AUDITORY SCENE ANALYSIS USING DEEP LEARNING APPROACHES

Thesis

Submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

SPOORTHY. V

(187126CO007)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE - 575 025

August, 2024

AUDITORY SCENE ANALYSIS USING DEEP LEARNING APPROACHES

Thesis

Submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

SPOORTHY. V

(187126CO007)

Under the guidance of

Prof. SHASHIDHAR G. KOOLAGUDI

Dept of CSE, NITK

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE - 575 025

August, 2024

DECLARATION

by the Ph.D. Research Scholar

I hereby declare that the Research Thesis entitled **Auditory Scene Analysis using Deep Learning Approaches** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy** in Department of Computer Science and Engineering is a bonafide report of the research work carried out by me. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

Spoorthy. V, 187126 187CO007

Department of Computer Science and Engineering

Place: NITK, Surathkal

Date: August, 2024

CERTIFICATE

This is to certify that the Research Thesis entitled **Auditory Scene Analysis using Deep Learning Approaches** submitted by **Spoorthy. V** (Register Number: 187126 187CO007) as the record of the research work carried out by her, is accepted as the Research Thesis submission in partial fulfilment of the requirements for the award of degree of **Doctor of Philosophy**.

16.8.24

Prof. Shashidhar G. Koolagudi

Research Supervisor-

Shashidhar G. Koolagudi
शशिधर ग । कुलगुडि

(Signature with Date and Seal)

Prof. K.Sreenivasa Rao

External Examiner

(Signature with Date)

19/8/24

Chairman - DRPC

(Signature with Date and Seal)

Chairman
DUGC / DPGC / DRPC
Dept. of Computer Science & Engineering
NITK- Surathkal
Srinivasnagar - 575 025

ACKNOWLEDGEMENTS

Giving thanks is the final step in completing my dissertation. In terms of both my technical and personal development, research has made me a better person. Without the help of the people who have assisted me, this kind of progress would not have been possible.

I want to start by sincerely thanking my research supervisor, Prof. Shashidhar G. Koolagudi, for his unwavering encouragement, direction, and support throughout the course of my Ph.D. and related research. I can honestly say that his extensive knowledge, support, and patience helped me get to this point. I appreciate him seeing my potential and offering me the chance to work with him and pursue my research under his guidance. The amount of consideration and time he has given during the stages of drafting research papers and thesis writing is impeccable. I could not have asked for a better mentor and advisor for my Ph.D. studies.

I would like to express my gratitude to Dr. Bhawana Rudra and Dr. Biswajit R. Bhowmik, members of my Research Progress Committee (RPAC), for their insightful reviews and recommendations on how to improve my work further. They made some incredibly insightful points that gave me new perspective on my work. I wish to show my appreciation to the Head of the Department Dr. Manu Basavaraju, former Head of the Department Dr. Shashidhar G Koolagudi and other faculty members for their continuous support. I appreciate the assistance of the entire teaching and non-teaching personnel of the NITK Computer Science Department throughout my research period.

The love, concern, and support I received from my closest friends Shubham, Rashmi, and Garima will always be treasured. Without their assistance, it would be impossible to even consider reaching this goal. I appreciate the technical and emotional assistance you provided. I'd like to thank my lab mates Pradyoth,

Nagarathna, Pravin, Alkha, Sneha, Swathi, and Vishal for the beneficial discussions, knowledge sharing, and collaborative efforts. I want to express my sincere gratitude to every teacher that instructed me in both high school and college. Without them, I would not be able to write this. I want to express my gratitude to Dr. Veena Thenkanidiyoor from the National Institute of Technology Goa in particular for introducing me to research in the first place.

Last but not least, I want to express my gratitude to my father Mr. C. Venkatesh, mother Mrs. Vasantha T, and sister Ms. Sindhura for their support and affection throughout my life. Words cannot explain how difficult those sacrifices were or how grateful I am for them. My family had to make numerous sacrifices for my studies and success. The person I want to thank the most is Mr. Vimal Maru, who has been my loving and supportive partner and who has always been there for me.

Finally, I thank the Almighty for giving me the strength and vision to move on with assurance. Thank you God.

Spoorthy. V

Place: NITK, Surathkal

Date: August, 2024

ABSTRACT

Auditory scene analysis (ASA) is a fundamental skill of the system that allows us to perceive and identify acoustic events in the environment around us. Automating ASA through computational devices such as hand-held smartphones or laptops is known as Computational Auditory Scene Analysis (CASA). This research is motivated by the significant number of real-time applications that ASA has. ASA can be used in context-aware mobile devices where the device can turn to silent mode when the owner enters a meeting or an ICU of a hospital. The environment/location in which a particular audio is recorded is known as an acoustic scene. The sounds occurring in a particular scene/location are called sound events. A combination of two or more acoustic events forms one acoustic scene. For example, given a meeting room as a scene, the sound events present in the scene are keyboard typing, mouse clicks, somebody speaking, and so on. The task of identifying the events present in an acoustic scene is known as Sound Event Detection (SED), and identifying the location of the source of the sound along with the event type is known as Sound Source Localization (SSL). Acoustic Scene Classification (ASC) is identifying a scene using sound cues and assigning a label to this scene.

In this thesis, three ASA tasks have been investigated, namely, SED, SSL, and ASC. One major challenge in identifying sound events is when they overlap at a given point in time. This type of event detection is said to be Polyphonic Sound Event Detection (PSED). In the existing works, the results obtained for the PSED task are less, and there is a considerable scope to improve the results. In this thesis, two new methods are proposed to perform PSED using spectral features and deep learning techniques. To perform PSED, a Mel-pseudo-based Constant Q-transform is proposed. The dataset considered to perform this task is TUT-Sound Event Detection (SED) 2016. The method resulted in an F1 score of 54% and an Error rate of 0.66. Once the event is detected, it is necessary to identify the source of the event. The presence of noise or the distance of the source can majorly affect the performance of the SSL.

Therefore, this work proposes Sound Event Localization and Detection (SELD) systems to estimate the Direction-of-Arrival (DOA) of the sound event and event type. In this research work, a channel-wise 'FusionNet' deep learning network is designed to perform the SELD task. The proposed model performs the tasks of SED and DOA estimation in one neural network model. The dataset considered to perform this task is TAU-NIGENS Spatial Sound Events 2020. The method resulted in an F-score of 81.2%, an Error rate of 0.23, and a Frame recall of 86.9%. Accurate event detection and localization in a particular surrounding will make identification of the scene a more straightforward task. However, a critical challenge in ASC is when the recording devices are different. In this case, there is a high chance of device distortion present in the audio recordings. Therefore, a device-robust ASC method is proposed to eliminate the device distortion in the audio recordings and improve the performance of the ASC task. Also, a different deep learning approach named Deep Fisher Network is also proposed to perform ASC. This method combines the working principles of traditional machine learning algorithms and deep learning algorithms. The dataset considered to perform this task is DCASE (Detection and Classification of Acoustic Scenes and Events) 2019 ASC Task 1(a). The best average accuracy achieved is 91%. Detailed experimental evaluation is carried out to compare the performance of each of the proposed approaches against baseline and state-of-the-art systems.

Keywords: Auditory Scene Analysis (ASA), Polyphonic Sound Event Detection (PSED), Sound Source Localization (SSL), Sound Event Localization and Detection (SELD), Acoustic Scene Classification (ASC), Time-Frequency Representations (TFRs), Deep Learning.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviations	Expansion
ACC	Accuracy
AED	Acoustic Event Detection
AE-CLBP	Adjacent Evaluation Completed Local Binary Patterns
ASA	Auditory Scene Analysis
ASC	Acoustic Scene Classification
ASR	Automatic Speech Recognition
CASA	Computational Auditory Scene Analysis
CNN	Convolutional Neural Network
CRANN	Convolutional Recurrent Attention Neural Network
CRNN	Convolutional Recurrent Neural Network
CQT	Constant Q-Transform
DCASE	Detection and Classification of Acoustic Scenes and Events
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DOA	Direction-of-Arrival
DRC	Dynamic Range Control
D-SELDNet	Depthwise Sound Event Localization and Detection Network
ER	Error Rate
FFNN	Feed Forward Neural Networks
FN	False Negative
FP	False Positive
FOA	First Order Ambisonics
FV	Fisher Vector
GCC	Generalized Cross-Correlation
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Units
GTCC	Gammatone Time Cepstral Coefficients
HIK	Histogram Intersection Kernel (HIK)
HMM	Hidden Markov Model
HoG	Histogram of Gradients
HPF	High-Pass Filter
HPSS	Harmonic Percussive Source Separation

Abbreviations	Expansion
IIR	Infinite Impulse Response
IR	Impulse Response
LBP	Local Binary Patterns
LK	Linear Kernel
LSTM	Long-Short Term Memory
LW-CRANN	Light-weight Convolutional Recurrent Attention Neural Network
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multi-Layer Perceptron
MP- CQT	Mel-Pseudo Constant Q-Transform
MR-TPNN	Modified Recurrent Temporal Pyramid Neural Network
MSE	Mean Squared Error
NMF	Non-Negative Matrix Factorization
PCA	Principal Component Analysis
PRE	Precision
PSED	Polyphonic Sound Event Detection
RBM	Restricted Boltzmann Machine
REC	Recall
ReLU	Rectified Linear Unit
RF	Random Forest
RIR	Room Impulse Responses
RNN	Recurrent Neural Network
SED	Sound Event Detection
SELD	Sound Event Localization and Detection
SELDNet	Sound Event Localization and Detection Network
SE-MobileNet	Squeeze-and-Excitation Mobile Network
SE-Net	Squeeze-and-Excitation Network
SSL	Sound Source Localization
SSR	Signed Square Root
STFT	Short-Time Fourier Transform
SVM	Support Vector Machines
TCRNN	Transposed Convolutional Recurrent Neural Network
TDOA	Time Difference of Arrival
TFR	Time-Frequency Representation
TN	True Negative
TP	True Positive
TPP	Temporal Pyramid Pooling
TUT	Tampere University of Technology
T-SELDNet	Transpose Sound Event Localization and Detection Network

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Humans have developed the ability to identify and locate various sound events in their environment solely through audio. This is known as Auditory Scene Analysis (ASA). Identifying a scene from audio recordings is a challenging task and is treated as the Computational Auditory Scene Analysis (CASA) (Wang and Brown, 2006). In recent years, smart technological devices have integrated into everyday life. Future human-device interactions will only improve and become more deeply intertwined. Machine audition/listening can provide one of the dimensions for interaction by enabling electronic devices to perceive and organically engage with the auditory environment around them. Context awareness is the ability of a device to automatically perceive and comprehend the events taking place around it without human input (Schilit et al., 1994). There are many real-life applications of ASA, namely smart homes, robotics, audio surveillance, context-aware mobile devices, music genre classification, etc. One of the important applications is detecting outlier activities such as gunshot or screaming in a particular indoor environment, etc. This can be done using audio surveillance that employs sound content analysis techniques for the detection of outliers (Peltonen et al., 2001).

In an acoustic situation when several sound events are occurring simultaneously and in close proximity to one another, it is difficult to perform recognition and localization tasks. The current technology does not have the complete ability to identify a sound or

the direction of the sound origin. Machines will be one step closer to human listening if they can recognize and locate sounds, making themselves audio-context-aware. For instance, a hearing-impaired person can benefit from an audio-context-aware device that helps them visualize sound occurrences. The main objective of this thesis is to create techniques that are inspired by the human auditory system to improve the performance to recognize polyphonic sound events, localize the origin of each sound event, and assign a semantic label to each event and its corresponding scene.

The three ASA tasks that are investigated in this thesis are Polyphonic Sound Event Detection (PSED), Sound Source Localization (SSL), and Acoustic Scene Classification (ASC). The details of these ASA tasks are given in below sub-sections:

1.1.1 Polyphonic Sound Event Detection

A recognizable sound activity in a particular audio is labelled as a sound event. These events act as good descriptors in an auditory scene, as they are used in understanding and analysing different scenes. Processing of sound events helps in identifying an auditory scene, for example, a restaurant scene consists of sound events such as cutlery, water running, people chatting, dish clinking, etc. The characterization of audio events in audio is known as Acoustic/Sound Event Detection (AED/SED). Acoustic events are divided into monophonic (isolated) and polyphonic (mixed) events. In real-life scenarios, polyphonic sound events are more commonly found and it is quite a challenging task to identify overlapped events as compared to non-overlapped events. It is also a difficult task to get an ordered sequence of the sound events in the case of overlapping. In both problems, based on the events present, a semantic label is attributed to an audio signal/scene. A basic architecture of a SED system is shown in Figure 1.1.

The audio signal input used for tasks like Automatic Speech Recognition (ASR) are in structured manner which means that the phonemes in a word or words in a sentence are arranged in a particular order to make it meaningful. However in real-life, the challenge incurred in the SED task is that the sound events present in an audio signal has no particular order for the occurrence. The random occurrence of a sound

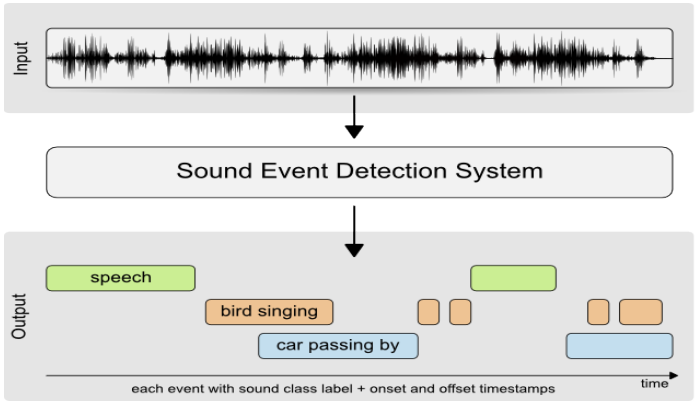

Figure 1.1: Block diagram of the Sound Event Detection (SED) system (Mesaros et al. 2016b)

event makes it difficult to identify the presence of a sound event. Also, there is a high degree of variation even in the same class of a sound event (Dennis 2014). Background noise and overlapping of multiple sound events are the major factors causing the variation. Additionally, events belonging to the same class can consist of different acoustic properties. For example, car horn can be different for different cars or dogs' bark may have different acoustic properties.

The main purpose of polyphonic SED is to identify the onset and offset times of a sound event and assign a corresponding label to the event in an audio recording. An SED system typically consists of two steps: Feature extraction and classification. The most common features used in SED task are frame-level features such as Mel-Frequency Cepstral Coefficients (MFCCs), Mel Spectrogram, and so on. These features are extracted from the input audio and fed to a classifier to get the semantic label for different sound events that are present in that particular frame. The evaluation of the proposed method is performed using two metrics, namely, F1 score and Error Rate (ER) (Mesaros et al. 2016a). F1 score is a metric that provides model's accuracy. This is computed using precision and recall. ER refers to a measurement of the degree of a model's prediction error.

1.1.2 Sound Source Localization

Many sound events are produced from different sources at the same time in a realistic setting, and they are distributed in space. For instance, if there is a fire outbreak in a

building, the sound events such as screaming for help, fire alarm, crying baby, etc., occurs simultaneously. This is an example of a critical scenario where sound events need to be identified and the place of source is to be detection (localization). The process of localizing the corresponding sources from where the sounds are being emitted (Direction-of-Arrival (DOA) estimation) is known as SSL. The joint task of detecting events and localizing them is Sound Event Localization and Detection (SELD). A basic architecture of an SELD system is shown in Figure 1.2.

