, and validation are always made within the same dataset.

The results obtained on each test set are summarized in Table III. We compute the mean of the metrics obtained from the cross-validation models of each dataset (G-SELD mean values). The metrics of the best model over the cross-validation process are also included in Table III to exhibit the best performance of G-SELD for each dataset (G-SELD best model). In the following sections we analyze our results, compare them with the reported SOTA approaches used as a baseline for each dataset, and analyze the learning curves.

1) Free Field Condition: We compare the metrics of the G-SELD system with the reported in [17] for the SELDnet and SELD-TCN approaches and evaluate them on the same dataset. As shown in Table III, the $\text{SELD}_{\text{score}}$, which takes into account the joint performance of all four metrics, was surpassed by the G-SELD system. Considering the G-SELD mean values for each metric, the LR and LE metrics associated with the class-sensitive localization were exceeded by 13.67% and 9.22 points, respectively, compared with SELDnet. They surpassed the performance of the SELD-TCN system by 8.07% and 7.72%, respectively. The LR and LE metrics of our best model surpassed in 13.70% and 9.30 points the SELDnet approach and in 8.10% and 7.80 points the SELD-TCN approach. However, the location-sensitive detection metrics (F_{score} and ER) did not surpass the values achieved by SELD-TCN, in 2% for the F_{score} and 0.03% for the ER. Nevertheless, the percentage of

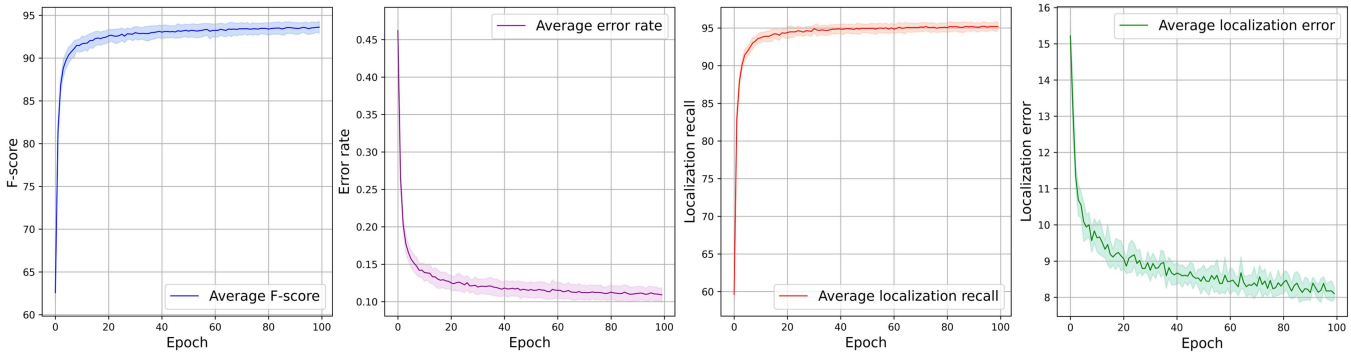

Fig. 6. Learning curves of the G-SELD system evaluated on a free field condition. ANSYN dataset.

TABLE III
PERFORMANCE EVALUATION OF THE G-SELD SYSTEM ON FOUR DATASETS THAT REPRESENT DIFFERENT ACOUSTIC SCENE CONDITIONS. EACH CONDITION IS COMPARED TO ITS BASELINE SOLUTION * †

Condition	Approach	F1↑	ER ↓	LR↑	LE↓	SELD↓
Free field (ANSYN)	G-SELD mean values	93.20	0.12	94.87	8.28	0.07
	G-SELD best model	93.50	0.11	**94.90**	**8.20**	**0.07**
	SELDnet	93.40	0.11	81.20	17.50	0.12
	SELD-TCN	**95.50**	**0.08**	86.80	16.00	0.09
Reveberant (REAL)	G-SELD mean values	78.02	0.30	78.78	11.94	0.21
	G-SELD best model	**79.80**	**0.29**	**79.80**	**11.70**	**0.19**
	SELDnet	74.10	0.39	48.20	38.10	0.34
	SELD-TCN	75.10	0.39	52.40	35.80	0.33
Background noise (L3DAS21)	G-SELD mean values	59.20	0.50	59.90	13.70	0.35
	G-SELD best model	**61.40**	0.47	**62.30**	13.60	0.33
	L3DAS21 Baseline [18]	45.00	-	40.00	-	-
Directional interferences (DCASE21)	G-SELD mean values	43.12	0.65	55.82	23.23	0.46
	G-SELD best model	**43.90**	**0.65**	**55.90**	**22.50**	**0.44**
	DCASE 2021 Baseline [15]	30.70	0.73	40.50	24.50	0.54

* The symbol '-' indicates that the information is not available.
† The arrows show if the metric improves by increasing or decreasing its value.

improvement obtained for the SSL task allows our system to maintain the best performance according to the joint $\text{SELD}_{\text{score}}$ metric.

