Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 4, 2025

Real-Time Sound Event Localization and Detection using IoT Raspberry PI Devices based on Single-Stage CRNN

[1]Syed Owais Shah, [2]Muneeba Darwaish, [*3]Mohsin Shah, [4]Muhammad Asad Khan, [5]Muhammad Shujaat

[1]*Department of Telecommunication, Hazara University Mansehra.*
[2]*Department of Telecommunication, Hazara University Mansehra*
[*3]*Department of Telecommunication, Hazara University, Mansehra.*
[4]*Department of Telecommunication, Hazara University Mansehra.*
[5]*Department of Telecommunication, Hazara University Mansehra.*

[1]owais@hu.edu.pk, [2]muneebadarwaish@gmail.com, [*3]syedmohsinshah@hu.edu.pk, [4]asadkhan@hu.edu.pk, [5]mshujaat749@gmail.com

DOI: https://doi.org/

Keywords
Sound event detection, Acoustic Scene, Single-stage CRNN, Spec-mix augmentation

Article History
Received on 02 April 2025
Accepted on 26 April 2025
Published on 28 April 2025

Corresponding Author:
Mohsin Shah

Abstract

Sound event localization and detection (SELD) plays a vital role in understanding the environment. Recently, the SELD problem has received increasing interest from the research community. The state-of-the-art models for the DCASE 2020 for the SELD task have achieved good performance in terms of accuracy, but these models are generally based on multistage deep neural networks, which require substantial computational power and memory, making them unsuitable for deployment on low-cost Internet of Things (IoT) edge devices. In this paper, we propose a single-stage convolutional recurrent neural network (SS-CRNN) designed for real-time implementation of SELD on resource-constrained devices like Raspberry Pi. We also conducted a comprehensive analysis of different feature representations in terms of both accuracy and computational efficiency. Our results demonstrate that the SS-CRNN outperforms other models on the DCASE 2020 SELD dataset in terms of real-time factor (RTF), with only a slight trade-off: a 1% reduction in frame recall performance and a 6-degree decrease in localization accuracy compared to state-of-the-art methods. Additionally, we employ SpecMix augmentation to further enhance the model's performance, which helps to boost our model's performance during training.

INTRODUCTION

Localization and event detection (SELD) has received increasing interest over the last few years due to its application in domestic activity monitoring [1], robot guidance [2], and speaker identification [3]. The SELD can be divided into two sub-tasks, i.e. Sound Event Detection (SED) and Direction of Arrival (DOA). In SED, the goal is to find the onset and offset time of sound events while also classifying these events. Whereas in DOA, the aim is to find the azimuth (Azi) and elevation (Elev) of the active sound event source [4]. So, the overall goal of the SELD is to find the active label and its location when an event is active. SED can be further divided into monophonic or polyphonic. In the monophonic SED, there is no overlapping of the sound events, whereas in the polyphonic SED, there is an overlapping of the sound events. Although monophonic sound events are relatively easier to work on, the polyphonic SED method is more useful for real-world applications, as it is more likely for these to contain several sound sources [5]. The main challenge for the SELD is intra-class variability, in which for instance footsteps or car horn events if repeated, may cause small shifts in the frequency and time domain.

SELD plays a vital role in environmental awareness and is useful in many applications, such as robotics that can detects and localizes the sound source and can also navigate the directions [2], bio-acoustic sensors in which autonomous recording units (ARUs) cap-ture sounds of wildlife animals for longer periods [6], sound localization that deter-mines the location of a sound source in three dimensions (two angles and distance) [7] and speech segregation to separate speech from noise [8]. For many of the above-mentioned real-world applications, SELD has to be implemented on low-end processing devices rather than standard computers [9].

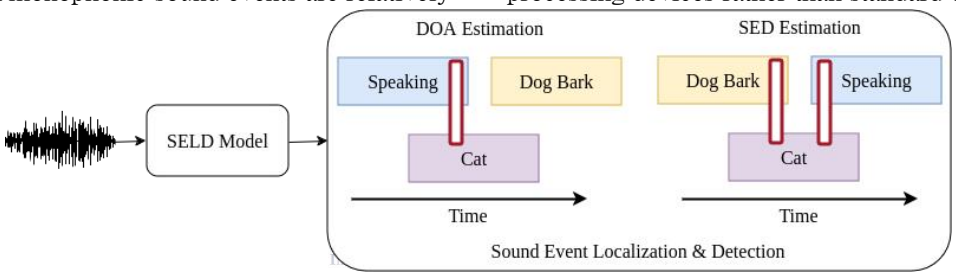

Figure 1: A General Illustration of SELD Model

The primary objectives of the Sound Event Localization and Detection (SELD) task are to (i) identify the classes of relevant sound events, (ii) determine the onset and off- set times for each occurrence, and (iii) localize the spatial position of each sound event while it is active. When an audio signal exhibits activity, a SELD system generates a temporal activation track for each target sound class, along with one or more corresponding spatial trajectories. The result is a spatio-temporal representation of the acoustic scene, which can be leveraged in a wide range of artificial intelligence applications, such as self-localization, environ- ment classification, navigation in the absence of visual input or with occluded targets, monitoring specific sound sources, smart home systems, scene visualization tools, and audio surveillance, among others. The detection and localization of overlapping sound events is referred to as polyphonic SELD, as illustrated in Figure 1.

Recently, the use of single-board computers (SBC) [10] have gained the notice of the reasearchers as they are cost- effective. When compared with the standard computers, the SBCs are cost-friendly, compact, and also engergy efficient which make them well suited to operate with batteries for longer durations in remote locations. The processing unit is the primary part of the SBC. Some of the SBCs that are readily available in market are, Raspberry Pi, BeagleBone, and Arduino [10]. The compute-intensive techniques that are challenging to implement on SBCs, such as deep learning for image classification, while SELD techniques can be imple-mented on these low-power SBCs.