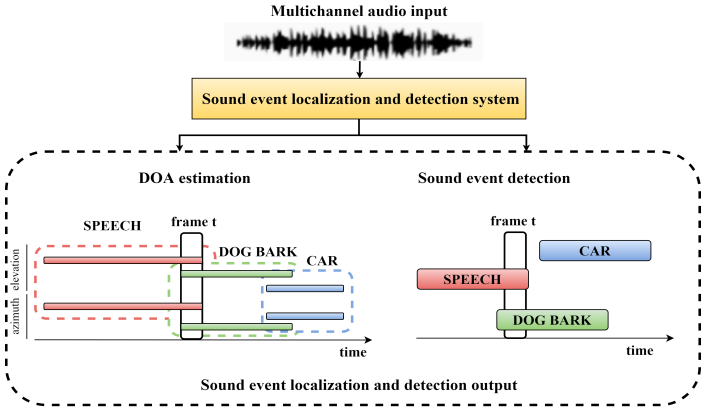

Figure 1.2: Block diagram of the Sound Event Localization and Detection (SELD) system

The DOA estimation of a particular sound event is performed using Azimuth and elevation angle. An azimuth is the angular measurement in a spherical coordinate system which represents the horizontal angle from a cardinal direction, most commonly north, and elevation angle is the angle between the horizontal plane and the line of sight, measured in the vertical plane. The calculation of Azimuth and Elevation angle is shown in Figure 1.3.

In the real-time environment, along with detecting an individual event and its source, it is also important to know the simultaneous or overlapped events. When several sound events are active simultaneously, providing a "label" to the event and adding the "location" of an auditory source in a scene is polyphonic SELD. A SELD task is more complex and challenging than performing SED or SSL individually. Two sets of metrics are employed to evaluate the performance of an SELD system. The first set of metrics was related to location-aware detection. If the prediction and the original reference have the same event class and the difference between them is less than the

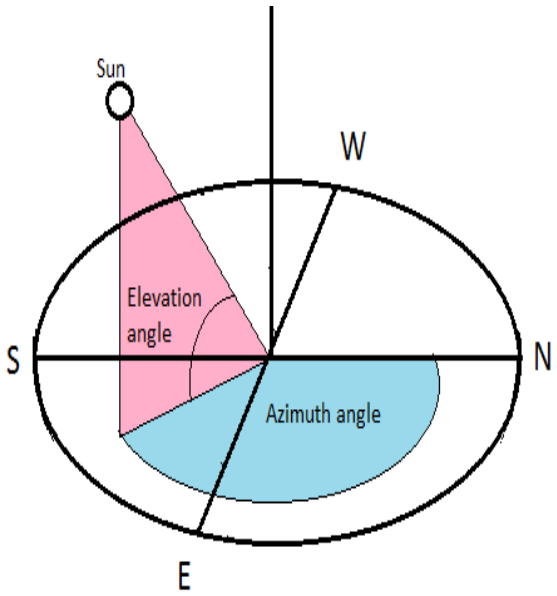

Figure 1.3: Illustration of Azimuth and Elevation angle (Patel and Patel 2023)

threshold value, the prediction is given a true value in these metrics. Error rate (ER) and F-score (F) are the corresponding metrics measured over one-second non-overlapping segments. In the ideal case, the error rate will be 0, and F-score will be 100% (Mesaros et al. 2019).

The second set of metrics is class-aware localization. The first is the localization error expressing average angular distance between predictions and references of the same class. The second is the localization recall metric expressing the true positive rate of how many of these localization estimates were detected in a class, out of the total class instances. In these metrics, the error is calculated between same-class predictions and references. Here, there is no distance threshold as in location-aware detection. Similar to the above metrics, in the ideal case of class-aware localization metrics, the localization error will be 0, and the localization recall value will be 100%.

The computation of ER is similar to Equation 2.4 given in subsection 1.1.1 and F is given in Equation 1.4.

In Equation 1.1, N denotes the active sound event classes in the reference labels, and S denotes substitution, i.e., the number of times the event was observed but at the wrong level. This is computed by combining the FPs(α) and FNs (γ) without correlating

which are FP substitutes and which FP. For each one-second segment t, the remaining FPs and FNs are inserted as I and deleted as D. The mathematical computations of S, D, and I are given by following Equations 1.1, 1.2, and 1.3.

$$S(t) = min(\gamma(t), \alpha(t)) \tag{1.1}$$

$$D(t) = max(0, (\gamma(t) - \alpha(t)) \tag{1.2}$$

$$I(t) = max(0, (\alpha(t) - \gamma(t)) \tag{1.3}$$

$$F = \frac{2.\sum_{t=1}^{T} TP(t)}{2.\sum_{t=1}^{T} TP(t) + \sum_{t=1}^{T} FP(t) + \sum_{t=1}^{T} FN(t)} \tag{1.4}$$

In Equation 1.4, TP denotes true positives, i.e., the number of active classes for both predictions and reference labels at the t^{th} one-second segment, FP denotes false positives, i.e., the sound event classes that were inactive in reference labels but active in the predictions, lastly, False Negatives (FN) are the opposite of FP, i.e., sound event classes that were inactive in predictions but active in reference labels.

The corresponding localization error and localization recall are computed as given in Equations 1.5 and 1.7.

$$FrameRecall = \frac{TP}{TP + FN} \tag{1.5}$$

The reference DOA estimates are given by x_G, y_G, z_G and the predicted DOA estimates are given by x_E, y_E, z_E, these are used to synthesize the dataset by utilizing the central angle $\sigma \epsilon [0, 180]$ computed using Equation 1.6. At the origin in degrees, the σ formed by the reference and predicted DOAs and is given by Equation 1.7.

$$\sigma = 2.arcsin\left(\frac{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}{2}\right).\frac{180}{\pi} \tag{1.6}$$

$$DOAError = \frac{1}{D}.\sum_{d=1}^{D} \sigma\left(\left(x_G^d, y_G^d, z_G^d\right), \left(x_E^d, y_E^d, z_E^d\right)\right) \tag{1.7}$$

1.1.3 Acoustic Scene Classification

Acoustic Scene Classification (ASC) is referred to as the task of identifying the environment in which the scene has been recorded. The scenes comprise the categories such as indoor (residence, restaurant/cafe), outdoor (park, metro station), and transportation (metro, bus, tram) (Barchiesi et al. 2015). This can also be said as making sense to the sounds or providing context to the environmental sound to make smarter devices. ASC plays an important role in the current generation where every device is automated. The main aim of ASC is to identify the environment in which an audio stream has been produced. Performance of an ASC system is mainly dependent on the high-level (audio clip level) representation of audio recordings and choice of features. A basic architecture of an ASC system is shown in Figure 1.4. ASC can be performed in two levels: Broader and finer levels.

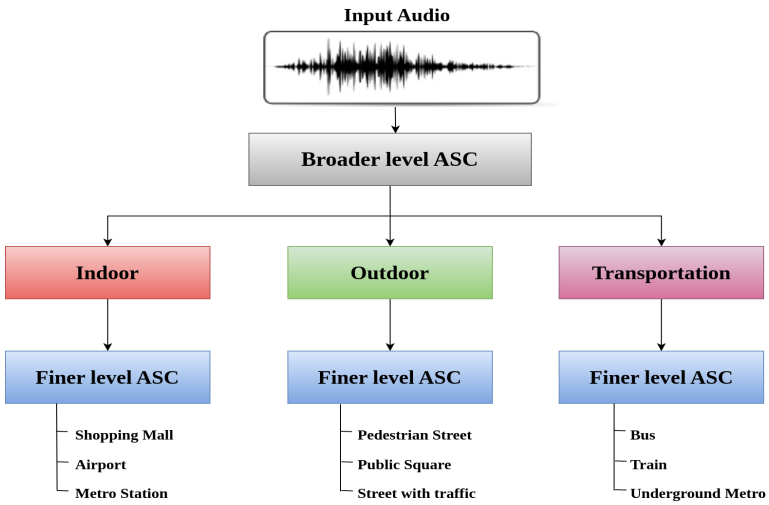

Figure 1.4: Block diagram of the Acoustic Scene Classification (ASC) system

The characteristics of acoustic scenes as signals are diverse and less structured as compared to normal music or speech signals in terms of structure of harmonics, stationarity of the signal and so on. The spectral and temporal structures of these audio samples have high variability. Therefore, there is a need to develop an efficient feature

extraction approach to characterize a surrounding scene.

Humans can differentiate acoustic scenes by using specific event cues that most commonly occur in a scenario (Peltonen et al. 2001). High variability and overlapping of multiple events in an acoustic scene make their classification more complex. Suitable classification approach is an essential need of the day for efficient classification.

The metric that is considered for evaluation on an ASC system is Accuracy (ACC). The accuracy metric gives the correct predictions of the model out of the total samples present in the dataset.

1.2 MOTIVATION

The research that is being conducted on CASA is still in its infant stages. There are various real-time applications on ASA. Making machines intelligent and sensitive to the surrounding environment is an important untrodden path in the contemporary research. Machine listening is a general area that has driven the attention of many researchers in the recent decade. The primary motivation behind choosing this research problem is that machine listening is the state-of-the-art technology that offers convenience to us. The tasks that are involved in machine listening have wide range of applications from day-to-day lives and in critical applications like military, Intensive Care Unit (ICU) in a hospital, etc. Nowadays, the majority of smart home gadgets recognize voice commands. They are also more intelligent and can hear sounds around them.

The design and development of machine listening systems has attracted a lot of attention over the last ten years. Detection and Classification of Acoustic Scenes and Events (DCASE) (Mesaros and Serizel 2013) is one such community consisting of publicly available datasets for different ASA tasks. Even though there are many contributions for various tasks, there are numerous challenges that we encounter in development of a real-time system that can sense the surrounding sounds.

One of the major challenges we encounter in ASA is the overlapping of sound events in an acoustic scene. For example, consider a scene in a bus, there can be people

talking on the bus or phone ringing in any scenario. Classifying this type of scene can be a complex task for a machine to separate sound events that occur simultaneously. There is much room for improvement in the performance of these systems by the use of different feature extraction and classification techniques, as the task of SED and ASC in the current approaches is limited to only specific feature representations. The existing approaches still need to address the issue of source localization of the audio recordings. It is critical in some surroundings to identify the source of the sound and also the type of sound. The device and the place the audio recording takes place play a major role in the performance of the SED and ASC system. Very limited research is present in the existing literature for device robust ASA systems. Thus, these limitations and challenges are the motivation for choosing this topic for my research work.

1.3 APPLICATIONS

There are many real-life applications of ASA, namely smart homes (Sehili et al. 2012), robotics (Aziz et al. 2019), audio surveillance (Chandrakala and Jayalakshmi 2019), context-aware mobile devices or personal digital assistants (Ma et al. 2003), music genre classification (Zieliński and Lee 2019), etc. In order to function without human intervention, electronic devices in smart homes must be able to detect sounds. Voice-based assistants are increasingly being utilized to manage various appliances in the home setting. ASA is to give robots a sense of their environment so that the necessary decisions would be taken by them.

Analysis of the surrounding audio environment may provide context-awareness capabilities to the devices when geolocalization and visual data are not available (Serizel et al. 2016). Smartphones or hand-held devices can sense their surroundings and switch to different modes. For example, suppose a person enters a meeting room or a hospital ICU, the smartphone should automatically switch to silent mode, or if any unusual sound, such as gun shot or a blast, is heard, the hand-held device should detect the sound as an unusual event/sound or not and essential alert indication is to be given. The devices should understand whether the person is in indoor or outdoor environments to control robotic wheelchairs. From a research perspective, the problem

of ASA is interesting as well as challenging and will be able to solve a real-life problem. Given an audio input, it is quite a challenging task to identify the scene for certain recordings as it can have multiple interpretations.

1.4 CHALLENGES

The field of ASA has many inherent challenges mainly because of its infancy. Some of the important ones are briefed below:

- **Unstructured signal type:** Analysis of an acoustic scene by using audio cues is a more challenging task compared to that of music or speech signals. The audio captured for any event detection or scene classification task generally does not contain any structured information like formants, cycles, harmonics, etc. in it. A speech signal normally consists of phonetic sequence (AbdelHamid et al. 2014) as opposed to a signal of environmental sounds. Additionally in a scene or event signal, there can be repeated isolated or overlapped sound events (Gemmeke et al. 2013). This makes the task of processing audio signal more challenging. Supportive models like language models are presently not known for audio signals of scenes for improving the performance.

- **Sound event origin in a scene:** The identification of the source of a sound event occurring in a particular scene is essential in the cases of critical events like gunshot or a person screaming. In such cases, it is necessary to locate from which place a particular sound event is occurring (Beltrán et al. 2015) for forensic needs. One of the major challenges in ASA is to localize the sound source along with identifying the type of event occurring in a particular surrounding.

- **Varying recording devices:** In the existing systems of ASA, the audio samples used are recorded with only one recording device. In a real-world scenario, it is always possible to encounter data recorded using different recording devices and hence an ASC system must be robust and device-independent to identify various acoustic scenes (Ozer et al. 2018). However recently, the data recorded with

multiple recording devices is made publicly available by DCASE challenges (Mesaros et al. 2018). This poses other challenges like domain adaptation (Gharib et al. 2018), data imbalance, etc.

- **ASA for low resource device applications:** It is a real challenge to deploy an ASA system in real-time due to issues like model interpretability and ambiguous event & scene annotations (Mesaros et al. 2018). The existing systems developed for ASA are built using deep learning paradigms, which are computationally expensive, making it difficult to deploy on the hand-held smart devices like mobile phones. Therefore, developing low-complex device robust SED, SSL, and ASC systems is still an issue that needs to be resolved.

- **Short duration event occurrence:** Short duration events like gunshots pose real problem due to insufficient length of an event required for analysis (Crocco et al. 2016).

- **No specific feature:** Presently different ASA tasks use conventional features that are normally proven to perform better in the case of Automatic Speech Recognition (ASR) and speaker recognition. However logically, these are not specific features for ASA. We need a thorough study of suitability of features for auditory scene analysis (Phan et al. 2017).

- **Inter-class similiarity:** Many times deciding an acoustic scene is confusing when multiple scenes contain similar sound events. For example, it is challenging to identify scenes such as a market and street traffic as these scenes carry very less discriminative information.

- **Lesser datasets:** The major dataset provider for ASA is the DCASE challenge organizer. Hence in the existing systems, the number of scenes and events is limited by the limitations of a dataset. Therefore, to build standard ASA systems for different tasks, better versatile datasets with more number of scenes and events of various real-world scenarios need to be collected.

1.5 SCOPE AND OBJECTIVES OF THE THESIS

In this research, different problems in auditory scene analysis (ASA) have been explored. ASA is a field of scene analysis which includes developing automated systems to detect and classify environmental sounds, classify auditory scenes, bird sound classification, audio tagging, sound source localization, domestic sound detection and classification, and so on. The scope of this research is to explore three ASA tasks, namely, polyphonic sound event detection, Sound Event Localization and Detection, and Acoustic Scene Classification. We aim to develop systems that will enhance the performance of the detection and classification of sound events and acoustic scenes.