The learning curves, shown in Fig. 6, present the metrics' evolution on the ANSYN dataset's validation split. The solid lines represent the average value calculated from the metrics of each cross-validation model for each epoch. The colored shadow that wraps around each curve represents two standard deviations below and above the mean, giving us an idea of the values' variability across the cross-validation combinations. These learning curves show that the model fitted the data since the validation metrics reached almost optimal values. The standard deviation reaffirms that the performance of our models is consistent over the cross-validation splits. Additionally, it is possible to evince that in the validation subset, the model reached a better performance on the localization task, as exhibited for the test set results in Table III.

2) Reverberant Environment: According to the evaluation results of the REAL dataset, presented in Table III, we can corroborate that modeling this dataset became more complicated than the ANSYN dataset due to the nature of the acoustic scene in which there are reflections produced by the use of real IRs captured in a reverberant space. However, our results surpassed all the metrics obtained with SELDnet and SELD-TCN systems. The LR was the metric with the most significant improvement, exceeding the SELDnet LR by 30.58% for the G-SELD mean value and by 31.60% for the G-SELD best model. The $\text{SELD}_{\text{score}}$ also corroborates the general best performance of our model compared with SOTA systems.

The learning curves are shown in Fig. 7, in which the validation curves suggest that the characteristics learned from training data were not enough to perfectly generalize our model to unseen data. The learning curves' evolution does not show overfitting, and the reached values are comparable with the metrics of the testing split. Additionally, the colored shadows that represent two standard deviations show that the cross-validation models result on validation metrics close to the mean, showing a reduced variability across models.

A plausible explanation for the drop in performance of the G-SELD system on REAL dataset, compared with ANSYN dataset, is that a substantial multi-path interference caused by room reverberation can significantly impact localization [39]. As concluded by [40], reverberation tends to smear the periodic components across time, and thus some time-frequency (T-F) samples in the reverberation tail are incorrectly assigned to the detected sources. However, our results are still competitive, considering the complexity of the scenario.

3) Reverberant Environment With Background Noise: Next in line, we use the L3DAS21 dataset, which IRs contain reflections from room boundary surfaces and office furniture. Moreover, this dataset includes constant background noise. The L3DAS21 Challenge baseline system [18] computed two metrics: F_{score} and LR, which can be compared with ours (Table III). As the baseline also computed precision of $P = 52.00\%$, we calculate and compare our average precision $P = 62.26\%$, demonstrating that all the comparable metrics were surpassed by more than 10% using the G-SELD system. As expected, the obtained metrics are lower than those achieved on ANSYN and REAL datasets since L3DAS21 is a more challenging scenario that combines reverberation with background noises. Authors in [41]

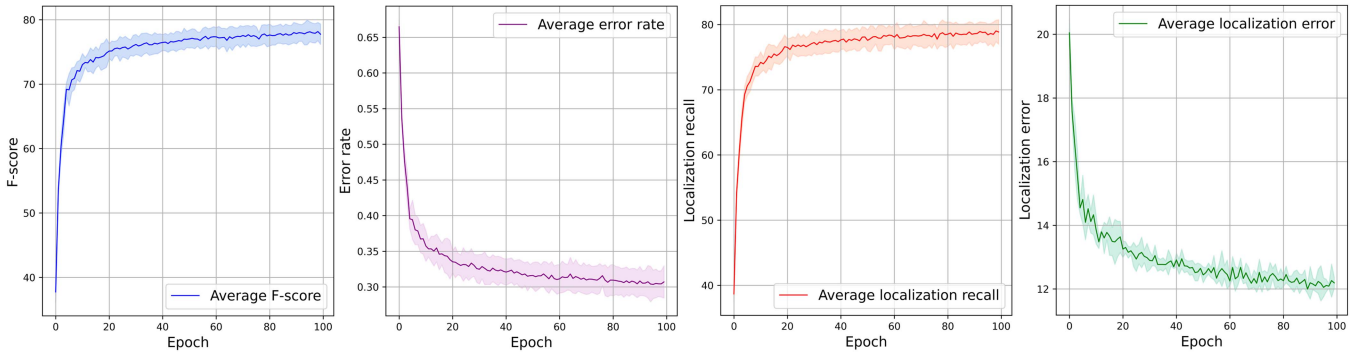

Fig. 7. Learning curves of the G-SELD system evaluated on a reverberant environment - REAL dataset.