The Internet of Things (IoT) has palyed an important role in our day to day lives, and it continous to shape our world in more impactful ways and make the surrounding environment smarter. Among the IoT techniques used in a smart environment, SELD is one of the technologies

that offers by interpreting the acoustic landscape of our daily lives. Our surroundings are filled with physical activities or events in the form of sounds, such as ambulances passing by, heavy traffic on roads, the opening and closing of doors, turning off the page, heartbeats, etc. These sound events carrying information about the soundings, and the human auditory system is capable to separate these sound events and recognize them. If we want to make such a smart system to act like a human's auditory system, then it must be able to separate sound events well enough to recognize and localize them as well.

In this paper, we propose an efficient algorithmic implementation for the Raspberry Pi to address the SELD challenge in real time. The Raspberry Pi is a portable, low-cost, and energy-efficient device that can be easily de- ployed for a wide range of applications. Although the top algorithms proposed in DCASE 2019, 2020, and 2021 for SELD task improved the performance in terms of sound event detection and classification as well as the direction of arrival estimation, most of these were generally based on complex and multi-stage deep neural networks. As a result, such algorithms are typically too computationally intensive to run in real-time on resource-constrained, low-cost IoT edge devices like the Raspberry Pi. We propose a single- stage CRNN, which is more hardware-friendly as compared to other proposed methods. Real-time factor (RTF) is an important metric used for the feasibility analysis of real-time implementations. Lesser the RTF, the better the execution speed. For real-time implementation, the RTF value must be less then one. Our proposed system has achieved a low enough RTF value between 0.5 and 0.7, with only a slight trade-off: a 1% reduction in frame recall performance and a 6-degree decrease in localization accuracy compared to state- of-the-art methods

The major contribution of this paper are as follows:

We propose a novel single-stage CRNN architecture which is efficiently implemented on Raspberry Pi for SELD in real-time (SELD-RT). The pro-posed system has low power consumption and low computational cost which makes it very suitable for real-world SELD applications. The best performers in DCASE 2019 2020 and 2021 use multi-stage CRNN, with higher computational complexity. Our single-stage CRNN architecture re-duces the computational complexity.

We highlight the discrepancies between four different features and figure out the best performance while con-sidering the RTF. We achieve SELD results by using only magnitude information of logmel energies.

The rest of the paper is organized as follows: Section II describes a comprehensive survey of the field, the proposed method is described in Section IV, experimental results, performance evaluation, embedding capacity, visual quality and comparison are presented in Section VI, and finally conclusion is drawn in Section VIII.

RELATED WORKS

Early approaches to SED methods utilize Mel Frequency Cepstral Coefficients (MFCC) [11], which often combine with Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) [12], also Non-negative Matrix Factorization (NMF) [13] has also been explored for the SED.

In recent years, deep learning has emerged as a promising technique for problems in almost every field. Convolutional Neural Networks (CNN) and Convolutional Recurrent Neural Networks (CRNN) have shown promising performance in the SED task [14]. The use of CNN and recurrent neural networks (RNN) has interestingly increased the field of audio signal processing, with their combination outperforming earlier SED models [15]. Besides, a combination of CNN and RNN outperforms previous SED models [16]. However, RNN suffers from vanishing gradients due to its long-term dependencies during training [17]. To address the vanishing gradient problem, gated recurrent units (GRU) and long short-term memory (LSTM) [18] are proposed. Both GRU and LSTM have a gating mechanism that selectively learns relevant information while discarding irrelevant information, making them effective in capturing temporal patterns that are relevant in the SED tasks [19]. Despite their advantages due to gating, deploying these models on low-cost, low-power hardware remains a significant challenge [20].

Several studies have addressed the task of sound event localization (SEL) or direction of arrival (DOA). Some

of the common approaches to estimate the DOA are steered-response power-phase transform (SRP-PHAT) [21], time- time-difference-of-arrival (TDOA) [22], multiple signal classification (MU-SIC) [23], and estimation of signal parameters via the rotational invariant technique (ES-PRIT) [24]. All of these approaches show trade-offs in terms of complexity and compatibility with different microphone array structures. For DOA analysis, among these approaches, MUSIC performs well; however, its effectiveness almost depends on accurate estimation of the number of active sources, which is not always feasible in the real-world scenarios. Moreover, it performs drops under low signal-to-noise ratio (SNR) and reverberation scenarios [25]. Many methods discussed address SED and DOA estimation separately. Several recent studies have proposed performing SED and DOA jointly. Kapka and Lewandowski proposed [26] a four-stage CRNN that contains two main tasks. These subtasks include detecting of active source and estimating of DOA when one sound source is active. The model gives state-of-the-art performance in the DCASE 2019 challenge in the SELD task. They [27] proposed a two-stage approach that performs SED and localization separately. The Logmel energies were used as features for SED, while the intensity vector and GCC-PHAT were used for the localization part. A masking method was used to enhance the DOA, which is based on SED. The Xu et al.

[28] used a conventional microphone array signal processing approach that performed well for the SELD task. They used a three-stage strategy to solve the SELD problem. Their method employed logmel, constant-Q transform (CQT), and the angular spectrum of four channels to solve the SELD problem. Zhang et al. The [29] used Spec-Augmentation [30], a data augmentation technique and post-processing step that is based on prior knowledge regularization (PKR) for the SELD task. They used logmel and short-time Fourier transform (STFT) representations as input features to the dual-branch CRNN. Nguyen et al [31] proposed the decoupling of detection and localization. They used GCC- PHAT and log-mel Bark Spectrogram as input features to the CRNN with mix-up augmentation. Wang et al [32] offer a comprehensive technological solution that includes post-processing, model ensembles, network training, and data augmentation. A more reliable prediction of SELD is produced using a model ensemble with many proposed designs. In addition, they employ a post-processing step, which applies various thresholds at different sound occurrences. In [33], the authors perform the SED part using the Cao et al. [27] model and the DOAs using the single-source histogram approach. The method in [34] built model ensembles by averaging the results of various models that have been trained under various settings, which include various input features, training folds, and model designs. With precise time-frequency mapping between the signal strength and the source of direction-of-arrival, the method in [35] developed a unique feature known as the spatial cue-augmented log-spectrogram (SALSA). The feature consists of stacked multi-channel log-spectrograms for each time-frequency bin on the spectrograms, the estimated direct-to-reverberant ratio, and a normalized version of the primary eigenvector of the spatial covariance matrix. Kang et al. [36] used a conformer-based model with audio channel swapping, which is used for data augmentation, which helps to enhance their model performance. Kim and Ko [37] use a residual connection in the CNN and self-attention mechanism, which is then ensembled to improve the SELD scores. Dong et al. in [38] proposed a model that performs SELD by unifying a framework that models SED, DOA, and distance estimation, which treats DOA and SED parts separately and combines them at the output stage. Yeow et al. [39] enhance the channel and attention mechanism by introducing a squeeze and excite conformer-based RasNet, which utilizes SALSA and data augmentation and data synthetic generation to boost the SELD performance.