The key objectives of this research work are stated below:

1. To perform comprehensive analysis of the literature on the methods of polyphonic SED, SSL and ASC tasks.

2. To identify key challenges occurring in the auditory scene analysis tasks.

3. To perform in depth analytical study and exploratory of different feature extraction techniques.

4. To propose different machine learning and deep learning approaches for performing various ASA tasks.

1.6 BRIEF OVERVIEW OF THESIS CONTRIBUTIONS

The key contributions of this thesis include enhancing the performance of three related ASA tasks: polyphonic Sound Event Detection (SED), Sound Source Localization (SSL) of overlapped acoustic events, and ASC in the case of mismatched recording devices.

Characterization and detection of different overlapped acoustic events

The most common features for detecting overlapped acoustic events are Time-Frequency Representations (TFRs) such as spectrograms. Visual information

from the spectrogram may be promising features for SED. However, spectrogram images fail to capture the short-duration sound events and also work better for isolated sound events as compared to overlapped sound events. To perform polyphonic SED, a Mel-pseudo-based Constant Q-transform is proposed. CQT is a signal transformation technique that has shown better results for western music signals (Fitzgerald et al. 2006). Therefore, an improvised CQT technique is used to identify sound events. These features are fed as input to hybrid neural networks. Also, a deep learning approach that captures temporal information from the given audio recording is proposed. Experiments show that the proposed features outperform traditional speech features, and the classifier outperforms traditional machine learning and deep learning classifiers (Spoorthy and Koolagudi 2023b).

Sound Source Localization of overlapped acoustic events

Sound Event Localization and Detection (SELD) is the spatial and temporal localization task of various sound events and their classification. Commonly, multitask models are used to perform SELD. In this research work, a channel-wise 'FusionNet' deep learning network is designed to perform the SELD task. The proposed model performs the tasks of SED and DOA estimation in one neural network model. A novel fusion layer is introduced in the conventional artificial neural network, where the input is fed channel-wise, and the outputs of all channels are fused to form a new feature representation. The proposed deep neural network utilizes separable convolution blocks in the convolution layers, making the network less complex in terms of resource utilization. In the existing models proposed for the SELD task, the channel-wise information in the input feature is overlooked. However, based on the experiments, it is observed that the channel-wise information in the input feature carries discriminative information, and it also provided improved results compared to the state-of-the-art SELD systems.

ASC in the case of mismatched recording devices

From the past two years, the datasets that are released for ASC by DCASE and other researchers consist of audio samples recorded with multiple devices bringing the problem closer to the real-world scenarios. In this research work, a device-robust ASC system is developed to normalize the distortions occured due to different recording

devices (known as device noise) especially of lower quality one. Here, a two-level ASC system is proposed which performs the classification of acoustic scenes at two sequential levels: At the broader or outer level, the acoustic scenes are classified into three classes, namely, indoor, outdoor and transportation; in the finer or inner level, the audio recordings are further classified into classes such as airport, bus, traffic street, and so on. The proposed model retrains the audio samples that are misclassified in broader level to achieve better finer level classification.

A different deep learning approach named Deep Fisher Network is also proposed to perform ASC. This method combines the working principles of traditional machine learning algorithms and deep learning algorithms. The results achieved from the proposed systems outperform the state-of-the-art ASC systems.

1.7 ORGANIZATION OF THE THESIS

The outcomes of the research work carried out are elaborated in 6 chapters. An outline of each chapter is given below.

- **Chapter 1 : The Introduction** covers introduction of various tasks under ASA. Polyphonic acoustic events, and their localization in an acoustic scene are briefly discussed. Motivation, applications, challenges during recognition of acoustic events and scenes are briefly mentioned. Chapter ends with the clearly articulated research contributions and thesis outline.

- **Chapter 2 : Literature Review** critically analyses the existing works, their scopes and limitations on the issues such as mainly available datasets, data augmentation approaches, features and their extraction, and relevance of classifiers employed during sound event detection, sound source localization and acoustic scene classification. After critical review, research gaps are identified, enumerated and aptly discussed. The common datasets used in this research work are introduced. Scope of the present work derived from the literature review is presented.

- **Chapter 3 : Characterization and detection of polyphonic acoustic events** elaborates different approaches of detecting different monophonic and polyphonic sound events present in given audio recordings. The feature extraction method named "Mel-Pseudo Constant Q-Transform" and deep learning methods named "Modified Recurrent Temporal Pyramid Neural Network" proposed to perform polyphonic SED are discussed. The performance of the proposed approaches is presented, analysed and discussed along with the necessary comparison with the state-of-the art approach.

- **Chapter 4 : Sound source localization and detection of acoustic events** covers a novel deep learning approach to perform SED as well as Direction-of-Arrival (DOA) estimation of different sound events. The features used to obtain event and its origin information are spectral features. A new deep learning model named "Channelwise FusionNet" is proposed to perform both SED and SSL tasks. The performance of the models is presented, analysed, and discussed along with appropriate comparison with state-of-the-art approach.

- **Chapter 5 : Acoustic scene classification for mismatched devices** discusses the new deep learning architectures proposed to perform scene classification for audio recordings that are collected from different recorders. The significance of each of the models is discussed. The features used to learn acoustic scene information are log Mel energies. Two new deep learning models, namely, "Deep Fisher Network" and "Bi-Level Lightweight ASC model" are proposed to perform the ASC task. To eliminate the device distortion present in different recording devices' audio recordings, an Adaptive Noise Reduction method is applied. The performance of the models is presented, analysed, and discussed along with appropriate comparison with state-of-the-art approach.

- **Chapter 6 : Conclusions and Future Scope** chapter is logically divided into three parts namely, summary, conclusion and future research directions. The sub-section summary highlights the important contributions of the work. Conclusion throws the light on the learning outcome of the complete research. We also have

highlighted the important additions to the existing knowledge base in the topics namely SED, SSL, and ASC. A well conducted serious and sincere research is always a beacon for future research. In this regard, we also tried to drag the attention of specific research groups to continue the work in future by listing some research pointers on the topic of this thesis.

CHAPTER 2

LITERATURE REVIEW

This chapter includes in detail the existing literature on the selected tasks of ASA. Sound event detection, localization of various sounds, and classification of different acoustic scenes are the major tasks in ASA. This chapter also includes sections on datasets used for different ASA tasks, signal preprocessing methods adopted, different data augmentation methods reported, feature extraction techniques, and classification models. A list of the major research gaps is provided which is obtained by reviewing the available literature at the end of the chapter.

2.1 DATASETS USED IN VARIOUS ASA TASKS

A suitable acoustic event/scene dataset is necessary for acoustic event/scene classification. The tasks of designing and collecting acoustic event datasets mainly depend on the research applications. For instance, a dataset with acoustic events such as gunshot, glass breaking and person screaming are mainly used in audio-based surveillance. Similarly, a dataset with acoustic events like applause, laugh, door knock, etc. are used to analyze the meeting room scenes. The survey presented in this section introduces the publicly available acoustic event and scene datasets for detection, localization and classification ASA tasks.

Table 2.1 contains popularly used polyphonic sound event datasets. Majority of them are Tampere University (TAU/TUT) sound event datasets, developed and released as parts of different editions of the "Detection and Classification of Acoustic

Scenes and Events" Workshop (DCASE) challenge. In this thesis, TUT Sound Events 2016 and 2017 datasets are used for the evaluation of proposed polyphonic SED approaches. TUT Sound Events 2016 dataset includes real-time acoustic events from home and residential areas (Mesaros et al. 2016b). TUT Sound Events 2017 dataset includes sound events recorded from a public street (Mesaros et al. 2017). These datasets are chosen for our experiments because the datasets contain sound events recorded from real-life scenarios. The solutions provided to the SED problem will help solve the problem of SED that is closer to real-world problems.

Most popular and commonly used datasets for localizing and detecting various sound events are listed in Table 2.2. This thesis considers TAU Spatial Sound Events 2020 dataset for evaluation of the proposed Sound Event Localization and Detection (SELD) system. The dataset contains multiple spatial sound-scene recordings, consisting of sound events of distinct categories integrated into a variety of acoustical spaces, and from multiple source directions and distances as seen from the different recording positions. The reason for choosing this dataset for the SELD task is that this is the only dataset on which both detection and localization tasks can be performed simultaneously.

Other widely used acoustic scene datasets are listed in Table 2.3. In this thesis, the task of classifying acoustic scenes recorded from mismatched recording devices (More than one device) is also addressed. For this task, we have used the TAU Urban Acoustic Scenes 2019, TAU Urban Acoustic Scenes Mobile, and TAU Urban Acoustic Scenes 2020 Mobile datasets released by DCASE challenge organizers. TAU Acoustic Scenes 2019 dataset consists of audios recorded from one microphone recorder (Mesaros et al. 2018). TAU Acoustic Scenes 2019 dataset consists of audios that are recorded from three recording devices: one independed microphone and two smartphones. The TAU Acoustic Scenes 2020 dataset consists of audios that are recorded from one microphone recorder, two smartphones, and one video-audio recorder (Heittola et al. 2020). The reason for choosing these datasets is that these datasets contain audio recordings that are recorded from multiple recording devices, which brings the ASC classification problem closer to the real-world ASC problem.

Table 2.1: Polyphonic sound event detection datasets

Sl.no	Dataset name	Year of recording	Details	Scope	Limitations	Ref.
1	Tampere University of Technology Sound events (TUT-SED) 2016 dataset	2016	Length in minutes-78, No. of audio samples-22 Recording scenes-Home & residential area, No. of events-20	The dataset consists of sound events recorded from home and residential areas. The dataset consists of both monophonic and polyphonic sound events.	Even though the dataset is recorded from real-life scenes, the number of audio samples is fewer.	Mesaros et al. (2016b)
2	Tampere University of Technology Sound events (TUT-SED) Synthetic 2016 dataset	2016	Length in minutes-566, No. of audio samples-18, No. of events-11	The dataset consists of synthetic mixture of different office sound events. Additional noises have been added to make the dataset more realistic.	The dataset is synthesized and differs from real-life scenario. The interclass-variability is less. The dataset's scope is limited to only one scene.	Lafay et al. (2017)
3	Tampere University of Technology Sound events (TUT-SED) 2017 dataset	2017	Length in minutes-70, No. of audio samples-24, Recording scene-Street, No. of events-6	This dataset contains acoustic events collected from real-life streets. The scope of the dataset is only limited to one acoustic scene-street.	The audio samples present are only 24 and the sound events present are only 6, which is very small.	Mesaros et al. (2017)
4	Joint sound event and scene dataset	2019	Length in minutes-1500, No. of acoustic scenes-10, No. of audio samples-3000, No. of events-32	This dataset consists of synthesized scenes created using real-world recordings of acoustic scenes from which different acoustic events are identified.	The dataset annotations are not validated. Therefore, this dataset cannot be used for performance evaluation.	Bear et al. (2019)

Table 2.2: Sound event localization and detection datasets

Sl.no	Dataset name	Year of recording	Details	Scope	Limitations	Ref.
1	Computational Hearing in Multisource Environments (CHiME) - Home dataset	2015	Length in minutes-408, No. of audio samples-1946, Recording scenes-Home, No. of events-7	The dataset consists of sound events and source information recorded in home scenarios.	The scope is limited to identifying the source origin of a sound event present in a scene and not detecting sound events.	Foster et al. (2015)
2	TAU Spatial Sound Events 2019 dataset	2019	Length in minutes-400, No. of audio samples-400, Recording scene-Academic building, No. of events-11	The dataset consists of sound events along with their source information, i.e., onset and offset of sound events and the azimuth and elevation angles for identifying the source of the sound events. Both event localization and detection tasks can be performed on this dataset.	The dataset recording has been performed inside an academic campus. Therefore, the number of sound events is limited as in comparison to real-world scenarios.	Adavanne et al. (2018a)
3	TAU-NIGENS Spatial Sound Events 2020 dataset	2020	Length in minutes-600, No. of audio samples-600, Recording scene-Academic building & common areas in TAU campus, No. of events-14	The dataset consists of audio samples of TAU Spatial Sound Events 2019 dataset. In addition to this, the dataset consists of audio recordings recorded from outdoor environment. To make the dataset more robust, additional ambient noise is added to the audio samples.	Similar to TAU Spatial Sound Events 2019 dataset, this dataset is also recorded in academic environment consisting of commonly accessible student areas such as cafeteria, library, and so on. However, real-world scenarios are not restricted to these sound events. Therefore, the number of sound events is very limited.	Politis et al. (2020a)

Table 2.3: Acoustic scene classification datasets

Sl.no	Dataset name	Year of recording	Details	Scope	Limitations	Ref.
1	LITIS-Rouen University Dataset	2015	Length in minutes-3026, No. of scenes-19,	This dataset consists of scenes such as busy street, bus, cafe, car, train station, kid game hall, student hall, restaurant, pedestrian street, shop, train, etc. Each audio is around 30-seconds length sampled at 22050 Hz.	The dataset contains audio recordings from different real-world scenes. However, the size of the dataset with respect to each scene is very limited.	Rakotomamonjy and Gasso (2015)
2	TAU Acoustic Scenes 2017 development dataset	2017	Length in minutes-52, No. of scenes-15, No. of audio samples-312, Recording device-Soundman OKM II Klassik/studio A3 microphone	The scenes present in this dataset are bus, city, forest, etc. The scenes were broadly categorized into indoor, outdoor, and transportation classes.	The dataset consists of different real-world recordings. However, the size of the dataset is limited with respect to each scene. It is difficult to generalize the performance of an ASC system with this limited sized dataset.	Adavanne et al. (2018a)
3	TAU Urban Acoustic Scenes 2018 development dataset	2018	Length in minutes-8640, No. of scenes-10, Number of audio samples-864, Recording device-Soundman OKM II Klassik/studio A3 microphone	The scenes present in the dataset are Airport, bus, pedestrian street, etc. The dataset consists of larger number of audio samples compared to previously released TAU 2017 dataset.	The dataset consists of recordings from multiple scenarios. However, the dataset consists of recordings only from one recording device. The system developed on this dataset may not show good results when input is given from the other recording device.	Mesaros et al. (2018)

Table 2.3: Acoustic scene classification datasets

Sl.no	Dataset name	Year of recording	Details	Scope	Limitations	Ref.
4	TAU Urban Acoustic Scenes 2018 Mobile development dataset	2018	Length in minutes-10080, No. of scenes-10, No. of audio samples-1008, Recording devices-Soundman OKM II Klassik/studio A3 microphone, Samsung Galaxy S7, and IPhone SE	This dataset consists of scenes similar to TUT urban acoustic scenes 2018 dataset. Along with that, additional two recording devices' audio samples are added in this dataset.	Even though, the audio samples are additionally recorded from different audio recorders, there is a huge imbalance in the number of recordings of the two devices. This can impact the performance of the ASC system.	Mesaros et al. (2018)
5	TAU Urban Acoustic Scenes 2019 dataset	2019	Length in minutes-14400, No. of scenes-10, No. of audio samples-1440, Recording device-Soundman OKM II Klassik/studio A3 microphone	The scenes present in this dataset are Airport, bus, pedestrian street, etc. The number of samples for each scene is increased in comparison to TAU 2018 dataset for main recording device.	The dataset consists of recordings from multiple scenarios. However, the dataset consists of recordings only from one recording device. The system developed on this dataset may not show good results when input is given from another recording device.	Mesaros et al. (2018)
6	TAU Urban Acoustic Scenes 2019 Mobile dataset	2019	Length in minutes-16560, No. of scenes-10, No. of audio samples-1656, Recording devices-Soundman OKM II Klassik/studio A3 microphone, Samsung Galaxy S7, and IPhone SE	The scenes present in the dataset are similar to that of TAU Urban Acoustic Scenes 2019 dataset. The dataset consists of audio samples recorded from different recording devices.	Although other audio recorders were also used to record the audio samples, there is a major difference between the number of recordings recorded with each device. The performance of the ASC system may be affected by this.	Mesaros et al. (2018)

Table 2.3: Acoustic scene classification datasets

Sl.no	Dataset name	Year of recording	Details	Scope	Limitations	Ref.
7	TAU Urban Acoustic Scenes 2020 Mobile dataset	2020	Length in minutes-23040, No. of scenes-10, No. of audio samples-23040, Recording devices-Soundman OKM II Klassik/studio A3 microphone, Samsung Galaxy S7, IPhone SE, and GoPro Hero5 Session, Simulated devices-6	The scenes present in this dataset are similar to TAU Urban Acoustic Scenes 2019 dataset. The dataset consists of recordings from multiple devices which are both real and simulated (synthetically generated) devices.	The dataset consists of recordings from multiple devices. However, the number of scenes is similar to previous datasets. To standardize an ASC system, more scenes are necessary.	Heittola et al. (2020)
8	TAU Urban Acoustic Scenes 2020 Three Class dataset	2020	Length in minutes-14400, No. of scenes-3, No. of audio samples-1440, Recording device-Soundman OKM II Klassik/studio A3 microphone	The dataset consists of 3 classes broadly divided into indoor, outdoor, and transportation scenes.	The dataset is limited to only 3-class classification. However, in real-world scenarios, more number of scenes need to be considered for more reliable ASC system.	Heittola et al. (2020)

2.2 POLYPHONIC SOUND EVENT DETECTION: A REVIEW

Polyphonic Sound Event Detection (SED) systems deal with the task of identifying overlapped multiple acoustic events in a continuous audio signal. Polyphonic SED is a machine listening problem that can be broadly implemented through two stages: The feature extraction or feature representation stage and the detection or classification stage. To extract meaningful distinguishable information about the sound events from the input signal is a the feature extraction step. Once the feature extraction step is completed, these features are fed as input to the classifier to detect the onset and offsets for labeling that event. An in-depth analysis of different features and classifiers used in the literature to perform polyphonic SED task is given in the subsections to follow.