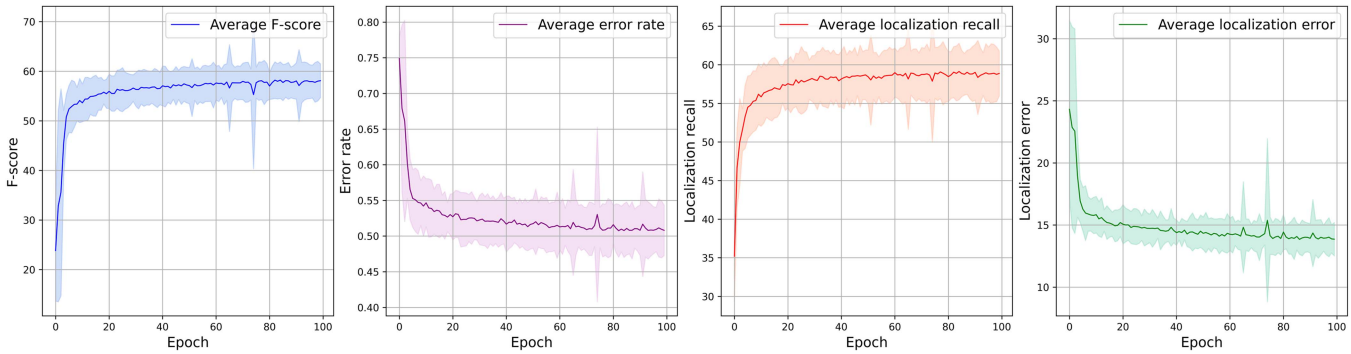

Fig. 8. Learning curves of the G-SELD system evaluated on a reverberant environment with background noise - L3DAS21 dataset.

demonstrated that early reflections produce phase misalignment that greatly decreases the ability to separate signals from noise. These facts help us understand the reasons for the decrease in the performance of our system in this environment. The mixed presence of reflections, background noise, and a 44% increment of sound classes to be identified reduced the performance of our system on the SELD task.

The learning curves of the G-SELD system applied to the L3DAS21 dataset are presented in Fig. 8. All curves reached almost flat slopes before completing 100 epochs, which means that the models could extract and learn features from original and slightly modified data during the number of training epochs.

We also note the presence of some peaks on the learning curves, which could be caused by the mini-batch gradient descent method used in Adam optimization. In other words, as training data is shuffled, the mini-batches may contain a more significant amount of unusual samples, causing a slight decrease in the metrics in a specific epoch. The colored shadows representing two standard deviations below and above the mean show that the cross-validation models produce slightly more variable validation metrics than the obtained for ANSYN and REAL datasets. This can be explained by the increased number of sound classes contained in a smaller set of audio samples of the L3DAS21 dataset.

4) Reverberant Environment With Moving Sound Sources and Directional Interferences: DCASE2021 is the most challenging dataset in which our G-SELD system was evaluated, as it includes all the complexities presented in ANSYN, REAL, and L3DAS21 datasets. Moreover, different challenging conditions were included to simulate difficult real-life situations. Moving sources were incorporated from about 500 sound event samples of 12 types, and an additional layer of directional interferences was selected from 400 sound events. The network is expected to learn to ignore interferences; if not, they will be considered false positives.

We compare our results with the metrics published by the baseline system of DCASE2021 Challenge [15]. As shown in Table III, all metrics were surpassed. The F_{score} and LR were exceeded in 12.42% and 15.32% respectively by the G-SELD mean values, and in 13.20% and 15.40% by the G-SELD best model. The ER was improved by 7% by the mean value and the best model, and the LE was surpassed by 1.27 and 2 points by the mean value and the best model, respectively. Our system reached an overall improvement of 10% according to the $\text{SELD}_{\text{score}}$. The LE exhibits that the inclusion of moving sources turns the DOA estimation more complex, such that the localization of several detected samples of sound events does not satisfy the threshold to be considered as a correct localization prediction. However, the improvements are promising, considering that the G-SELD network architecture is dealing with the SELD problem without dividing it into specific branches for the localization and detection subtasks, maintaining conceptual simplicity on the implemented modifications.