OVERVIEW OF DCASE 2019 DATASET

The TAU Spatial Sound Events 2019 dataset [40] provides two datasets for the SELD task. The first set is recorded in four-channel first-order ambiosonic (FOA), which captures a full 3D area (W,X,Y,Z), and the second set consist of four-channel directional microphone (MIC) arrays. The FOA dataset is obtained by converting 32 channels of the Eigenmike using filters

based on anechoic array response measurements. In contrast, the MIC recordings do not need any additional filtering and are obtained using 6, 10, 26, and 22 channels of Eigenmike arrays. FOA and MIC both contain 400 one-minute-long audio recordings, which are sampled at 48 kHz with a 30 dB signal-to-noise (SNR) ratio for audio. These audio recordings were captured in five indoor locations having impulse responses (IRs), resulting in 504 unique combinations of azimuth and elevation. These IRs are then convolved with 11 events of the DCASE 2016 task 2 dataset [41]. The 400 recordings in the development dataset were organized into four cross-validation splits, each containing 100 audio files. In practice, MIC is easier to use due to its direct catering approach, whereas FOA requires a spherical microphone array and additional spatial encoding but provides a more complete representation of the 3D sound field through Ambisonics.

Table 1: *Overview Of Related Work In Sound Event Localization And Detection (SELD)*

Literature (Year)	Architecture / Strategy	Input Features
Kapka et al. [26] (2019)	Four-Channel CRNN with multi-tasking approach	STFT, Complex Spectrogram
Cao et al. [27] (2019)	Two-stage CRNN (SED followed by DOA estimation)	Log-Mel, Intensity Vectors, GCC-PHAT
Xue et al. [28] (2019)	Three-stage CRNN framework	Log-Mel, CQT, Angular Spectra
Zhang et al. [29] (2019)	CRNN with SpecAugment and PKR	Log-Mel, Complex STFT
Nguyen et al. [31] (2019)	Two-stage CRNN with mix-up augmentation	Bark Log-Mel, GCC-PHAT
Wang et al. [32] (2020)	CNN + CRNN hybrid architecture	Log-Mel, Intensity Vectors, GCC-PHAT
Nguyen et al. [33] (2020)	Ensemble of Two-stage CRNNs	Log-Mel, Complex Spectrogram
Shimada et al. [34] (2021)	RD3Net with TFRN and D3Block ensemble	PCEN Spectra, IDP, Magnitude, cosIPD, sinIPD
Nguyen et al. [35] (2021)	CRNN ensemble with eigenvector augmentation	Log-Mel, Eigenvector-Augmented Spectra
Kang et al. [37] (2023)	Confromer with data augmentation	Log-Mel
Kim et al. [37] (2023)	CNN with self attention	Log-Mel
Dong et al. [37] (2025)	Resnet-18 with conformer	Log-Mel, intensity vector
Yeow et al. [37] (2024)	ResNetConformer architectures with Squeeze-and-	Yeow et al. [37] (2024)

We utilzed the development dataset and four cross-validation splits provided by DCASE 2019 task 3, as outlined in Table 2. Each splits contain the audio recordings and their corresponding metadata, which specifies the onset and offset time of the sound events and their spatial locations in terms of azimuth and elevation angles.

PROPOSED METHODOLOGY

THE RASPBERRY PI 4 AND RESPEAKER

The primary objective is to develop an acoustic sensor network (ASN) that has the capability to not only detect active sound events but also localize the events by using spatial information. Although the best algorithms proposed in DCASE 2019 to 2025 have significantly improved the performance in terms of sound event detection as well as the direction of arrival estimation, most of these approaches rely on complex multistage deep neural networks, as shown in Table 1. Furthermore, in dataset creation, eigenmikes were used, which cost around ($3,999).

Due to their high computational requirements and dependence on costly hardware, these algorithms are not well-suited for deployment in low-cost and resource-constrained devices. In contrast, the goal of this work is to explore and propose affordable and efficient solutions for the deployment of ASN using low-power embedded systems. Specifically, we propose a setup based on the Raspberry Pi 4 paired with the ReSpeaker microphone

array, which offers a cost-effective alternative for real-world ASN deployments. The deployment of the ASN model can be divided into three groups depending on sensor functionality and the cost associated with it.

High-End Commercial Devices: These devices are made for accuracy and reliability and are often used in commercial or industrial monitoring applications. These devices may cost up to 11, 000. An example of these devices is Bruel and Kjaer [42], which costs around 15, 000.

Mid-Range Research and Commercial Devices: This category falls in the mid-range and is used for research as well as commercial applications. These devices have a moderate price tag and cost around $600. Examples of such devices are Libelium's Waspmote Plug & Sense and RU-MEUR [9].

Low-Cost Academic Prototypes: These sensors are designed for large-scale deployment that offering a low cost of around ($150). These devices are mainly built by university research groups [43].

Recently, advances in low-cost and low-power microphone technology and networking have provided a platform for real-time data processing with deep learning. For the quality of ASN at low cost, it must provide at least the following features [44].

Recent developments in wireless networking and low-cost, low-power microphone arrays have made it possible to create deep learning-powered real-time ASN monitoring systems. A low-cost ASN must fulfil the following essential conditions to be efficient and scalable.