2.2.1 Features

Different features used for extracting event information from the audio signal are listed in Table 2.4. The most common features used for capturing event information from the audio recordings were Mel-Frequency Cepstral Coefficients (MFCC) and their velocity and acceleration variants (Rajapakse and Wyse 2005), Mel spectrograms (Mesaros et al. 2015), Time-Difference of Arrival (TDOA) (Adavanne et al. 2017), and Mel spectra (Bisot et al. 2017a). However, log Mel band energies were discovered to be more useful in polyphonic SED systems developed in the last three years (Xia et al. 2018),(Vesperini et al. 2019).

2.2.2 Classifiers

The classifiers used to obtain the event onset and offset along with event type are given in Table 2.4. In the earlier SED models, Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) were popularly used for polyponic SED task (Rajapakse and Wyse 2005). The GMM-HMM based models learn and model the individual sound event using Gaussian mixtures and estimate the onset and offset times of the event using HMM's states' Viterbi algorithm. Non-Negative Matrix Factorization (NMF) based source separation has also exhibited better performance in SED (Mesaros et al. 2015), (Bisot et al. 2017a).

Research activities of Polyphonic SED have recently adopted the advanced deep learning algorithms because of significant improvement in the performance and their capability to extract high-level shift invariant features along with the long temporal information from the audio recordings. Some of the machine learning and deep learning based methods used for Polyphonic SED are Feed Forward Neural Networks (FFNN) (Cakir et al. 2015), Convolutional Neural Networks (CNN) (Xia et al. 2018), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) (Adavanne et al. 2017), Convolutional Recurrent Neural Network (CRNN) (Chatterjee et al. 2020), Transposed Convolutional Recurrent Neural Network (TCRNN) (Chatterjee et al. 2020), and so on.

Table 2.4: Summary of important features and classifiers used for polyphonic sound event detection

Sl.no	Title	Approaches	Limitations	Remarks
1	Generic Audio Classification Using a Hybrid Model Based on GMMs and HMMs (Rajapakse and Wyse 2005)	Dataset: TUT-SED 2009 Features: MFCCs and their velocity coefficients (Dimension-39), Classifier: GMM-HMM (3 HMM states with 16 Gaussian mixture components for each class)	The method has provided low accuracy with large error rate and also at a given time, the method can detect only one prominent event.	The GMM-HMM model performed well for sound classes consisting of more number of audio samples. However, the hybrid GMM-HMM model could not classify all the sound events.
2	Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations (Mesaros et al. 2015)	Dataset: TUT-SED 2009, Features: Spectrum (Window length-100 ms, overlap-50%, FFT-1024), Classifier: Non-negative Matrix Factorization (NMF) (Non-negative dictionaries are obtained from training data)	This method requires careful tuning of threshold value which is a time-consuming task.	Method for identifying overlapping events in large databases without building unique models for every class is presented in this work. An annotation matrix is generated with overlapping sound events and has the benefit of being able to train directly on complex audio with annotated overlapping sound events.
3	Overlapping sound event detection with supervised nonnegative matrix factorization (Bisot et al. 2017a)	Dataset: TUT-SED 2016, Features: Mel Spectrum (No. of bands-40, Window length-40 ms, overlap-50%), Classifier: Non-negative Matrix Factorization (NMF) using Logistic Regression (Non-negative dictionaries and projections are obtained from training data)	The method does not perform well for overlapped sound events.	In this method, both supervised and unsupervised NMF methods are investigated for polyphonic SED. The method showed improved performance for deep learning based NMF model.

Table 2.4: Summary of important features and classifiers used for polyphonic sound event detection

Sl.no	Title	Approaches	Limitations	Remarks
4	Polyphonic sound event detection using multi label deep neural networks (Cakir et al. 2015)	Dataset: TUT-SED 2009, Features: Log-Mel Energies (Window length-40 ms, overlap-50%), Classifier: Multi-label DNN (Hidden layers-2, No. of units-800, Optimizer-Stochastic Gradient Descent)	The method requires large amounts of data for proper training for overlapped sound events.	The work proposes the use of multi-label DNNs for polyphonic sound event detection in realistic settings. It was found that multi-label DNN classification with post-processing based on median filtering could detect overlapping sound occurrences with a high degree of accuracy.
5	Sound event detection in multichannel audio using spatial and harmonic features (Adavanne et al. 2017)	Dataset: TUT-SED 2016, Features: Log-Mel Energies (Window length-40 ms, overlap-50%), Pitch (For each of the channels, top three pitch values, and its respective periodicity values are extracted at every frame in 100- 4000 Hz frequency range), Time difference of arrival (TDOA)-Median of multi-window TDOA's extracted from stereo audio, Classifier: Recurrent Neural Network (RNN) (LSTM layer)	The pitch and TDOA features provide very little event information. Thus, no much improvement has been observed by using these features.	The paper proposes the use of RNN-LSTM networks when combined with spatial and harmonic features for multi-label sound event detection. It was observed that the multi-channel features outperformed the baseline system with mono-channel features in a substantial manner.

Table 2.4: Summary of important features and classifiers used for polyphonic sound event detection

Sl.no	Title	Approaches	Limitations	Remarks
6	Polyphonic sound event detection by using capsule neural networks (Vesperini et al. 2019)	Dataset: TUT-SED 2016 and 2017, Features: Log-Mel Energies (No. of Mel bands-40, Window length-40 ms, overlap-50%), Classifier: CapsuleNet(CapsNet) (Maximum number of routing iterations-100, Optimizer-AdaDelta, Epochs-100)	The training of CapsNet requires extensive computational resources. Also, F-score is not reported in this work, which is an important metric for SED task.	This work introduced the adaptation of CapsNet for sound event detection task. The method works better for limited data as well. However, the time and space complexity of the method is high as compared to traditional CNN models.
7	A deep neural network-driven feature learning method for polyphonic acoustic event detection from real-life recordings (Mulimani et al. 2020)	Dataset: TUT-SED 2016, Features: Log-Mel energies (No. of Mel bands-60, Window length-25 ms, overlap length-10 ms), Classifier: DNN-driven feature learning model (2 projection layers to learn features, CNN layer-processes a small local region of the input using the group of filters)	The combination of CNN layer and projection layers in the work resulted in improved performance. However, there is no proper justification of working of projection layer for performance improvement.	In this work, a Deep Neural Network (DNN)-driven feature learning method is proposed for polyphonic Acoustic Event Detection (AED). The proposed method learns from spectral features and extracts high-level features for polyphonic AED. The proposed DNN model is invariant to the size of the datasets.
8	Polyphonic sound event detection using transposed convolutional recurrent neural network (Chatterjee et al. 2020)	Dataset: TUT-SED 2016, Features: Mel-IFgram (Instantaneous Frequeny), Classifier: Transposed Convolutional Recurrent Neural Network (TCRNN) (3 2D transposed convolution layers, 2 Bidirectional Gated Recurrent Unit (GRU))	The reported results have improved F1 score, however, the Error rate obtained from this method is high.	A deep learning neural network is developed using the combination of transposed convolution and GRU layers. The use of Mel-IFgrams in this work has provided improved performance as compared to conventional Mel spectrogram features.

2.3 SOUND SOURCE LOCALIZATION: A REVIEW

Sound Source Localization (SSL) is the task of determining the position (location) or direction of the sound source with respect to the microphone. In the human auditory system, the localization of sound sources is based on auditory cues, mainly the significant characteristics of incoming audio signals aid in localizing sound sources. Generally, there are two channels in the recorded signals from which auditory cues may be extracted: binaural and monoaural signals. Monoaural cues are the variations in the audio signals that use only one ear. The distance between the sound source and one of the ears, the frequency contents, the impact of echoes and reverberations on the incoming audio signal, etc., are the important information available with monoaural representation of the audio. Binaural Cues, also referred to as inter-aural differences, represent the variations in the audio signals received by the two ears or audio sensors. Aural difference is an important piece of information in SELD. Some of the widely used extraction methods and classifiers used for source localization and detection of sound events are discussed in detail in the subsections to follow.

2.3.1 Features

Different features used to perform the SELD task are given in Table 2.5. To localize the sound events present in a surrounding, one of the most common features used is acoustic intensity vectors (Cao et al. 2019). They give the signal's net acoustic energy flux and are computed for each of the Mel-bands in the spectrogram obtained from the signal. The sound event detection from the audio signal is performed using phase and magnitude spectrogram (Adavanne et al. 2018a), (Kapka and Lewandowski 2019). The combination of these two features is used to perform both localization and detection of sound events. Either monaural or binaural channel is used to compute log Mel band energies. Some of the binaural features are dominant frequency feature and Time Difference of Arrival (TDOA) features (Adavanne et al. 2017). Based on the position and orientation of microphones (the sound source) in a binaural environment, TDOA features are calculated.

2.3.2 Classifiers

To develop a system performing the localization and detection tasks, a multitask learning model is needed. In the literature, deep learning models are used for localization and the detection task. Some widely used classifiers are given in Table 2.5. CNN and RNN architectures are combined to form CRNN model (Adavanne et al. 2018a), which performed better than the systems developed using individual CNN and RNN architectures. The CRNN architecture consolidates the properties of CNN by extracting higher-level shift-invariant features and RNN learns long term temporal information of the audio recordings. CRNN may be considered as the state-of-the-art deep learning model for sound source localization and detection, as the approach has reported the best performance. Till now, use of CRNN architecture has been used in numerous existing SELD systems (Politis et al. 2020b). However, CRNNs require a large dataset for training; if the dataset is not sufficiently large, then this model may encounter problem of overfitting.

Table 2.5: Summary of important features and classifiers used for sound event localization and detection

Sl.no	Title	Approaches	Limitations	Remarks
1	Sound event localization and detection of overlapping sources using convolutional recurrent neural networks (Adavanne et al. 2018a)	Dataset: TAU Spatial Sound Events 2019, Features: Phase and Magnitude spectrogram (No. of FFT-1024, overlap-50%), Classifier: Convolutional Recurrent Neural Network (CRNN)(Convolution layers, ReLU activation function, Bidirectional GRU)	The method learns both event source information and the event type information in the same neural network model. The resultant error rate and DOA error are higher.	This paper proposes a CRNN to track sound events in relation to time and concurrently identify and localize them. Phase and magnitude spectrograms are used as the input feature by the network. By utilizing this non-method-specific features, the method becomes generic and may be effortlessly extended to various array structures.
2	Sound source detection, localization and classification using consecutive ensemble of CRNN models (Kapka and Lewandowski 2019)	Dataset: TAU Spatial Sound Events 2019, Features: Magnitude and Phase spectrograms (Window size-40 ms, overlap-50%, No. of FFT-1024), Classifier: 4 CRNNs	The method worked better for one task at a time and the performance degradation was observed if SSL and SED are performed simultaneously.	In this work, there was minimal feature engineering and the pure magnitude and phase spectrograms of the First Order Ambisonics format were used as input to the CRNN model.
3	Polyphonic sound event detection and localization using a two-stage strategy (Cao et al. 2019)	Dataset: TAU Spatial Sound Events 2019, Features: Log mel spectrograms & Generalized cross-correlation Phase Transform (GCC-PHAT) (No. of FFT-1024, Window length-20 ms, overlap-50%, No. of Mel bands and delays in GCC-PHAT-60), Classifier: Ensemble CRNN (Stage 1-SED branch, Stage 2-DOA Estimation)	Even though a two-stage network was introduced to perform the SED and DOA estimation tasks separately, the ensemble model is computationally complex.	In this work, a transfer learning approach is investigated, where the first stage performs SED and the information of this stage is transferred to the second stage. The method displayed improved performance as compared to baseline system.

Table 2.5: Summary of important features and classifiers used for sound event localization and detection

Sl.no	Title	Approaches	Limitations	Remarks
4	Sound event detection and direction of arrival estimation using residual net and recurrent neural networks (Ranjan et al. 2019)	Dataset: TAU Spatial Sound Events 2019, Features: Phase and Magnitude spectrogram (Window size-40 ms, overlap-50%), Classifier: ResNet RNN (2D convolution layer with 64 filters, ReLU activation function, 4 identity blocks with 3 filters, GRU layer).	The ResNet RNN method does not perform well when SED and DOA estimation tasks are jointly performed.	In this work, the classification and localization of sound events is achieved through the use of a 2-stage ResNet architecture when combined with RNN. Data augmentation and post-processing approaches greatly enhance the suggested model's performance, particularly for the DOA task the estimation of DOA is done with low error and 90% frame recall.
5	Event-independent network for polyphonic sound event localization and detection (Cao et al. 2021)	Dataset: TAU Spatial Sound Events 2020, Features: Log-Mel Spectrogram for SED (Window length-5 seconds, overlap-80%), Intensity vector for DOA estimation (Acoustical energy direction of a s ound wave, computed for x,y,z directions), Classifier: Event-independent network (1D convolutional layers, Batch normalization, average pooling)	The results exhibited improved F1 score. However, the network performs three tasks in parallel, making the network more computationally complex.	This paper proposes an end-to-end event-independent network for polyphonic sound event detection and localization. Polyphonic events are handled by the network as multi-track problems, where each track consists of a maximum of one event and its related DOA. Event activity detection encompasses the feature embedding information from both SED and DOA, hence is able to predict on-set times of events more accurately.