The cross-validation models' variability, represented by the standard deviation shown in the learning curves of Fig. 9 increased compared with previous datasets. However, we consider it tolerable since this dataset includes a wider variety of sounds and DOAs in a few audio samples.

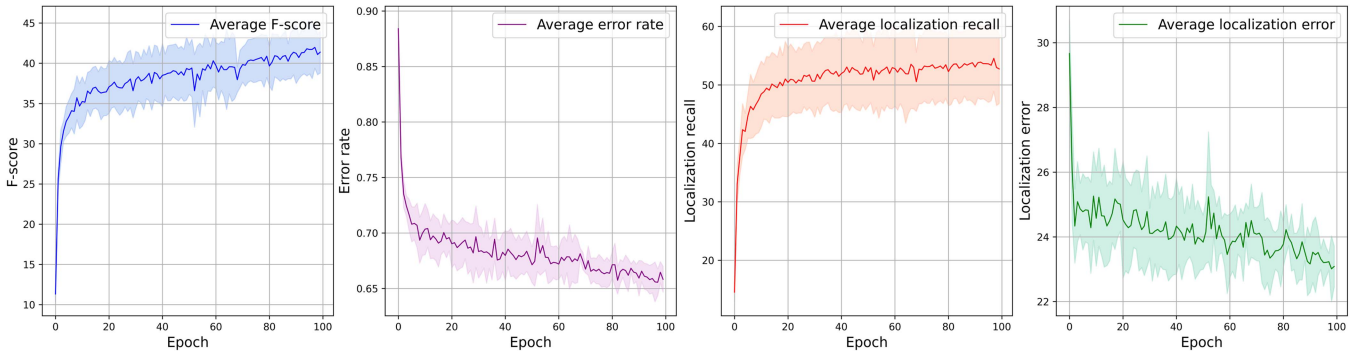

Fig. 9. Learning curves of the G-SELD system evaluated on a reverberant environment with moving sound sources and directional interferences - DCASE2021 dataset.

We identified that the LR was the less aggravated metric compared to the increasing difficulty of the databases. This demonstrates that the G-SELD model can extract a significant amount of information from the feature vectors, which leads to the detection of a significant quantity of samples that contain sound events, even in challenging scenarios.

The difficulty related to directional interference increases when the target sound is similar to the sound that should be considered as an interference (inter-class similarity problem). We identified the mentioned problem in the DCASE2021 database, in which *engines* and *fire* sounds are used as interferences. Then, considering that a characteristic sound related to fire is a fire alarm, the system learned from many samples that an alarm-like sound should be considered interference. Therefore, in case the system detects a sound with comparable characteristics, it will wrongly disregard this sound as an interference.

C. Ablation Study

We conducted an ablation study to analyze the individual contributions of the Gammatone based SELD (G-SELD) system. However, we highlight that the G-SELD system as a whole encompasses all the proposed improvements, whose results were presented in Sections III-A and III-B. The k-fold cross-validation technique was also used to train different split combinations of each dataset. The results in this section represent the mean value of the metrics obtained from the test split of the cross-validation scheme.

1) Gammatone Vs. Mel Filter Banks: The SELD-DCASE2021 architecture proposed in [15] was used to evaluate our hypothesis of using a gammatone filter bank instead of a Mel filter bank for obtaining a better performance on the SELD task. For this experiment, we changed the filter bank while maintaining fixed all other parameters related to the preprocessing stage. The results presented in Table IV were obtained for the test fold of the ANSYN dataset. This dataset was selected due to the high metrics achieved by baseline systems such as SELDnet and SELD-TCN. We noted that as metrics reach near-perfect values, it becomes more difficult to get improvements. Then, we sought to prove that just changing the filter bank applied to the spectrogram in the preprocessing stage results in a performance improvement in a dataset that has already reached nearly perfect values. Note that although small, all metrics show an improvement. We did not use the G-SELD model architecture, since we wanted to guarantee that the filter bank change alone leads to better performance.