The ASN will have the ability to monitor SELD with a comparable level of performance in the form of SED and DOA metrics. Doing digital signal processing (DSP) tasks in real-time and also performing wireless audio transmission. Overall cost must not exceed $100.

In this work, we propose a robust and efficient model with low computational and power cost, a real-time SELD system (SELD-RT). The SELD-RT model continuously monitors acoustic environments, detects the active sound events and their azimuth and elevation angles for localization, making it suitable for low-cost real-world scenarios.

Table 2; *Cross-Validation Setup For Model Performance Evaluation*

Fold	Training Set	Validation Set	Test Set
1	Folds 3, 4	Fold 2	Fold 1
2	Folds 4, 1	Fold 3	Fold 2
3	Folds 1, 2	Fold 4	Fold 3
4	Folds 2, 3	Fold 4	Fold 1

In contrast to other SBC devices, such as BeagleBone Black, ODROID C1+, and Tinker Board, which range between 47−70. The Raspberry Pi comes in a complete package for building an ASN because of its processing power, RAM, storage options, built-in WIFI, and support for Secure Shell (SSH) for remote management.

The SBC device used in our proposed system is the Raspberry Pi 4, which has a low cost and low power consumption and features a 1.5 GHz quad-core CPU, 4GB of RAM, and a 40-pin GPIO interface, all at a cost of approximately $59. It runs on the open-source operating system (OS) Raspbian [45] and its SSH allows for remote updates and maintenance of the model. Due to its capabilities, the Raspberry Pi 4 is chosen for real-time audio signal processing and the deployment of deep learning models in embedded environments.

Table 3 presents a comparison of popular single-board computers (SBCs) commonly used in edge computing and sensor network applications. Among these, the Raspberry Pi 4 offers the best balance of performance, connectivity, and affordability. It features a quad-core CPU, 4 GB of RAM, and built-in Wi-Fi, making it well-suited for real-time signal processing and machine learning tasks. Its low cost ($59), small form factor, and active open-source community further support its adoption in scalable, low-power acoustic sensor networks

Considering the main purpose of our work, which is to employ a low-cost and efficient model for ASN, we also proposed a microphone array device, ReSpeaker 4-Mic array[46], that can be integrated with Raspberry Pi 4. ReSpeaker has a tetrahedral array and is widely used for

audio capturing in sound applications. It is easily available and has a low cost ($20-25). Figure 2 illustrates the complete flowchart for the SELD- RT. Audio event signals are captured by the Re-speaker array, which contains 4 microphones that cover 360 degrees in azimuth. These audio signals can be digitized and then fed to our pre-trained model in a Raspberry Pi 4 that performs SED and DOA.

The proposed technique for SELD-RT is evaluated using the TAU Spatial Sound Events 2019 dataset [40]. This benchmark dataset is used for training and testing of the proposed CRNN system to assess the feasibility of SELD-RT. Additionally, the Real-Time Factor (RTF) is analyzed to provide a more accurate evaluation of the system's real-time performance.

SELD FEATURES

Most of the acoustic sound classification signals are monaural. However, in the case of multi-channel signals, events are well spatially distributed, and this spatial information can be used for localization of the event. The first joint SELD was introduced in DCASE 2019 task 3. In the aspect of research, SL and SED are different. In DCASE 2019 task 3 most common features for sound event detection (SED) were Log- mel, constant Q transform (CQT), and for sound localization (SL) GCC-PHAT, and intensity vector was used. Log-mel and CQT are like human auditory perception, which is based on a non-linear frequency scale.

To obtain the lag time between two microphones, GCC is used. GCC-PHAT removes the impact of the amplitude, and the cross-power spectrum is whitened, leaving only the phase. For robustness, the popular PHAT weighting scheme can be used to obtain a unity gain for all frequency components, while preserving phases which contain the actual delay information. The GCC-PHAT can be expressed as

$$GCC_{i,j}(t,\tau) = \mathcal{F}^{-1}_{f\to\tau}\frac{X_i(f,t)X_i^*(f,t)}{|X_i(f,t)||X_j(f,t)|} \quad (1)$$

Where $\mathcal{F}^{-1}$ [1] is the inverse-FFT from f to τ, $X_i(f,t)$ is the short-time Fourier Transform (STFT) of the *ith* microphone signal, and $(\cdot)^*$ denotes the conjugate.

For feature extraction, we use the librosa library. We use different features derived from MIC datasets. The features for SED are log mel energy and CQT, and for DOA calculation, GCC PHAT, and intensity vector are used. CQT is expressed as

$$CQT(\gamma) = f_{min} \times 2k\gamma/B \quad (2)$$

where f_{min} is the minimum frequency, γ is the filter index, and B is the number of bins per octave. We make four combination features for the SELD-RT task. To reduce RTF, we only consider the Mic dataset and four combinations of features. We have used these combinations because for Task 3 of DCASE 2019, most of the participants have also used these features.

From mic data-sets, we used the following combination of features ;

- Log Mel energy (4 channels) and intensity vector (3 channels) (LI)
- CQT (4 channels) and intensity (3 channels) vector (CI)
- Log Mel energy (4 channels) and GCCPHAT (6 channels) (LG)
- Log Mel energy (4 channels) , intensity vector (3 channels) and GCCPHAT (6 channels) (LIG)

The parameters for the extraction of the feature are listed in Table 4. As there is a trade-off between computational complexity and system performance [47]. We have used 64 mel bins for feature extraction, which is observed as the optimal filter size for the proposed model.