Table 2.5: Summary of important features and classifiers used for sound event localization and detection

Sl.no	Title	Approaches	Limitations	Remarks
6	Audio event detection and localization with multitask regression network (Phan et al. 2020)	Dataset: TAU Spatial Sound Events 2020, Features: Log-Mel magnitude spectrogram (Window size- 40 ms, overlap-50%, No. of Mel bands-64), Intensity vectors, Classifier: Multi-regression network (6 convolution layers, 5 max pooling layers, ReLU activation function, Bidirectional RNN)	The method has an Error rate of 0.60 which is very high as the good performing SELD system will have an error value near to or equal to 0.	This paper proposes a joint modeling technique to solve the sound event detection and localization task as a multitask regression issue, allowing the Mean Square Error loss to be applied uniformly for both subtasks. The proposed network has a self-attention mechanism and CRNN design, which is popular for sound event detection.
7	The USTC-iFlytek system for sound event localization and detection of DCASE 2020 challenge (Wang et al. 2020)	Dataset: TAU Spatial Sound Events 2020, Features: Log-Mel spectrogram (Window size- 40 ms, overlap-50%, No. of Mel bands-64), Intensity vectors (x,y,z directions), Classifier: Ensemble of deep learning models (CNN, RNN, ResNet, Xception, factorized time delay neural network (TDNNF))	To achieve high-level feature representations, different deep learning architectures are used. In this work, deep learning models are used for both feature learning process, and SELD task, making the SELD system complex in terms of both time and space.	This work proposes an SELD method where ensemble of several deep learning models are investigated to achieve improved performance. Additional improvements in DOA and SED estimation are obtained by post-processing and using model ensemble technique.

2.4 ACOUSTIC SCENE CLASSIFICATION: A REVIEW

Acoustic Scene Classification (ASC) is referred to as the task of identifying the environment in which the scene has been recorded. This section includes the tasks of characterizing and automatically detecting the acoustic scenes using approaches such as signal processing, feature extraction, and machine learning (McAdams and Bigand 1993). In this section, the feature extraction and classification methods that are commonly used for classifying acoustic scenes are discussed.

2.4.1 Features

Some of the important features that are used for ASC are listed in Table 2.6. In the literature, the most common set of features are spectrogram features. In the early ASC systems, the Mel-Frequency Cepstral Coefficients (MFCC) are commonly used. Even after rigorous research, there are no specific features that have been proposed to identify events and scenes. The feature representations used for scene processing are most commonly used in the other speech tasks such as ASR, speaker recognition, and so on. Recently, Image-based features are being extracted from Mel-spectrograms. The most common features extracted are Local Binary Patterns (LBP), Histogram of Oriented Gradients (HoG), and Adjacent Evaluation Completed LBP (AE-CLBP) (Abidin et al. 2017; Yang and Krishnan 2017). In recent trends, there has been an increased usage of deep learning models for feature extraction process as well, rather than only classification or prediction purposes. Single dimensional feature maps extracted from SoundNet (Singh et al. 2018), Deep features extracted using DNN (Jung et al. 2020), deep audio embeddings (Xie et al. 2022), and so on are some of the feature extraction initiatives through DNNs. Deep learning models tend to learn high-level features hierarchically from the input fed to the network i.e., more complex features are learnt from the deeper layers.

2.4.2 Classifiers

Some of the widely used classifiers used for ASC task are given in Table 2.6. The classifiers used to develop ASC systems can be broadly categorized into two types, namely, traditional or conventional and deep learning classifiers. Traditional classifiers

can be further categorized into generative and discriminative classifiers. In the traditional classifiers, we have the machine learning methods such as Support Vector Machines (SVM) (Ye et al. 2015), Non-negative Matrix Factorization (NMF) (Bisot et al. 2017b), Random Forest (RF) (Pham et al. 2021), GMM-HMM-based classifiers (Vuegen et al. 2013), Adaboost (Fonseca et al. 2018), Linear Discriminant Analysis (LDA) (Fonseca et al. 2018), k-means clustering algorithm (Salamon and Bello 2015), Multilayer Perceptron (MLP), and so on. ASC systems have also used Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) to learn the discriminative information from the various audio scene recordings (Vuegen et al. 2013). These methods maybe called traditional methods because these classifiers require hand-crafted features to learn the discriminant information to classify various acoustic scenes. Generally, the performance of the traditional classifiers depletes in the case of larger datasets. These limitations of the traditional methods are overcome by deep learning models. The deep neural network models learn the features from the input data, and the final decision is taken. The models work well with larger datasets as well. Some of the popular and common deep learning models used to develop an ASC system are CNN (Han and Lee 2016), CRNN (Hu et al. 2021), Deep Neural Network (DNN) (Xu et al. 2016), Long-Short Term Memory (LSTM) (Leng et al. 2020), VGGNet (Leng et al. 2020), Gated Recurrent Units (GRUs) (Ren et al. 2018b), Restricted Boltzmann Machine (RBM), and so on. The combination of two deep learning architectures also is used to model as ASC system (Hu et al. 2021).

Table 2.6: Summary of important features and classifiers used for acoustic scene classification

Sl.no	Title	Approaches	Limitations	Remarks
1	Knowledge distillation in acoustic scene classification (Jung et al. 2020)	Dataset: TAU Urban Acoustic Scenes 2019, Features: Deep features using DNN (Dimension: 64x10), Classifier: Student-teacher DNN (Ensemble of multiple DNNs)	The method in this work assigns a soft label to the input that is trained using DNN. However, the predicted label is not verified with the annotation provided in the dataset. Therefore, the results obtained may not be reliable.	This work addresses two aims by adapting a knowledge distillation framework for the ASC task. One approach involved uses soft labels to represent the similarity of acoustic properties between various classes. The other goal was to distill different DNNs into a single DNN in order to create a single DNN that performs similarly to a score-level ensemble of numerous DNNs. The method improved class-wise performance of ASC task.
2	Deep mutual attention network for acoustic scene classification (Xie et al. 2022)	Dataset: TAU Urban Acoustic Scenes 2019 and TUT Acoustic Scenes 2017, Features: Mel-based spectrogram, Gammatone-based spectrogram, Bark-based spectrogram, and the CQT based spectrogram (Window size-40 ms, overlap-50%), Classifier: Deep mutual attention Network (DMA-Net)	The model works better for fusion of two or more time-frequency features. Therefore, the performance of the network is majorly dependent of multiple features.	In this method, the spectral features are fed to two CNNs and the intermediate features from the hidden layers are aggregated to get the final scene classification results. The method displayed improved performance for feature aggregation for different configurations of CNNs.
3	A two-stage approach to device-robust acoustic scene classification (Hu et al. 2021)	Dataset: TAU Urban Acoustic Scenes 2020 Mobile, Features: Log-mel energies along with their velocity and acceleration coefficients ((Window size-40 ms, overlap-50%), Classifier: ResNet, FCNN (fully convolutional neural network) and fsFCNN (frequency sub-sampling FCNN)	The method resulted in lower accuracy of 81.9% for three-class classification. If the 3-class classification performance is less then it affects the performance of 10-class classification.	In this work, a two-stage ASC is proposed. First stage is a broader-level scene classification and second stage is finer level scene classification. The method resulted in better performance with ensemble of deep residual networks.

Table 2.6: Summary of important features and classifiers used for acoustic scene classification

Sl.no	Title	Approaches	Limitations	Remarks
4	A Hybrid Approach to Acoustic Scene Classification Based on Universal Acoustic Models (Bai et al. 2019)	Dataset: TAU 2018 ASC, Features: MFCCs are used to model GMM-HMM to build acoustic segment model (ASM) (Window size-40 ms, overlap-50%, HMM-3 states, GMM-50 mixtures), Classifier: SVM, DNN (No. of hidden layers-3, No. of neurons-512, dropout rate-0.2)	The highest average accuracy obtained by the proposed method with hybrid DNN approach is 66.1%. For ASC task, this accuracy is less to generalize or scale this system for different devices' audio samples.	The method uses Hidden Markov Model to undermine the acoustic units in an acoustic scene. The method uses traditional feature learning approach and trained using SVM/DNN model.
5	Domain adaptation neural network for acoustic scene classification in mismatched conditions (Wang et al. 2019b)	Dataset: TAU Urban Acoustic Scenes 2019 Mobile, Features: Log-Mel energies (Window size-64 ms, Overlap-15 ms, No. of Mel banks-64/128), Classifier: Domain adaptation neural network (DANN) (4 convolution block, ReLu activation function, Dense layer with softmax activation function)	Even though the domain adaptation neural neural was able to mitigate the performance difference between mismatched recording devices, the overall accuracy of the ASC system is only 60.1%.	In this work, DANN is proposed which performs supervised domain adaptation to handle mismatched recording device data for ASC. DANN minimizes the classification error and meanwhile reduces the network capacity on discriminating the source data from the target data by adversarial training.
6	Feature alignment for robust acoustic scene classification across devices (Zhao et al. 2022)	Datasets: TAU Urban Acoustic Scenes 2019 and 2020 Mobile, Features: Log-Mel spectrogram and their acceleration and velocity coefficients (Window size-64 ms, overlap-25%, No. of Mel banks-64), Classifier-Two stream CNN (6 convolution block, Fully connected layer, softmax activation function)	The method uses a feature alignment technique to avoid misclassification in the case of mismatched recording devices. However, the overall accuracy obtained is only 66.5%.	In this work, a two-stream CNN is trained for two features, Log-Mel spectrogram and delta-deltas separately. A feature alignment technique is adopted to reduce the data-imbalance issue between mismatched recording devices.

2.5 RESEARCH GAPS

Some important research gaps, identified from the above review are listed below.

- The features currently being used for polyphonic sound event detection are well accepted ones in the domain of speech features. These features may not be suitable for acoustic event processing. Therefore, there is a need to identify more specific features that precisely characterize the acoustic events.

- Real-time sound events are normally overlapped themselves and with different background noises. Frame-based speech specific features are highly sensitive to noise. Therefore, there is a need for features that work well in noisy conditions.

- The neural network models require larger datasets for training. Therefore, there is a need for developing the models that work better for smaller datasets.

- The systems in the literature, proposed for sound source localization and detection exhibit comparatively poor performance in terms of error rate and F-score. There is a huge scope for improvement in the area of localization and detection of events.

- The ASC systems do not perform well when mismatched recordings recorded from different recording devices are fed to the system. Thus, there is a need to develop device robust ASC system.

2.6 PROBLEM STATEMENT

Based on the research gaps listed above, the following problem is formulated for the present research task:

Auditory scene analysis from audio data using deep learning approaches.

From the above problem statement, three research objectives are derived for the present research work:

2.6.1 Research Objectives

1. **Characterization and detection of different polyphonic events:**

   The task is to detect various acoustic events. The events can be monophonic or polyphonic. The first step is to identify the onset and offset times of the event and then detect the class to which the event belongs. This thesis explores different feature extraction methods such as Constant Q-Transform (CQT), Pseudo CQT, and Short-time Fourier Transform (STFT). Different deep learning models, such as Convolutional Recurrent Neural Network (CRNN) and Temporal Recurrent Neural Network, have been explored to detect sound events. Two polyphonic SED datasets, namely, TUT Sound Events 2016 (TUT-SED 2016) and TUT Sound Events 2017 (TUT-SED 2017), have been used to evaluate the proposed methods. The evaluation metrics considered are the Error rate and F1-score. The output of the SED system will be the onset and offset of a sound event in a given time frame and the type of sound event.

2. **Sound source localization and detection of acoustic events:**

   Identifying position or location of a sound event source is essential for certain critical applications. For example, a gunshot is heard in a surrounding. It is crucial to identify the sound event as a gunshot and the source of the sound. In this objective, we aim to develop a Sound Event Localization and Detection (SELD) system that identifies the event and provides the Direction-of-Arrival (DOA) estimation of the sound event. For this task, two features are used for the SED and DOA estimation tasks. For the SED task, log Mel energies are used, and for the DOA estimation task, intensity vectors are used. The features are fed as a multichannel input to deep learning models, namely, Channelwise FusionNet and different variants of SELDNet. To evaluate the performance of the proposed methods, Tampere University (TAU) Spatial Sound Events 2020 dataset is used. The metrics considered for performance evaluation are ER, F1-score, DOA error, and frame recall. The outcome of the SELD system is the sound event onset and offset, along with their sound type and the DOA estimation of the sound events.

3. **Acoustic scene classification for mismatched devices:**

   Acoustic Scene Classification (ASC) identifies a surrounding and assigns a semantic label to an audio recording. The existing systems proposed for ASC perform better if the audio recordings are recorded from one recorder, and the system's performance is depleted if audio recordings from other devices are fed to the ASC system. Therefore, in this thesis, we present device-robust ASC methods. The features used are Mel-Frequency Cepstral Coefficients and log Mel energies. Different classifiers, namely, Deep Fisher Network, Convolutional Attention Recurrent Neural Network, and a bi-level lightweight deep learning model, are proposed to perform device robust ASC task. Three datasets are used to evaluate the ASC systems' performance: TAU Urban Acoustic Scenes 2019, TAU Urban Acoustic Scenes Mobile 2019, and TAU Urban Acoustic Scenes 2020. The performance metric used to evaluate the ASC system is accuracy. The outcome of the ASC system is a scene label for a given audio recording.

2.7 COMMON RESOURCES USED IN THIS WORK

In this section, the details of datasets and the evaluation metrics that are considered for conducting experiments and evaluating performance are mentioned.

2.7.1 Resources for Polyphonic Sound Event Detection

Two datasets, namely, TUT-SED 2016 and TUT-SED 2017 are used for the evaluation of the proposed polyphonic SED systems. The details of datasets and performance metrics are given below:

A. TUT-SED 2016

The audio recordings present in the dataset are recorded from two real-life scenarios, namely, residential area and home (Mesaros et al. 2016b). The training material contains 16 event classes, provided as isolated sound events, 20 recordings per class. The details of the sound events present in each scene and the number of instances each sound event occurs in the recordings are given in Table 2.7. In order to get variability in the dataset, the recordings are captured at different locations, i.e.,

Table 2.7: Sound events present in Residential area and Home scenes of 'TUT-SED 2016' development dataset

Sl.No	Scenes	Events	No. of Instances
1	Residential Area	Objects banging	23
		Bird singing	271
		Car passing by	108
		Children shouting	31
		People speaking	52
		People walking	44
		wind blowing	30
2	Home	Object rustling	60
		Object snapping	57
		Cupboard	40
		Cutlery	76
		Dishes	151
		Drawer	51
		Glass jingling	36
		Object impact	250
		People walking	54
		Washing dishes	84
		Water tap running	47
3	Total no. of recordings		40
4	Duration of each recording		3-5 minutes long

from various homes and streets. This is done to avoid inter-class similarity amongst the audio recordings. The dataset consists of binaural audio recordings from each location for about 3-5 minutes long, summing up to a total of 78 minutes of audio. The annotations of these recordings have been performed manually. The residential area recordings consist of seven different annotated sound events and the home recordings consist of eleven different annotated sound events. A four-fold cross validation is performed to evaluate the proposed method, so that each audio recording is used at least once for testing of the model. The dataset split is given predefined by the DCASE challenge organizers.

B. TUT-SED 2017

The audio recordings present in the dataset are recorded from different streets (Mesaros et al. 2017). A 3-5 minute audio was recorded at each location. The details of the sound events present in the street scene and the number of instances each sound event occurs in the recordings are given in Table 2.8. The recording setup comprises a binaural Soundman OKM II Klassik/studio A3 electret in-ear microphone and a Roland Edirol R-09 wave recorder with a 44.1 kHz sampling rate and 24-bit resolution. TUT-SED 2017 dataset consists of sound classes such as brakes squeaking, car, children, large vehicle, people speaking, and people walking. These classes were recorded in the traffic areas. In order to get variability in the dataset and to avoid inter-class similarity, the recordings are captured from different locations, i.e., from various streets.