TABLE IV
COMPARISON BETWEEN THE USE OF GAMMATONE AND MEL FILTER BANKS USING THE SELD-DCASE2021 ARCHITECTURE †

	F1↑	ER↓	LR↑	LE↓	SELD ↓
gammatone filter bank	86.40	0.21	89.20	10.70	**0.12**
Mel filter bank	85.80	0.23	88.60	11.00	0.13

† The arrows show if the metric improves by increasing or decreasing its value.

2) Inclusion of a TCN Block: As previously explained, the G-SELD architecture contains four types of blocks: CNN, TCN, RNN, and FC. In order to visualize each group of blocks' contribution to the SED task, we apply the t-SNE visualization technique that reduces a high-dimensional feature vector into a two or three-dimensional map [42]. In this experiment, we restricted our data to samples that contain just one sound event at the same time to simplify the clusters' visualization. The ANSYN or REAL datasets could be used for this experiment since they provide a split of data containing audio with sound events happening one at a time. As in the previous experiment, we selected ANSYN dataset because it is more challenging to get improvements in a database that has reached near-perfect metrics.

Fig. 10 shows the t-SNE representations of the output vectors taken after each group of blocks in the G-SELD architecture, with a perplexity value of 50. It is possible to recognize a clustering process that begins with the CNN blocks and finishes with the FC layers. However, despite being close to each other, the class-coincident samples of the CNN output are better clustered after passing through the TCN block. The clustering evidences how valuable the use of the TCN block is in the G-SELD architecture. Then, the RNN and FC layers, as final stages of the network, are used for learning temporal dependencies of data and reducing its dimensionality, respectively, which also contributes to the clustering evolution through the model.

3) Data Augmentation: We also experimented with the stage of data augmentation, aiming to demonstrate that our G-SELD

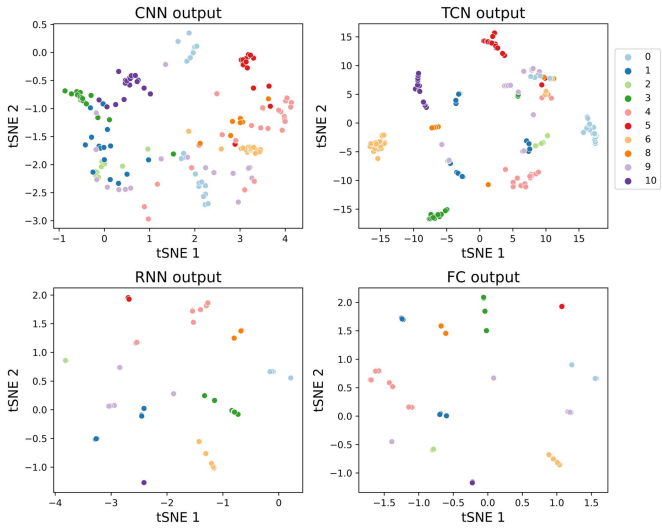

Fig. 10. t-SNE representation of the CNN, TCN, RNN, and FC outputs for a sample of the ANSYN dataset. The colors represent the sound classes.

TABLE V
COMPARISON OF THE METRICS IMPROVEMENT WITH AND WITHOUT DATA AUGMENTATION †

	F1↑	ER↓	LR↑	LE↓	SELD↓
Data augmentation	58.75	0.50	59.60	13.83	**0.35**
No data augmentation	52.70	0.55	56.50	17.00	0.39

† The arrows show if the metric improves by increasing or decreasing its value.

system can improve the metrics even without this technique. The best-ranked results in Task 3 of the DCASE2021 challenge [15] showed that using data augmentation techniques applied to spectrograms results in a performance improvement on the DCASE2021 dataset [43], [44]. Therefore, we decided to conduct our experiment in a dataset that has not yet been used for this comparison. Concretely, we use the L3DAS21 dataset to explore the impact of data augmentation in the overall improvement of our model.

For this trial, we used the L3DAS21 dataset and the G-SELD architecture. The neural network was first trained without data augmentation and then using the three data augmentation techniques detailed in Section II-C. The results for the test fold are shown in Table V. It is possible to compare two metrics of our results with the published for the L3DAS21 Challenge baseline system: the $F_{\text{score}} = 45.0$ and the $LR = 40.0$ were improved by the G-SELD system without data augmentation at 7.7% and 16.5% respectively and by 13.75% and 19.60% with the use of data augmentation. In conclusion, the G-SELD system improves the metrics of the SELD task even without the data augmentation stage.