Table 3: *Comparison of Common SBC Devices for ASN Deployment*

SBC Device	Retail Price	CPU	RAM	Wi-Fi
Raspberry Pi 4	$59	Quad-core (1.5 GHz)	4 GB	Yes
ASUS Tinker Board	$70	Quad-core (1.8 GHz)	4 GB	Yes
BeagleBone Black	$55	Single-core (1 GHz)	0.5 GB	Yes
Odroid c1+	$37	Quad-Core (1.5 Ghz)	1 Gb	No

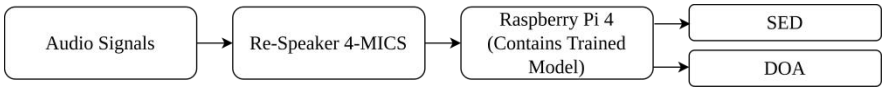

Figure 2: Block diagram of the Re-speaker model

MODEL ARCHITECTURE

We have trained a single-stage CRNN for SELD task implementation in a Raspberry Pi. By using variance and mean, they are normalized before feeding as input features to the single-stage CRNN-based model. CRNN-1 are inspired by VGGNeT, which has four group convolutional layers and has the ability to learn both inter and intra-channel features. Each group include two 2D Conv layers with a filter size of 3×3, a stride size of 1×1 and a padding size of 1×1 and average pooling of 2×2. We applied batch normalization to speed up the process and for stabilization. The Relu function is used as a non-linear function. After four group convolutional layers, we applied Global averaging pooling to reduce the computational cost in CRNN-1. Finally, the two layers of the GRU which are connected to the fully connected layers. GRU's are used to learn temporal information from audio. GRU consist of reset and update gates that can store relevant information and remove unnecessary information. These gates also help in removing the gradient descent problem. The output layer for SED is sigmoid as an activation function; the output will be in the [0,1] range. In order to improve overall and to minimize model over-fitting, we have used a data augmentation technique which is inspired from spec- augmentation known as spec-mix. For optimization, we have used Adam as optimizer and Binary Cross-Entropy Loss is used as a loss function. Our learning rate is set to 0.001 for 1000 iterations, and then it is reduced by 10 % and a batch size of 32 is used. Overview of the SELD-RT model CRNN is shown in Figure 3.

EXPERIMENTAL RESULTS AND DISCUSSIONS

We describe the experimental setup in this section, which makes use of the Pytorch module and the SELD dataset. On a 64-bit PC with an Intel Core i5-7300HQ quad-core CPU (6 MB cache, up to 3.5 GHz), 24 GB of 2400 MHz DDR4 RAM, 128 GB of SSD storage, 1 TB of 5400RPM SATA storage, and a GTX 1060 6 GB graphics card, all of the implementation programs for training were built in an experimental Python environment utilising Pytorch. While using the same Python environment and Pytorch for deep learning, the trained model is transferred to the Raspberry Pi 4 (OS Raspbian).

The proposed Single-stage CRNN contains three main parts, as shown in Figure 4. In the first part, we have extracted features on the Raspberry Pi and noted RTF of each feature. In the second part, we have trained our model by using a GPU and in the last part, we transfer our trained model to a Raspberry Pi 4 and inference our model on the Raspberry Pi 4. For real-time justification, we have used RTF, which is given in Equation 3.

$$RTF = \frac{T}{D} \tag{3}$$

where D is the input of duration, and T is the time. If RTF is less than 1, it means processing is in real time.

PERFORMANCE EVALUATION

SELD contains two subtasks, SED and DOA; therefore, separate evaluation metrics are used for each task. The SED performance is evaluated using the F-score and the error rate (ER) [40]. Ideally, the F1-score should be 1, and the ER should be 0. F1-Score is expressed as

$$F_1 - Score = \frac{2 \times \sum_{k=1}^{K} TP}{\sum_{k=1}^{K} TP + \sum_{k=1}^{K} FP + \sum_{k=1}^{K} FN} \tag{4}$$

Where K is the total number of window segments, and k is

the single window segment. TP(k) is the true positive, FP(k) is the false positive, and FN(k) is the false negative. ER is expressed as

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \tag{5}$$

Where $N(k)$ is the total number of active sound events in reference, $S(k)$ Substitution, $D(k)$ Deletion and $I(k)$ Insertion, defined as follows:

$$S(k) = \min\left(FN(k), FP(k)\right) \tag{6}$$

$$D(k) = \max\left(0, FN(k) - FP(k)\right) \tag{7}$$

$$I(k) = \max\left(0, FP(k) - FN(k)\right) \tag{8}$$

DOA error and frame recall are used to evaluate the performance of DOA [3].

$$DOA_{error} = \frac{1}{\sum_{t-1}^{T} D_k^t} \sum_{t=1}^{T} H(DOA_R \cdot DOA_E) \quad (9)$$

Where T is recording length, DOA_R Reference DOA_S at time frame t, DOA_E Estimated DOA_S at time frame t and H is the Hungarian Algorithm

$$Frame - recall = \sum_{t=1}^{T} I(D_R^t = D_E^t)/T \quad (10)$$

Where I is an indicator. Its output value is 1 when $D_R^t = D_E^t$ else it's 0.

Table 4: *Parameters for Features*

Parameter	Value
Sampling rate	32 KHz
Window size	1024
Hop size	320
Minimum frequency	52 Hz
Mel bins	64 bins