Table 2.8: Sound events present in Street scene of 'TUT-SED 2017' development dataset

Sl.No	Scene	Events	No. of Instances
1	Street	Brakes squeking	52
		Car	304
		Children	44
		Large vehicle	61
		People speaking	89
		People walking	109
2	Total no. of recordings		21
3	Duration of each recording		3-5 minutes long

C. Performance Metrics

An SED system is evaluated using two performance metrics, namely, F1 score and Error rate (ER). F1 score gives the accuracy of the model and ER gives the model's prediction error. F1 score is used as a primary evaluation metric which is calculated within a single time frame of one-second length. If the temporal position of an event in the system output overlaps with the temporal position of an identically labeled event in the ground truth, the event is said to have been successfully identified. The statistical values to represent false positive (FP), true positive (TP), and false negative (FN) need to be calculated for each event.

TP: correctly detected events

FP: events in the system output that are not correct according to the ground truth

FN: events in the ground truth that have not been correctly detected

S: events in system output that have correct temporal position but incorrect class label

D: events in ground truth that are neither correct nor substituted

I: events in system output that are neither correct nor substitutions

N: number of events in the ground truth

Equations 2.1 and 2.2 are used to compute precision (PRE) and recall (REC). The precision and recall computed are further used to calculate the F1 score (F1) as given in Equation 2.3 (Mesaros et al. 2016a).

$$PRE = \frac{TP}{TP + FP} \tag{2.1}$$

$$REC = \frac{TP}{TP + FN} \tag{2.2}$$

$$F1 = \frac{2.PRE \times REC}{PRE + REC} \tag{2.3}$$

The evaluation of the SED system is performed using a second metric ER. The statistics used for calculating ER are (1) the number of substitutions (S), (2) deletions (D), (3) insertions (I) and (4) number of events in the ground truth (N). The total ER is computed using statistics as shown in Equation 2.4 (Mesaros et al. 2016a).

$$ER = \frac{S + D + I}{N} \tag{2.4}$$

ER metric provides the misclassification rate of the onset and offset of sound events. The ER value ranges from 0 to 1; for an ideal SED system, the lesser the ER, the better the system performs. The F1 score metric shows how well the model detects the presence of an event, i.e., given a time frame, the sound events present in that time

frame need to be detected. The F1 score value ranges from 0 to 1; for an ideal SED system, the higher the F1 score, the better the SED system performs.

2.7.2 Resources used for Sound Source Localization of different overlapped acoustic events

The details of the dataset and the performance metrics considered for evaluation of SELD task are given below:

A. TAU-NIGENS Spatial Sound Events 2020

The proposed model for the SELD task is evaluated using the TAU-NIGENS (TAmpere University - Neural Information processing group GENeral Sounds) Spatial Sound Events 2020 dataset (Politis et al., 2020a). The audio recordings present in the dataset consist of different sound events that are recorded in multiple locations and also from different recording positions. The details of the sound events present in the dataset are given in Table 2.9. Two types of recording formats were used to record the dataset, namely, Microscopic array (MIC) and First-order Ambisonics (FOA). In this thesis, the experiments are performed on the FOA dataset. Each sound in the dataset is either stationary or a moving sound source in a room. Case Time-Variant Room Impulse Responses (RIRs) are used with the moving sound sources. A trajectory is associated for each sound event of its Direction-of-Arrival (DoA) to the recording point along with the event's temporal onset and offset. An Eigenmike spherical microphone array is used to obtain the real-life Impulse Response (IR) while creating a dataset that appears realistic. The Eigenmike was surrounded by a Genelec G Three loudspeaker for play-back. The dataset consisted of 14 sound events and 600 one-minute long audio recordings in the development dataset. The standard evaluation split used in the competition (DCASE 2020 SELD) is considered in this work. Among six dataset splits, 3, 4, 5, and 6 are used for training, Split 2 is used for validation, and split 1 (unseen data) is used for testing the performance of the proposed methods.

B. Performance Metrics

The evaluation of an SELD system is performed using two set of metrics, namely, location-aware detection and class-aware localization. The location-aware detection

Table 2.9: Sound events present in 'TAU-NIGENS Spatial Sound Events' development dataset

Sl.No	Scene	Events	No. of Instances
1	General	Alarm	49
		Baby crying	40
		Crash	50
		Dog barking	45
		Engine	39
		Female screaming	45
		Female speech	100
		Fire	51
		Footsteps	42
		Knocking	40
		Male scream	31
		Male speech	100
		Phone ringing	40
		Piano	42
2	Total no. of recordings		100
3	Duration of each recording		1 second - 5 minutes long

metrics are used for SED task in the SELD system. While computing the metrics, the location of a sound event is considered as well, therefore, these metrics are location-dependent. The metrics are Error rate (ER) and F1 score. The equations for computating of ER and F1 score are mentioned in the SED performance metrics (Sub-section 2.7.1).

The class-aware localization metrics are used for localization task in the SELD system, but are now classification-dependent, meaning that they are computed only across each class only, instead of across all outputs. The localization error (DOA Error) and localization frame recall are computed as given in Equations 2.5 and 2.7. The difference in angular degrees between the predicted and reference DOAs is known as the DOA error. In order to account for time frames where the number of estimated and reference DOAs are unequal, the frame recall metric is used.

$$FrameRecall = \frac{TP}{TP + FN} \tag{2.5}$$

The reference DOA estimates are given by x_G, y_G, z_G and the predicted DOA estimates are given by x_E, y_E, z_E, these are used to synthesize the dataset by utilizing

the central angle $\sigma\epsilon[0, 180]$ computed using Equation 2.6. D denotes the number of DOA ($d = 1...D$). At the origin in degrees, the σ formed by the reference and predicted DOAs and is given by Equation 2.7 (Adavanne et al. 2018b).

$$\sigma = 2.arcsin\left(\frac{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}{2}\right).\frac{180}{\pi} \tag{2.6}$$

$$DOAError = \frac{1}{D}.\sum_{d=1}^{D}\sigma\left(\left(x_G^d, y_G^d, z_G^d\right), \left(x_E^d, y_E^d, z_E^d\right)\right) \tag{2.7}$$

The DOA estimation metrics are essential for the SELD task to identify how well the SELD system estimates a sound source's x, y, and z coordinates. DOA error metric gives the misclassification rate of the sound events' source. The second DOA estimation metric, Frame Recall, gives the true positives. Therefore, the higher the recall, the lower the false positives/alarms. These two metrics are essential for DOA estimation.

2.7.3 Resources used for Acoustic Scene Classification

The performance of ASC systems has been evaluated using three ASC datasets. They are TAU Urban Acoustic Scenes 2019 development dataset, TAU Urban Acoustic Scenes 2019 Mobile development dataset, and TAU Urban Acoustic Scenes 2020 Mobile dataset. The details of these datasets and the performance metrics used for the evaluation of ASC system are given below:

A. TAU Urban Acoustic Scenes 2019 development dataset

The dataset consists of ten acoustic scenes, namely, Airport, bus, metro, metro station, public square, park, shopping mall, street traffic, street pedestrian, and tram (Mesaros et al. 2018). The dataset includes 14400 binaurally recorded audio recordings. Each class consists of 1440 audio recordings of length 10 seconds each. The audio segments are sampled at a rate of 48000 Hz and stored with 24-bit resolution. The device used for audio recordings is Soundman OKM II Klassik/studio A3, electret binaural microphone and a Zoom F8 audio recorder. The details of the acoustic scenes along with the number of recordings present in the dataset are given in Table 2.10.

Table 2.10: Acoustic scenes present in 'TAU-Urban Acoustic Scenes 2019' development dataset

Sl.No	Scene	No. of Recordings
1	Airport	1440
2	Indoor shopping mall	1440
3	Metro station	1440
4	Pedestrian street	1440
5	Public square	1440
6	Street with medium level of traffic	1440
7	Travelling by a tram	1440
8	Travelling by a bus	1440
9	Travelling by an underground metro	1440
10	Urban park	1440
11	Total No. of recordings	14400
12	Length of each audio file	10 seconds long

B. TAU Urban Acoustic Scenes 2019 Mobile development dataset

Mobile dataset consists of 40 hours of data of TAU Urban Acoustic Scenes 2019 development dataset along with the additional data recorded with two different mobile devices (Mesaros et al. 2018). The dataset consists of audio segments totaling up to 65 hours. Capturing discriminative information of events in this dataset is more challenging because of multiple recording devices and their characteristics. In this thesis, a total of 10265 audio segments of 10 different acoustic scenes were used for training the model and a total of 4185 audio segments were used for evaluation of the model. The audio segments are sampled at a rate of 48000 Hz and stored with 24-bit resolution. The main device used for audio recordings is Soundman OKM II Klassik/studio A3, electret binaural microphone and a Zoom F8 audio recorder. The other devices used for recording are Samsung Galaxy S7, IPhone SE, and GoPro Hero5 Session. The recordings from the different devices are time synchronized (recorded simultaneously). The details of the acoustic scenes along with the number of recordings for devices A, B, and C present in the dataset are given in Table 2.11.

C. TAU Urban Acoustic Scenes 2020 Mobile dataset

This dataset consists of audio recordings of ten acoustic scenes that are recorded from 12 European cities (Heittola et al. 2020). The recordings are collected from 4

Table 2.11: Acoustic scenes present in 'TAU-Urban Acoustic Scenes Mobile 2019' development dataset

Sl.No	Scene	No. of Recordings		
		Device A	Device B	Device C
1	Airport	1440	54	54
2	Indoor shopping mall	1440	54	54
3	Metro station	1440	54	54
4	Pedestrian street	1440	54	54
5	Public square	1440	54	54
6	Street with medium level of traffic	1440	54	54
7	Travelling by a tram	1440	54	54
8	Travelling by a bus	1440	54	54
9	Travelling by an underground metro	1440	54	54
10	Urban park	1440	54	54
11	Total	14400	540	540
12	Total No. of recordings	16550		
13	Length of each audio file	10 seconds long		

different devices. Device A is Soundman OKM II Klassik/studio A3, electret binaural microphone and a Zoom F8 audio recorder using 48kHz sampling rate and 24-bit resolution, device B is a Samsung Galaxy S7, device C is an iPhone SE and device D is a GoPro Hero5 Session. Addition to this, some audio recordings are generated synthetically for 11 mobile devices using the original audio recordings. The simulated audio recordings are generated using impulse responses recorded with real devices, and additional dynamic range compression. A recording from device A is processed through convolution with the selected S_i impulse response, then processed with a selected set of parameters for dynamic range compression (device-specific). In the development dataset, the recordings of only 10 cities are present and the remaining 2 cities are present only in the evaluation dataset. The development dataset consists of data from 9 devices: 3 real devices (A, B and C) and 6 simulated devices (S1-S6). The total amount of audio in the development set sums up to 64 hours of data. The training/test split of the dataset is of 70% and 30% respectively. The details of the acoustic scenes along with the number of recordings for devices A, B, C, and simulated devices S1-S6 present in the dataset are given in Table 2.12.

Table 2.12: Acoustic scenes present in 'TAU-Urban Acoustic Scenes 2020' development dataset

Sl.No	Scene	No. of Recordings			
		Device A	Device B	Device C	Devices S1-S6
1	Airport	1440	54	54	216
2	Indoor shopping mall	1440	54	54	216
3	Metro station	1440	54	54	216
4	Pedestrian street	1440	54	54	216
5	Public square	1440	54	54	216
6	Street with medium level of traffic	1440	54	54	216
7	Travelling by a tram	1440	54	54	216
8	Travelling by a bus	1440	54	54	216
9	Travelling by an underground metro	1440	54	54	216
10	Urban park	1440	54	54	216
11	Total	14400	540	540	2160
12	Total No. of recordings	23040			
13	Length of each audio file	10 seconds long			

D. Performance Metric

The metric that is considered for evaluation on an ASC system is Accuracy (ACC). The accuracy metric gives the correct predictions of the model out of the total samples present in the dataset. To compute the accuracy, TP, TN, FP, and FN are used. The equation of Accuracy is given in 2.8.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}$$

The use of accuracy in ASC is crucial. This metric is helpful in analyzing the performance of an ASC system. The accuracy measure helps assess the reliability of the ASC system.

2.8 SUMMARY

In this chapter, the review of existing systems for polyphonic SED, SSL and ASC tasks is presented in detail. The chapter summarizes the different feature extraction methods and classification methods that are used to perform various ASA tasks. From the literature review, the research gaps are drawn out and the problem statement for

our current research work is framed. The problem statement is further elaborated into three objectives. The common resources that are used for the experiments are given at the end of this chapter. In the next chapter, the methods proposed for first objective - characterization and detection of polyphonic acoustic events are discussed in detail.

CHAPTER 3

CHARACTERIZATION AND DETECTION OF POLYPHONIC ACOUSTIC EVENTS

This chapter includes two new methods to perform polyphonic Sound Event Detection (SED). The first method uses Mel-Pseudo Constant Q-Transform (MP-CQT) features and a Convolutional Recurrent Neural Network (CRNN) classifier. The second method uses a modified recurrent temporal pyramid neural network for classification of events based on CQT spectrograms. The dataset contains both monophonic (non-overlapped) and polyphonic (overlapped) events.

3.1 POLYPHONIC SOUND EVENT DETECTION USING MEL-PSEUDO CONSTANT Q-TRANSFORM AND DEEP NEURAL NETWORK

The primary purpose of polyphonic SED is to identify the onset and offset times of a sound event and assign a corresponding label to the event in an audio recording. This is performed in two steps, namely, sound representation and classification. The first step, sound representation involves extraction of frame-level features from the sound signal to obtain a feature vector $x_t \epsilon \mathbb{R}^F$, where $\mathbb{R}$ represents a real number, t represents the time frame and F represents the number of features per frame. This work uses a MP-CQT technique to extract frame-level features from the polyphonic audio signal. The second step is the classification of the sound events. In this step, probabilities $p(y_t(k)|x_t, \theta)$ are estimated for different event classes $k = 1, 2, ...K$ in a frame t, where θ represents the parameters of the classifier. The classifier parameters θ are trained by supervised

learning and the target outputs y_t for each frame are obtained from the onset/offset annotations of the sound event classes. Based on the probabilities obtained for the sound events, the binarization is applied using thresholding to achieve event activity predictions $\hat{y} \epsilon \mathbb{R}^k$. The thresholding is performed over a constant value (0.5). Based on the presence of an event in the particular time frame t, the value of that event is given either 0 or 1. The model that has been trained is used to identify an event present in real-life scenarios.

In this work, a Mel-Pseudo Constant Q-Transform (MP-CQT) based technique is proposed to perform polyphonic SED and effectively learn the sound events in the case of both small and large datasets. The block diagram of the proposed method is given in Figure 3.1. A pseudo CQT technique is adapted to extract features from the audio files using Mel-scale to imitate human perception of sound. The detection of the sound events is performed using a Deep Neural Network (DNN). The architecture of the DNN is a combination of convolutional and recurrent layers. The ensemble of Pseudo CQT and Mel-scale has produced a set of distinguished and discriminative features framewise for processing multiple overlapped events. The robustness of the system while processing noisy audio signals is also tested. The proposed method has significantly improved the performance over the state-of-the-art systems.