Based on the experiments presented in the last sections, we show that each proposed improvement in the G-SELD system is a valuable addition to the whole performance of the system.

IV. CONCLUSION

In this work, we used a deep learning approach to develop a system for sound event detection and localization in spatial audio.

A combination of acoustic features inspired by the human auditory system and IVs containing phase information were implemented to provide appropriate cues for estimating the location in time and the direction of arrival of a sound event. It was demonstrated that gammatone filters are a viable alternative to modify the frequency linear resolution of the spectrogram since they model the tonotopic frequency distribution produced in the cochlea.

Based on a deep learning model that includes CNN and RNN layers, the architecture of our model is improved by incorporating a TCN block that is capable of learning core features in the structure of sequential data, due to its ability to capture long-term dependencies. This modification generates a deeper feature extraction, producing a more significant number of trainable parameters.

In summary, the G-SELD system was evaluated on four databases that provide different ambient conditions, from a controlled environment without reflections to various reverberant scenes. The G-SELD system maintains a good performance for polyphony up to level three in anechoic and reverberant environments. The performance decays when background noises and directional interferences are included in addition to the target classes because the system must learn to overlook those specific types of sounds. However, our results surpassed the ones obtained using the baseline systems proposed along with each dataset, maintaining a conceptual simplicity of the network architecture.

REFERENCES

[1] C. Colby, "Perception of extrapersonal space: Psychological and neural aspects," in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. New York, NY, USA: Pergamon, 2001, pp. 11205–11209.
[2] A. Ronzhin, A. Ronzhin, and V. Budkov, "Audiovisual speaker localization in medium smart meeting room," in *Proc. IEEE 8th Int. Conf. Inf., Commun. Signal Process.*, 2011, pp. 1–5.
[3] M. Yağanoğlu and C. Köse, "Real-time detection of important sounds with a wearable vibration based device for hearing-impaired people," *Electronics*, vol. 7, no. 4, Apr. 2018, Art. no. 50.
[4] I.-C. Yoo and D. Yook, "Automatic sound recognition for the hearing impaired," *IEEE Trans. Consum. Electron.*, vol. 54, no. 4, pp. 2029–2036, Nov. 2008.
[5] M. Kushwaha, X. Koutsoukos, P. Volgyesi, and A. Ledeczi, "Acoustic source localization and discrimination in urban environments," in *Proc. IEEE 12th Int. Conf. Inf. Fusion*, 2009, pp. 1859–1866.
[6] E. Browning, R. Gibb, P. Glover-Kapfer, and K. Jones, "Passive acoustic monitoring in ecology and conservation," WWF Conservation Technology Series 1(2), WWF, Woking, U.K., Tech. Rep., Oct. 2017. [Online]. Available: http://dx.doi.org/10.25607/OBP-876
[7] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.
[8] A. Marti, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 134, no. 4, pp. 2627–2630, 2013.
[9] P. Townsend, "Enhancements to the generalized sidelobe canceller for audio beamforming in an immersive environment," Ph.D. dissertation, Univ. of Kentucky, Lexington, KY, USA, 2009.
[10] A. Xenaki and P. Gerstoft, "Compressive beamforming," *J. Acoustical Soc. Amer.*, vol. 136, no. 1, pp. 260–271, 2014.
[11] Q.-H. Huang, Q. Zhong, and Q.-l. Zhuang, "Source localization with minimum variance distortionless response for spherical microphone arrays," *J. Shanghai Univ. (English Edition)*, vol. 15, no. 1, pp. 21–25, 2011.