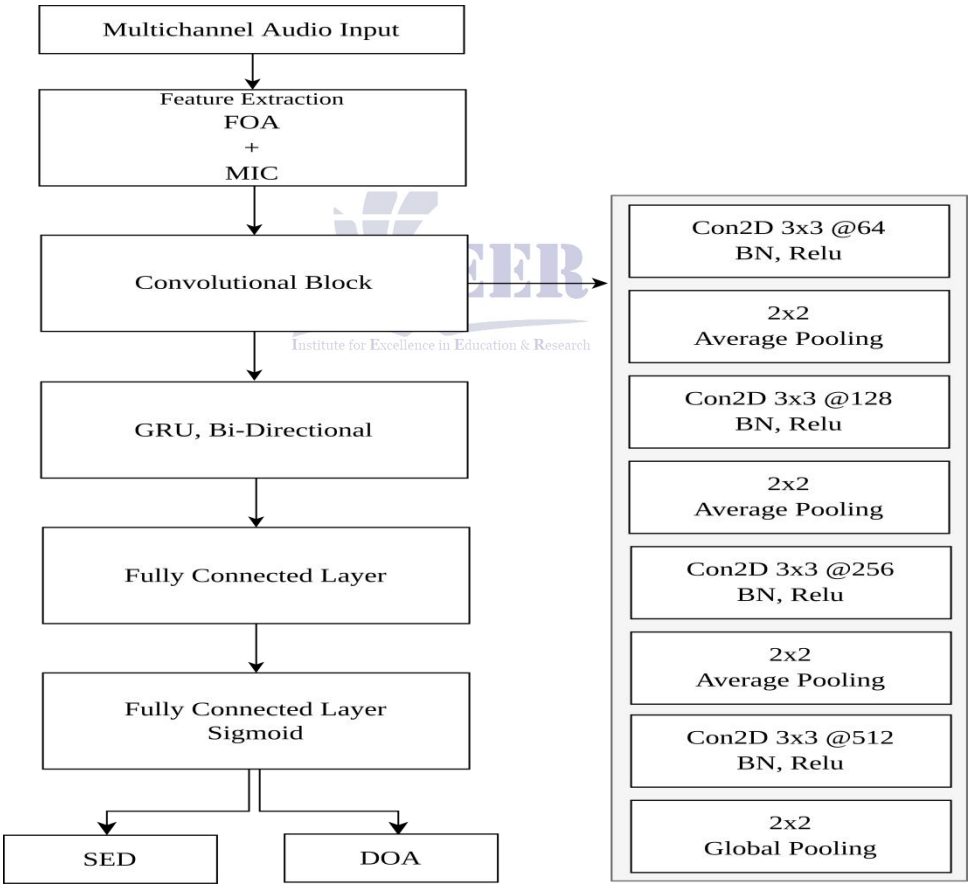

Figure 3: Overview of the SELD Model Convolutional Recurrent Neural Network

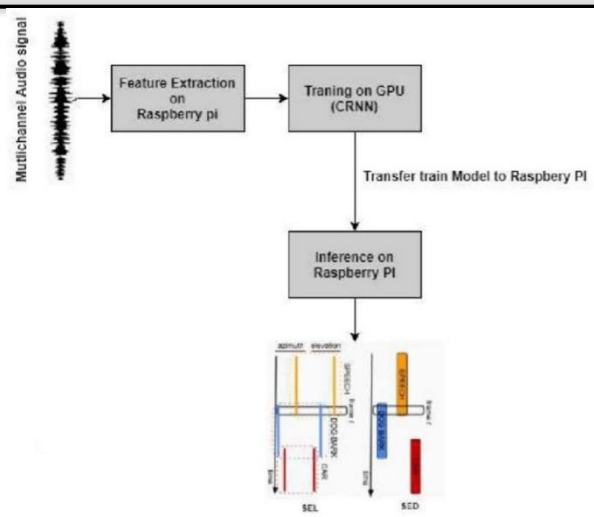

Figure 4: Block Diagram Of The Proposed System

VII. RESULTS AND DISCUSSIONS

As our main goal for the SELD task is to make it computationally efficient, therefore, we have devised our model for hardware efficiency and ease of implementation in Raspberry Pi. We have compared our results with the works of Cao. et al model [27] in terms of the best feature and real-time. As illustrated in Figure 5, the best feature for SED, which is used as input to CRNN-1a, is LG, having an F-score of 0.901 and an error rate of 0.163, while Cao. et al. model has an F-score of 0.90 and an error rate of 0.16. In DoA performance, the proposed model with CI as input feature has DOA error of 14.1^{O} and LG has a feature that has the best frame recall of 0.87, while Cao. et al. model has a DOA error of 9.8^{O} and a frame recall of and having of 0.863. The best RTF is achieved by LI 0.5, and the large RTF value is achieved by LGI has an RTF of 0.7, while Cao. et al. The Surrey model has achieved RTF of 1.2, which is not real-time. Furthermore, if we increase the number of channels, then the RTF value for inference also increases. Another observation is that although LI and CI have the same number of channels, i-e 7, LI is a more robust feature than CI.

All the features proposed, which is given as input to CRNN- 1a has RTF less than one, which shows their robustness. In their findings presented in [48], it was demonstrated that the DOA and frame recall were successfully achieved by solely utilizing the magnitude values of logmel energy. We have compared our model's results, which were also accomplished by utilizing magnitude information, specifically LI. Our LI results surpassed both the SED and SELD scores of [48] and performed favorable compared to the baseline of DCASE 2019 [40]. An error rate of 0.16 and an F1-score of 90.0 were achieved. Additionally, a DOA of 15.8 and a frame-recall of 88.0 were obtained.

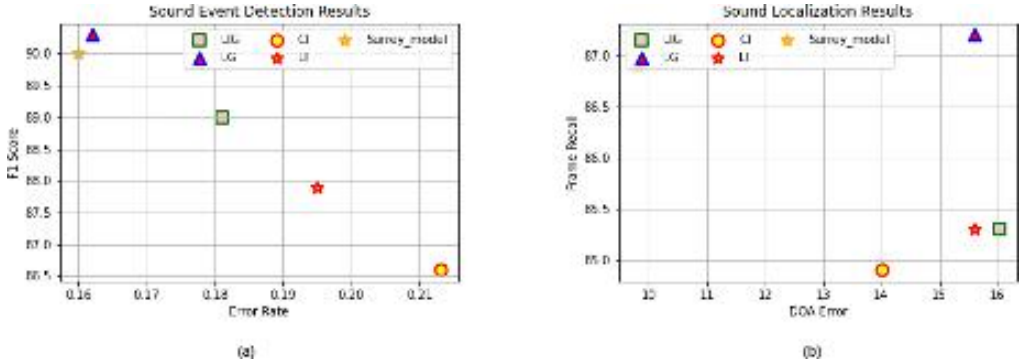

Figure 5: Overall Results of Proposed Model with RTF of different features. (a) Sound Event Detection Results, (b) Sound Localization Results

Table 5: *Comparison Using Magnitude As Input*

SELD Model	Error rate	F1-score	DOA	Frame-Recall
CNN5-Avg [48]	0.33	80.7	54.4	77
CNN9-Avg [48]	0.32	80.5	44.0	77.1
CNN9-Max [48]	0.34	79.4	45.6	76.3
CNN13-Avg [48] +	0.42	72.8	42.8	71.4
Baseline [40] +	0.34	79.9	28.5	85.4
Proposed CRNN LI-Feature +	0.16	90.0	15.8	88.0