The reason for choosing frame-level features for performing the task of polyphonic SED is the instantaneous nature of some events. Polyphonic SED is a multi-label classification problem in which a single audio recording may contain multiple events. Sound events like glass breaking, gunshots, or door thuds are impulsive/sudden and end within shorter time. However, the events like rain or traffic occur for a longer time. Therefore, the methods used for classification must preserve the temporal contextual information from the features to get better discrimination. Thus, the input features in this work are chosen to be the frame-level features $X_{t:t+T-1}$, where T represents the number of frames in the audio, and the output targets are given by $Y_{t:t+T-1}$ for the frames t to $t+T-1$. The notations $X_{t:t+T-1}$ and $Y_{t:t+T-1}$ are denoted as X and Y, respectively for the sake of ease of reading through the rest of the chapter.

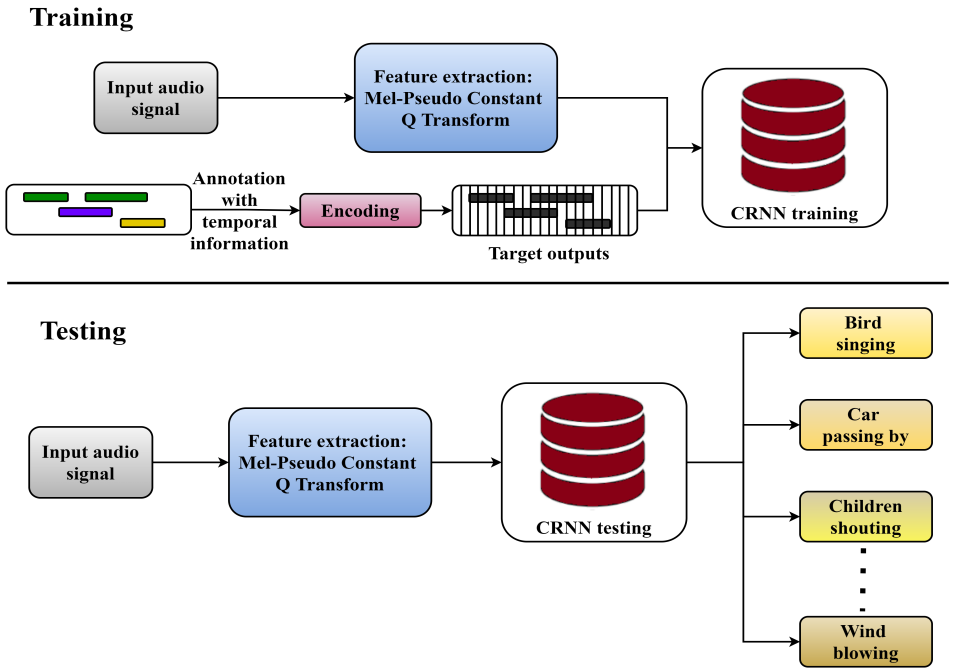

Figure 3.1: Block diagram of the proposed polyphonic SED method with Mel Pseudo CQT spectrograms as features and CRNN as classifier

3.1.1 Feature Extraction using Mel-Pseudo Constant Q-Transform

Information about the different sound events is captured using the Time Frequency Representations (TFRs) of the audio signals. In the literature, feature representations such as spectrogram or mel spectrogram are most commonly used TFRs to identify and classify sound events. The other popularly used TFR for SED task is Constant-Q Transform (CQT), in which the frequency axis is logarithmic. The advantage of using CQT for the SED task is that it offers better temporal resolution for high frequencies and better spectral resolution for lower frequencies (Fitzgerald et al. 2006). However, the pseudo (faster method of CQT computation) CQT performs better in the case of sparse audio signals and provides better information for different pitch values present in a signal. This motivated us to use pseudo CQT to perform the polyphonic SED task, as the sound events have sparse matrix representation and spread at different frequency bands.

CQT is closely related to Discrete Fourier Transform (DFT). This technique was introduced to map the scale of Western music. The key difference between DFT and CQT is that the former one uses frequency scaling with constant spacing, whereas, in

the case of the latter, the frequencies are geometrically distributed. A constant ratio Q is obtained between the central frequency of a band and the frequency resolution $f_t - f_{t-1}$. Hence, the name Constant-Q Transform (Fitzgerald et al. 2006). The equations are taken from the mentioned reference for CQT computation.

Given a minimum frequency of a signal as f_0 for the CQT, the center frequencies of each energy band for t number of filters can be calculated using Equation 3.1:

$$f_t = f_0 2^{(\frac{t}{b})} (t = 0, 1, ...) \tag{3.1}$$

where b is the number of bins per octave. The fixed ratio of the center frequency to bandwidth equation is given by Equation 3.2:

$$Q = \left(2^{\frac{1}{b}} - 1\right)^{-1} \tag{3.2}$$

The required bandwidth of each frequency band can be obtained by choosing the window length which is computed using Equation 3.3:

$$N_t = Q^{\frac{f_s}{f_t}} \tag{3.3}$$

where f_s is the sampling frequency of the audio signal. The CQT of an audio signal is computed using Equation 3.4:

$$X_t = \frac{1}{N_t} \sum_{n=0}^{N_t - 1} W_{N_t}(n) x(n) exp^{-j2\pi Q n / N_t} \tag{3.4}$$

where $x(n)$ represents the n^{th} sample of the audio signal in the time domain, W_{N_t} represents the window function of length N_t. In DFT domain, the transform is equivalently written as given in Equation 3.5:

$$X = MY \tag{3.5}$$

where M denotes the matrix that is a resultant of DFT on each column and Y denotes the DFT of vector x. The pseudo CQT is performed by applying l_1 normalization on the Y. L1 normalization is the normalization technique that modifies

the dataset values so that the sum of the absolute values in each row will always be up to 1. Once the CQT is computed, a Mel scale is applied on the coefficients. The Mel scale was developed to try and scale frequency data in a way which more closely resembles the way humans perceive sound. Therefore, the method is named as Mel-Pseudo CQT (MP-CQT). An illustration of the extraction of MP-CQT spectrogram is shown in Figure 3.2. Pseudo CQT for a window length of 10 milliseconds is presented in Figure 3.3.

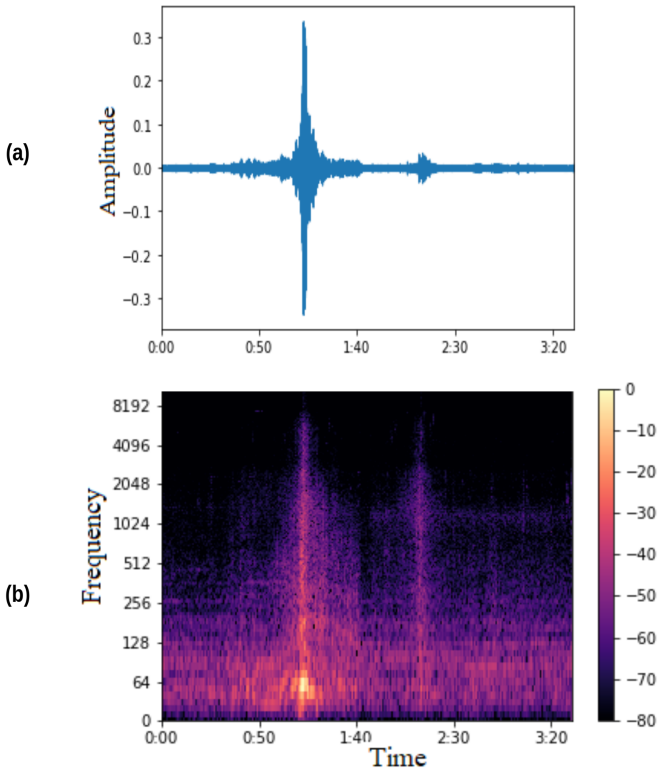

Figure 3.2: Illustration of Mel Pseudo CQT spectrogram generation (a) Input audio signal, (b) Spectrogram obtained from Pseudo CQT after applying Mel scale

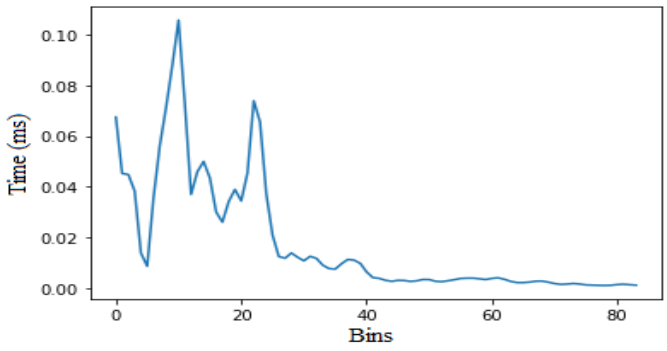

Figure 3.3: Pseudo CQT signal of a single window of 10 millliseconds

The spectrogram features extracted from MP-CQT are fed as input to the CRNN deep learning model. The model learns high-level [1] features, from which the sound event information is extracted. The final outcome of the model is the onset and offset of the event along with the event label.

3.1.2 Event Detection using Convolutional Recurrent Neural Network

The deep neural network architecture considered in this work is a combination of CNN and RNN, which is known as a CRNN model. This architecture learns the spatial information from the feature representation through the convolutional layers and the temporal information from the recurrent layers. The combination of the CNN and RNN harnesses the strengths of both architectures, eventually resulting in an improvement in the performance during the polyphonic Sound Event Detection (SED).

The network architecture contains a pipeline of layers: Initially, the pseudo CQT features are fed to the convolutional layer through two-dimensional convolutional filters. Once the features are obtained after passing them to an activation function, a non-overlapping max-pooling operation is applied to the features. The output of the previous layers is normalized using the batch normalization technique. In order to capture the temporal information in the bands, recurrent layers are used in the network. The recurrent layers need a one-dimensional input. Therefore, the output obtained from the last max-pooling layers is converted to one-dimensional input using permute and reshape layers. Here, the output of max-pooling layer is stacked with a frequency axis, and fed to the Gated Recurrent Unit layer (GRU). This layer takes two inputs: one is the output of the current frame of the previous layer, and the other is the output of the previous frame of the current layer. The last layer is the dense layer, which is a feed-forward layer with sigmoid as an activation function.

The optimal layer configuration is achieved by experimenting on different number of filters for the convolutional layers. The experiments were performed on 8, 16, 32, and 64 filter sizes. The arrangement of the filters in the max-pooling layer is initially

[1] High-level features are built on top of low-level features in the deep neural network, i.e, the initial layers in the network learns the information from the input fed to the network. The higher level features are extracted from the hidden layers in the neural network using the lower level features from the preceding layers.

5×5, followed by 4×4, and last max-pooling layer has 2×2 sized filter which will result in the reduction of MP-CQT features in three stages. Initial input is 40 bands and then transformed through the network.

In order to get the best performance, other hyperparameters such as activation function, dropout rate, number of hidden units in the dense layer, number of GRUs in the recurrent layers are emperically fine-tuned. The optimization function used for the CRNN architecture is Adam where the learning rate is set to 0.001 (Kingma and Ba 2015). The loss function used in the architecture is Binary Cross-entropy as this function tends to work better for multi-class classification problem (He et al. 2015). The threshold value is set to 0.5 for the event activity probabilities to achieve a binary activity matrix which is used for computing the reference metrics on the basis of the ground truths. The binary activity matrix is an output consisting of 0 and 1 values, where 0 indicates the absence and 1 indicates the presence of an event in a recording. The hyperparameters that provided the best performance are reported in this work. The layer configuration of the neural network is given in Figure 3.4.

The feature extraction performed in this work has been done using Python library Librosa (McFee et al. 2015) and the neural network architecture is designed using deep learning package Keras with Theano as backend (Chollet et al. 2018; Team et al. 2016).

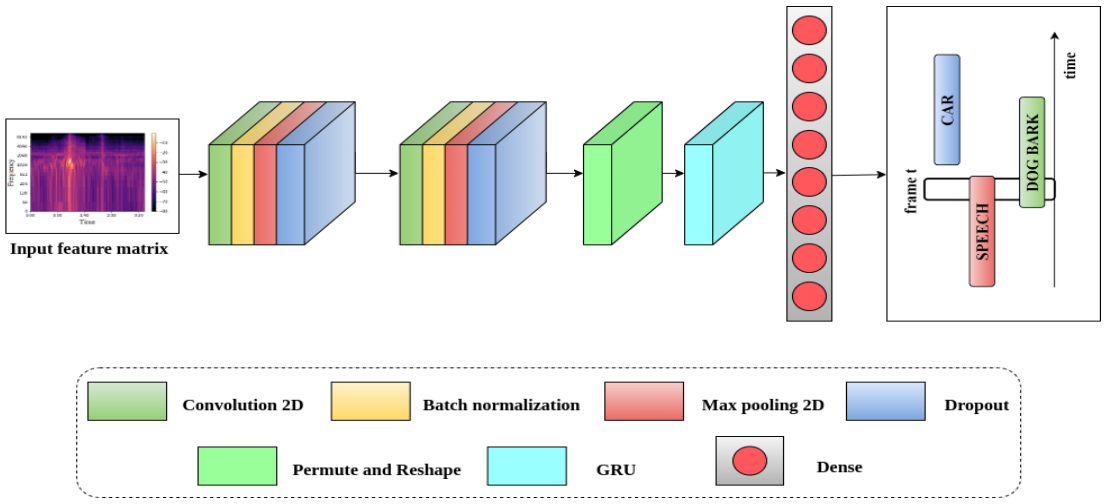

Figure 3.4: Layer configuration of the CRNN architecture

3.1.3 Performance Evaluation

In order to evaluate the performance of the proposed method, the datasets used are TUT Sound Events 2016 (TUT-SED 2016) and TUT Sound Events 2017 (TUT-SED 2017). The results are presented by comparing the provided annotated reference labels and the outputs obtained from the proposed method.

The results are reported in Tables 3.1 and 3.2. To make the proposed system more robust, additive noise is added to the audio recordings and tested using the proposed model. The results obtained with additional noise are also presented in Tables 3.1 and 3.2. Analysis of different transformation techniques such as Short-Time Fourier Transform (STFT), CQT, and the proposed approach Mel Pseudo CQT is presented in the sections to follow:

A. Performance of the proposed approach

Table 3.1 illustrates the results obtained for different filter sizes for clean and noisy audio recordings. The results obtained from different noise (Signal-to-Noise Ratio (SNR)) levels is presented in the table. The variation in the filter size in the network architecture exhibited differences in both evaluation metrics-error rate (ER) and F1-score (F1). The average value for clean audio is achieved using a filter size of 64 with ER of 0.660 and F1-score of 54.0%. The average ER and F1 obtained for noisy audio is 0.684 and 52.3% respectively. As it can be observed from the Table 3.1, the results obtained for the filter size of 8 are the lowest and the results obtained for the filter size 64 are highest. There was a slight decrease in the ER value for filter sizes 128 and 256. Also, there was no significant improvement in F1 measure as well. This indicates that the proposed model works better for the lower receptive field of the filter i.e, when the filter sizes are set less than or equal to 64. Event information present in a feature is only for a certain region. This type of feature information is captured by filter sizes of 16, 32, and 64. Also, normally, if there is less amount of data, the model tends to overfit (Mulimani and Koolagudi, 2019b).