[12] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. IEEE 26th Eur. Signal Process. Conf.*, 2018, pp. 1462–1466.
[13] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Proc. Audio Eng. Soc. 138th Conv.*, 2015, p. 6.
[14] 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.
[15] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Proc. Detection Classication Acoustic Scenes Events*, 2021, pp. 125–129.
[16] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 915–919.
[17] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif, and B. Yang, "SELD-TCN: Sound event localization & detection via temporal convolutional networks," in *Proc. IEEE 28th Eur. Signal Process. Conf.*, 2021, pp. 16–20.
[18] 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.
[19] Z.-Q. Wang and D. Wang, "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1–5.
[20] S. Mondal and A. D. Barman, "Human auditory model based real-time smart home acoustic event monitoring," *Multimedia Tools Appl.*, vol. 81, no. 1, pp. 887–906, 2022.
[21] Y. R. Leng, H. D. Tran, N. Kitaoka, and H. Li, "Selective Gammatone filterbank feature for robust sound event recognition," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 2246–2249.
[22] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, "Listen carefully and tell: An audio captioning system based on residual learning and Gammatone audio representation," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2020, pp. 150–154.
[23] Y. Jin, H. Su, C. Xu, and Q. Guo, "Application of Gammatone filter bank to active noise control algorithm," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput.*, 2017, pp. 1–5.
[24] A. Roginska and P. Geluso, Eds., *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*, 1st ed. Evanston, IL, USA: Routledge, 2018, pp. 53–54.
[25] A. Mesaros et al., "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 379–393, Feb. 2018, doi: 10.1109/TASLP.2017.2778423.
[26] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 1041–1044.
[27] 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.
[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] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The nigens general sound events database," 2019, *arXiv:1902.08314*.
[30] D. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Int. Speech Commun. Assoc.*, 2019, pp. 2613–2617.
[31] L. Mazzon, M. Yasuda, Y. Koizumi, and N. Harada, "Sound event localization and detection using FOA domain spatial augmentation," DCASE2019 Challenge, Tech. Rep., Jun. 2019.
[32] D. Rho, S. Lee, J. Park, T. Kim, J. Chang, and J. Ko, "A combination of various neural networks for sound event localization and detection," DCASE2021 Challenge, Tech. Rep., Nov. 2021.
[33] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[35] K. Tyagi, S. Nguyen, R. Rawat, and M. Manry, "Second order training and sizing for the multilayer perceptron," *Neural Process. Lett.*, vol. 51, pp. 963–991, 2020.
[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[37] 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.
[38] X. Zhong and W. A. Yost, "How many images are in an auditory scene?," *J. Acoust. Soc. America*, vol. 141, no. 4, Apr. 2017, Art. no. 2882.
[39] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays* (ser. Digital Signal Processing), M. Brandstein, D. Ward, A. Lacroix, and A. Venetsanopoulos, Eds. Berlin, Germany: Springer, 2001, pp. 157–180.
[40] J. Woodruff and D. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 806–815, Apr. 2013.
[41] D. Griesinger, "What is "proximity," how do early reflections and reverberation affect it, and can it be studied with LOC and existing binaural data?," in *Proc. 22nd Int. Congr. Acoust.*, 2016, pp. 1–5.
[42] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
[43] K. Shimada et al., "Ensemble of ACCDOA-and EINV2-based systems with D3Nets and impulse response simulation for sound event localization and detection," 2021, *arXiv:2106.10806*.
[44] T. N. T. Nguyen, K. Watcharasupat, N. K. Nguyen, D. L. Jones, and W. S. Gan, "DCASE 2021 task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection," 2021, *arXiv:2106.15190*.

Karen Rosero (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering and telecommunications from Army Polytechnic School, Sangolquí, Ecuador, in 2020 and the M.Sc. degree in electrical engineering from the School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Campinas, Brazil, in 2022. She is currently a Ph.D. Student with The University of Texas at Dallas, Richardson, USA. Her research interests include spatial audio, deep learning, music information retrieval, affective computing, and multimodal signal processing.

Felipe Grijalva (Senior Member, IEEE) received the B.S. degree in electrical engineering and telecommunications from the Army Polytechnic School, Sangolquí, Ecuador, in 2010 and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Campinas, Campinas, Brazil, in 2014 and 2018, respectively. He is currently a Professor with the Universidad San Francisco de Quito, Quito, Ecuador. He was with signal processing, machine learning, and computer vision applications.

Bruno Masiero (Member, IEEE) received the B.S. and M.Sc. degrees in electrical engineering from the University of São Paulo, São Paulo, Brazil, in 2005 and 2007. He is an Assistant Professor with the School of Electrical and Computer Engineering (FEEC), University of Campinas, Campinas, Brazil. His research interests include application of modern digital signal processing techniques in audio and acoustic applications, such as spatial sound acquisition and reproduction, acoustic imaging, characterization of acoustic materials, and development of audiological assessment tools. In 2012, he was the recipient of the Ph.D. in engineering by the RWTH Aachen University, Aachen, Germany. During 2019–2022, he was a Member of the Board of the International Commission for Acoustics.