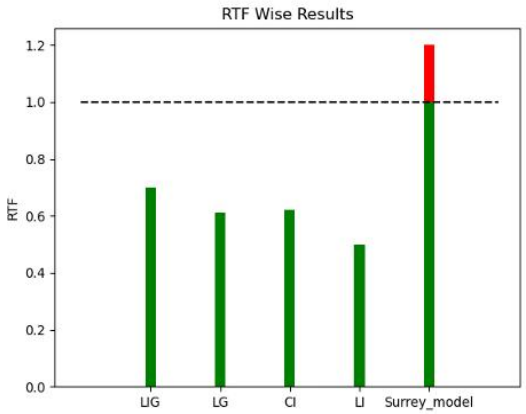

Figure 6: Overall Results of Proposed Model with RTF of different features

CONCLUSION

In this Paper, we have introduced more robust and hardware-friendly models that can use SBC devices like Raspberry Pi for the SELD task. We also study different DNN models and propose a novel single-stage model which can be deployed in a Raspberry Pi 4 to perform inference. Four different combinations of features are proposed, which are extracted from Mic datasets. All proposed features have RTF less than 1 on the Raspberry Pi as shown in Figure 5. Our experiment plays with the effect of GRU layers by increasing their layers, which boosts the performance of SELD. The spec-mix also increases SELD performance. The single-stage CRNN proposed is highly in favour of embedded systems and some useful applications like robotics, speech enhancement, etc., with low-power and low-cost resources. We have evaluated our model using the DCASE 2019 task 3 development dataset by comparing the F1 score, error rate, DOA error, DOA frame recall, and RTF.

REFERENCES

[1] Sharnil Pandya and Hemant Ghayvat. Ambient acoustic event assistive framework for identification, detection, and recognition of unknown acoustic events of a residence. Advanced Engineering Informatics, 47:101238, 2021.

[2] Siqi Zheng, Weilong Huang, Xianliang Wang, Hongbin Suo, Jinwei Feng, and Zhijie Yan. A real-time speaker diarization system based on spatial spectrum. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7208–7212. IEEE, 2021.

[3] Andrew Francl and Josh H McDermott. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. Nature human behaviour, 6(1):111–133, 2022.

[4] Oktay Akpolat. Deep learning and probabilistic approaches for financial time series analysis. The Florida State University, 2021.

[5] Vincent Lostanlen, Aurora Cramer, Justin Salamon, Andrew Farnsworth, Benjamin M Van Doren, Steve Kelling, and Juan Pablo Bello. Birdvox: Machine listening for bird migration monitoring. bioRxiv, pages 2022– 05, 2022.

[6] Rui Yan, Cheng Wen, Shuran Zhou, Tingwei Guo, Wei Zou, and Xiangang Li. Audio deepfake detection system with neural stitching for add 2022. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 9226–9230. IEEE, 2022.

[7] Lukas Grasse. Biologically-inspired auditory artificial intelligence for speech recognition in multi-talker environments. University of Lethbridge (Canada), 2021.

[8] Ashish Sutradhar, Md Harunur Rashid Bhuiyan, Faria Tashnim Mazumder, Pritom Goswami, Saiful Islam Salim, Tarik Reza Toha, and Shaikh Md Mominul Alam. Devising a dust and noise pollution monitoring system for textile industry. In Proceedings of the 8th International Conference on Networking, Systems and Security, pages 77–82, 2021.

[9] Félix Gontier, Vincent Lostanlen, Mathieu Lagrange, Nicolas Fortin, Catherine Lavandier, and Jean-François Petiot. Polyphonic training set synthesis improves self-supervised urban sound classification. The Journal of the Acoustical Society of America, 149(6):4309–4326, 2021.

[10] Panuwit Nantasri, Ekachai Phaisangittisagul, Jessada Karnjana, Surasak Boonkla, Suthum Keerativittayanun, Anocha Rugchatjaroen, Sasiporn Usanavasin, and Takahiro Shinozaki. A light-weight artificial neural network for speech emotion recognition using average values of mfccs and their derivatives. In 2020 17th International conference on Electrical Engineering/electronics, computer, telecommunications and information technology (ECTI-CON), pages 41–44. IEEE, 2020.

[11] Nam Kyun Kim and Hong Kook Kim. Polyphonic sound event detection based on residual convolutional recurrent neural network with semi-supervised loss function. IEEE Access, 9:7564–7575, 2021.

[12] Teck Kai Chan, Cheng Siong Chin, and Ye Li. Non-negative matrix factorization-convolutional neural network (nmf-cnn) for sound event detection. arXiv preprint arXiv:2001.07874, 2020.

[13] Muhammad Salman Khan, Mohsin Shah, Asfandyar Khan, Amjad Ald- weesh, Mushtaq Ali, Elsayed Tag Eldin, Waqar Ishaq, and Lal Hussain. Improved multi-model classification technique for sound event detection in urban environments. Applied Sciences, 12(19):9907, 2022.

[14] Yang Bai, Li Lu, Jerry Cheng, Jian Liu, Yingying Chen, and Jiadi Yu. Acoustic-based sensing and applications: A survey. Computer Networks, 181:107447, 2020.

[15] Kashif Ahmad and Nicola Conci. How deep features have improved event recognition in multimedia: A survey. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(2):1–27, 2019.

[16] Inas Abuqaddom, Basel A Mahafzah, and Hossam Faris. Oriented stochastic loss descent algorithm to train very deep multi-layer neural networks without vanishing gradients. Knowledge-Based Systems, 230:107391, 2021.

[17] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. Artificial Intelligence Review, 53(8):5929–5955, 2020.

[18] Ays¸egül Özkaya Eren and Mustafa Sert. Audio captioning using gated recurrent units. arxiv preprint arXiv:2006.03391, 2020.

[19] Jacob R Stevens, Ashish Ranjan, Dipankar Das, Bharat Kaul, and Anand Raghunathan. Manna: An accelerator for memory-augmented neural networks. In Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, pages 794–806, 2019.

[20] De-Bing Zhuo and Hui Cao. Fast sound source localization based on srp- phat using density peaks clustering. Applied Sciences, 11(1):445, 2021.

[21] Na Zhu and Tamim Reza. A modified cross-

correlation algorithm to achieve the time difference of arrival in sound source localization. Measurement and Control, 52(3-4):212–221, 2019.

[22] Ruchi Pandey, Santosh Nannuru, and Aditya Siripuram. Sparse bayesian learning for acoustic source localization. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4670–4674. IEEE, 2021.

[23] Amir Masoud Molaei, Bijan Zakeri, and Seyed Mehdi Hosseini Andargoli. A one-step algorithm for mixed far-field and near-field sources localization. Digital Signal Processing, 108:102899, 2021.

[24] Marco Sewtz, Tim Bodenmüller, and Rudolph Triebel. Robust music-based sound source localization in reverberant and echoic environments. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2474–2480. IEEE, 2020.

[25] Sławomir Kapka and Mateusz Lewandowski. Sound source detection, localization and classification using consecutive ensemble of crnn models. arxiv preprint arXiv:1908.00766, 2019.

[26] Yin Cao, Turab Iqbal, Qiuqiang Kong, Miguel Galindo, Wenwu Wang, and M Plumbley. Two-stage sound event localization and detection using intensity vector and generalized cross-correlation. DCASE2019 Challenge, Tech. Rep, 2019.

[27] Wei Xue, T Ying, Z Chao, and D Guohong. Multi-beam and multi- task learning for joint sound event detection and localization. DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Challenge, 2019.

[28] Jingyang Zhang, Wenhao Ding, and Liang He. Data augmentation and prior knowledge-based regularization for sound event localization and detection. DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Challenge, 2019.

[29] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779, 2019.

[30] Thi Ngoc Tho Nguyen, Douglas L Jones, Rishabh Ranjan, Sathish Jaya-balan, and Woon Seng Gan. A two-step system for sound event localization and detection. arXiv preprint arXiv:1911.11373, 2019.

[31] Qing Wang, Huaxin Wu, Zijun Jing, Feng Ma, Yi Fang, Yuxuan Wang, Tairan Chen, Jia Pan, Jun Du, and Chin-Hui Lee. The ustc-iflytek system for sound event localization and detection of dcase2020 challenge. IEEE AASP Chall. Detect. Classif. Acoust. Scenes Events, 2020.

[32] Thi Ngoc Tho Nguyen, Douglas L Jones, and Woon-Seng Gan. Ensemble of sequence matching networks for dynamic sound event localization, detection, and tracking. In DCASE, pages 120–124, 2020.

[33] Kazuki Shimada, Naoya Takahashi, Yuichiro Koyama, Shusuke Takahashi, Emiru Tsunoo, Masafumi Takahashi, and Yuki Mitsufuji. Ensemble of accdoa-and einv2-based systems with d3nets and impulse response simulation for sound event localization and detection. arXiv preprint arXiv:2106.10806, 2021.

[34] Thi Ngoc Tho Nguyen, Karn Watcharasupat, Ngoc Khanh Nguyen, Douglas L Jones, and Woon Seng Gan. Dcase 2021 task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection. arXiv preprint arXiv:2106.15190, 2021.

[35] Sang-Ick Kang, Kyongil Cho, Myungchul Keum, and Yeonseok Park. The distillation system for sound event localization and detection of dcase2023 challenge. Technical report, DCASE2023 Challenge, June 2023.

[36] Gwantae Kim and Hanseok Ko. Data augmentation, neural networks, and ensemble methods for sound event localization and detection. Tech. Report of DCASE Challenge, 2023.

[37] Yuxuan Dong, Qing Wang, Hengyi Hong, Ya Jiang, and Shi Cheng. An experimental study on joint modeling for sound event localization and detection

with source distance estimation. arXiv preprint arXiv:2501.10755, 2025.

[38] Jun Wei Yeow, Ee-Leng Tan, Jisheng Bai, Santi Peksi, and Woon-Seng Gan. Squeeze-and-excite resnet-conformers for sound event localization, detection, and distance estimation for dcase 2024 challenge. arXiv preprint arXiv:2407.09021, 2024.

[39] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. A multi-room reverberant dataset for sound event localization and detection. arXiv preprint arXiv:1905.08546, 2019.

[40] Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D. Plumbley. Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(2):379–393, 2017.

[41] Douglas Manvell and Phil Stollery. New techniques to determine specific noise for increasing the effectiveness of continuous unattended noise monitoring systems. In INTER-NOISE and NOISE-CON Congress and Conference Proceedings, volume 249, pages 3642–3649. Institute of Noise Control Engineering, 2014.

[42] Ilkka Kivelä, Chao Gao, Jari Luomala, Jukka Ihalainen, and Ismo Hakala. Design of networked low-cost wireless noise measurement sensors (issn 1726-5479). International Journal on Sensors & Transducers, 9:171–190, 2010.

[43] Juan P Bello, Claudio Silva, Oded Nov, R Luke Dubois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. Communications of the ACM, 62(2):68–77, 2019.

[44] Petr Gallus and Petr Frantis. Security analysis of the Raspbian Linux operating system and its settings to increase resilience against attacks via network interface. In 2021 International Conference on Military Technologies (ICMT), pages 1–5. IEEE, 2021.

[45] Bharath Sudharsan, Sree Prem Kumar, and Rakesh Dhakshinamurthy. Ai vision: Smart speaker design and implementation with object detection custom skill and advanced voice interaction capability. In 2019 11th International Conference on Advanced Computing (ICoAC), pages 97–

[46] 102. IEEE, 2019.

[47] Qiuqiang Kong, Keunwoo Choi, and Yuxuan Wang. Large-scale midi-based composer classification. arXiv preprint arXiv:2010.14805, 2020.

[48] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D. Plumbley. Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems. arXiv preprint arXiv:1904.03476, 2019.