In this work, we also evaluated the robustness of the model. A model is said to be robust, when the performance of the model is not affected by the noise. Robust

Table 3.1: Polyphonic SED performance in terms of Error and F1 scores of Mel-pseudo CQTs as the features and CRNN model as a classifier in the case of clean and noisy audio recordings of TUT-SED 2016 dataset

Clean Audio			**Noisy Audio**			
Metric	**Filter Size**	**Values**	**Performance in Different Noise (SNR) Levels**			
			20 dB	**10 dB**	**0 dB**	**Average**
Error	8	0.674	0.696	0.713	0.745	0.707
F1	8	52.8%	51.2%	49.8%	47.1%	50.2%
Error	16	0.667	0.692	0.715	0.748	0.705
F1	16	52.9%	51.6%	49.8%	47.9%	50.5%
Error	32	0.661	0.688	0.711	0.724	0. 696
F1	32	53.1%	52.5%	50.7%	48.3%	51.1%
Error	64	**0.660**	0.684	0.692	0.703	0.684
F1	64	**54.0%**	52.9%	51.9%	50.5%	52.3%
Error	128	0.662	0.693	0.695	0.717	0.691
F1	128	53.8%	51.1%	51.4%	49.9%	51.5%
Error	256	0.679	0.703	0.779	0.793	0. 738
F1	256	53.3%	51.6%	49.8%	48.1%	50.7%

Table 3.2: Polyphonic SED performance in terms of Error and F1 scores of Mel-pseudo CQTs as the features and CRNN model as a classifier in the case of clean and noisy audio recordings of TUT-SED 2017 dataset

Clean Audio			**Noisy Audio**			
Metric	**Filter Size**	**Values**	**Performance in Different Noise (SNR) Levels**			
			20 dB	**10 dB**	**0 dB**	**Average**
Error	8	0.532	0.598	0.617	0.626	0.613
F1	8	74.5%	73.6%	73.0%	71.8%	72.8%
Error	16	0.488	0.494	0.503	0.532	0.509
F1	16	75.8%	75.1%	74.6%	73.8%	74.5%
Error	32	0.327	0.364	0.409	0.510	0.427
F1	32	78.6%	77.8%	77.0%	76.2%	77.0%
Error	64	**0.210**	0.278	0.284	0.302	0.288
F1	64	**80.1**%	79.3%	78.7%	78.0%	78.6%
Error	128	0.228	0.259	0.307	0.416	0.327
F1	128	80.0%	79.7%	78.2%	78.1%	78.6%
Error	256	0.243	0.262	0.297	0.313	0.290
F1	256	79.1%	78.4%	76.2%	75.1%	76.5%

system is expected to yield a better polyphonic SED system in real-life scenarios. In order to simulate a noise, an additional Gaussian noise is added to all the input audio recordings of TUT-SED 2016 and 2017 development datasets and the algorithms are re-run. Different SNR levels such as 0 dB, 10 dB, and 20 dB are added to the input audio recordings. The addition of noise in the audio displayed considerable performance in terms of both ER and F1 measure. From Tables 3.1 and 3.2, it can be observed that the best performing result is achieved from the filter size of 64 with average ER of 0.684 and 0.210 and average F1-score of 52.3% and 80.1% for TUT-SED 2016 and 2017 datasets respectively. The reason for the proposed Mel pseudo CQT features performing well is that CQT features are able to capture lower frequencies at higher resolution even in noisy background and Fourier Transform fails to do so. It can be stated that, even in noisy conditions, the proposed method performed at par or better as compared to clean conditions.

B. Analysis of different spectral representations

Different spectral representations of two audio recordings used in this study are shown in Figure 3.5. They are Short time Fourier Transform (STFT), Constant Q-Transform (CQT), and the proposed Mel Pseudo CQT. The most commonly used spectral representation is STFT in the literature for the polyphonic SED task. In the Figure 3.5, the circles mark the events occurring in the audio recording. The spike in the audio gives the indication of a sudden event with high frequency but of short duration. The white circles in the spectrograms are the event information captured. It can be observed that, STFT captures only some of the event information and also has the background noise. CQT has also exhibited very similar spectrogram as of STFT but background noise is significantly reduced. In Mel Pseudo CQT, event information is captured better and the background noise is observed to be minimal as compared to STFT and CQT transformation techniques. Therefore, it can be said that Mel Pseudo CQT based spectrogram features can be used to improve the performance of the polyphonic SED systems.

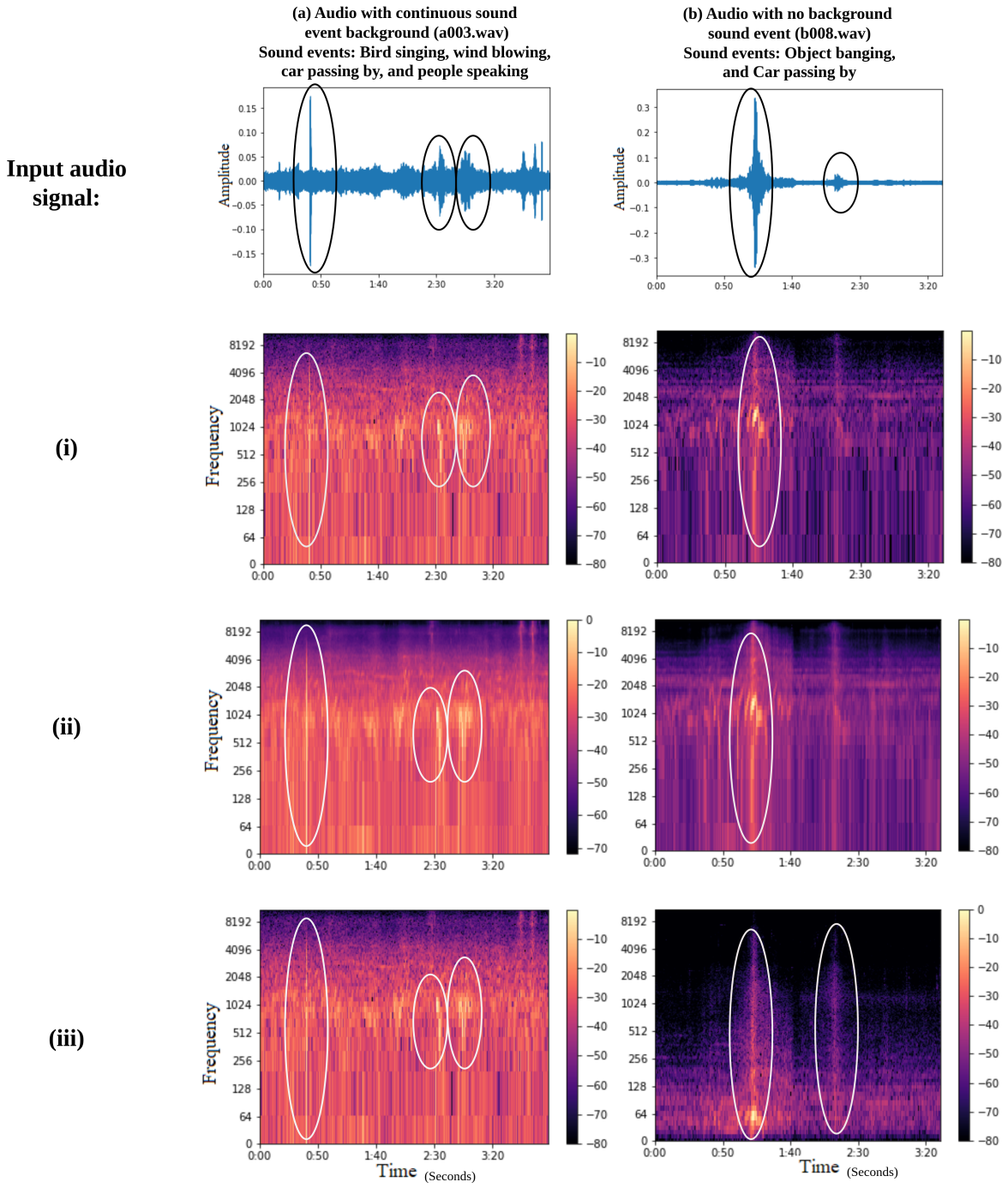

Figure 3.5: Different spectral representations, (i) STFT, (ii) CQT, and (iii) Mel Pseudo CQT (proposed method) for two different audio files chosen from TUT-SED 2016 dataset

C. Performance comparison of the proposed method with state-of-the-art SED methods

The proposed method is compared with the state-of-the-art polyphonic SED methods that use advanced classifiers like FNN, CNN, CRNN, DNN, and TCRNN. The majority of these studies have used conventional feature representations like log-mel band energies in their work. The key reason of performance improvement in the proposed work is the usage of Mel Pseudo CQT features that help in capturing better event information as compared to log Mel band energies. Table 3.3 contains the average ER and F1-score values of the other state-of-the-art methods. A significant improvement is observed in both average error rate and F1-score from the baseline method. The proposed method has achieved better results as compared to the existing other state-of-the-art approaches proposed for polyphonic SED.

3.1.4 Contributions and Limitations

In this work, a Mel-Pseudo CQT for feature extraction along with CRNN as classifier are used to perform polyphonic SED. The method extracts high level discriminative features using pseudo CQT technique and is compacted by applying mel scale on the transformed features. The extraction of CQT features is more advantageous as compared to traditional Fourier Transform based spectrogram features. The reason is that human beings perceive the sound in logarithmic scale which aligns with how our auditory system processes sound intensity and frequency; and CQT tends to capture this sound information, however, the Fourier Transform uses fixed frequency bins. Therefore, the proposed Mel Pseudo CQT features capture better event information as compared to log Mel band energies which are used in existing polyphonic SED systems. However, there is one disadvantage of using CQT features. The computation complexity of CQT features is more as compared to log Mel band energies.

The obtained features are then fed to a combination of CNN and RNN architecture that complements both architectures by learning the time and frequency features from the input feature representations. The overall framework resulted is used to get the event activity probabilities.

Table 3.3: Comparison of F1 score and error rate of different approaches on the TUT-SED 2016 dataset. (Legend: ER- Error rate, F1- F1 score)

Sl no.	Title	Method	F1	ER	Remarks
1.	TUT database for acoustic scene classification and sound event detection (Mesaros et al. 2016b)	Features: MFCCs, Classifier: GMM	23.7%	0.91	The system was given as baseline for TUT-SED 2016 challenge. The system uses MFCC features and CNN as classifier. This system was the first to introduce deep learning models to perform SED and resulted in better performance as compared to traditional GMM-HMM model for SED task.
2.	Convolutional recurrent neural networks for polyphonic sound event detection (Cakır et al. 2017)	Features: Log Mel band energies, Classifier: CRNN	30.3%	0.95	The combination of convolutional and recurrent layers resulted in better learning of both spatial and temporal information of log mel band energies for polyphonic SED
3.	Polyphonic sound event detection using transposed convolutional recurrent neural network (Chatterjee et al. 2020)	Features: Mel-IFgram features, Classifier: TCRNN	38.7%	0.92	A transposed convolution layer was introduced in this work trained with Mel-IFgram (Instantaneous Frequeny) features to perform polyphonic SED. The results displayed improved F1 score; however, the ER is high.
4.	A deep neural network-driven feature learning method for polyphonic acoustic event detection from real-life recordings (Mulimani et al. 2020)	Features: Log Mel band energies, Classifier: DNN-driven feature learning	44.5%	0.71	The combination of CNN layer and projection layers in the work resulted in improved performance in their work from the baseline.
5.	Polyphonic Sound Event Detection Using Mel-Pseudo Constant Q-Transform and Deep Neural Network (Spoorthy and Koolagudi 2023b)	Proposed Features: Mel-Pseudo CQT spectrograms, Classifier: CRNN	**54.0%**	**0.66**	In the proposed approach, a new Mel-Pseudo CQT technique based spectrogram features are used to train on CRNN deep learning model. As events present in the audio recordings of the dataset are sparse in nature, the features displayed better performance in terms of both F1 score and ER.

The pseudo CQT technique has reported better discriminative features than conventional STFT + log mel band energies. A significant improvement is observed by the proposed method in terms of both average ER and F1-scores specifically for polyphonic SED.

The take away point of this work is that the combination of different transformation techniques and classifiers is effective in the case of polyphonic SED. The audio recordings for SED task are highly diverse based on the recording environment. Example, if there is busy traffic road, there are more sound events present in the recording, unlike a silent park, where less sound events may be present. The occurrence of events is usually sparse (meaning one or more sound events can occur for a short burst of time and then there can be silence in the rest of the signal) and rest of the audio is normally silence. Therefore, use of pseudo CQT technique for feature extraction resulted in better discriminative spectrograms resulting in improved performance in polyphonic SED.

3.2 POLYPHONIC SOUND EVENT DETECTION USING MODIFIED RECURRENT TEMPORAL PYRAMID NEURAL NETWORK

The schematic diagram of the proposed polyphonic SED system is shown in Figure 3.6. The system is divided into two important phases, namely, feature extraction and detection of sound events with the use of neural network architecture. The features used to identify different sound events are spectrogram features extracted using Constant-Q Transform (CQT). CQT-based spectrograms are fed as input to the neural network architecture. The network proposed to identify and detect sound events is the Modified Recurrent Temporal Pyramid Neural Network (MR-TPNN). The network uses a temporal pyramid pooling layer to get fixed-dimensional output even though the input features have variable dimensions. The recurrent layer used in the network is Long-Short Term Memory (LSTM). Detailed information about each of the phases is given in the subsections to follow.

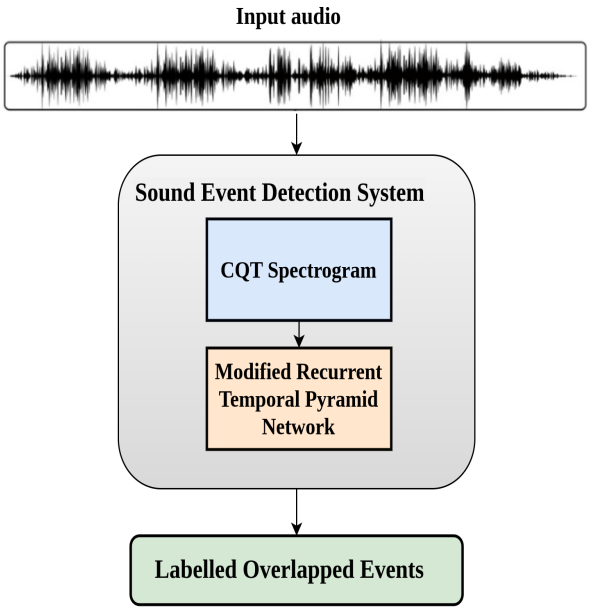

Figure 3.6: Schematic diagram of the polyphonic SED using CQT spectrogram as features and Modified Recurrent Temporal Pyramid Network as classifier

3.2.1 Feature Extraction from CQT-based Spectrogram

Log Mel-band energies are the features used in the development of the baseline system applied on TUT Sound Events 2017 challenge (Mesaros et al. 2017). The spectrograms generated in the baseline system use Fast Fourier Transform (FFT) as the transform function. The transformation function used in this work is CQT which is a time-frequency representation in which all frequency bins have the same Q-factors (ratios of their centre frequencies to bandwidths). The frequency bins are geometrically spread out in the CQT. Due to the fact that the CQT is essentially a wavelet transform, lower frequencies have better frequency resolution and higher frequencies have better time resolution (Schörkhuber and Klapuri 2010).

The transform may be viewed of as a series of filters, where each filter f_y has a spectral width δf_y that is a multiple of the filter f_y that came before it as given in Equation 3.6:

$$\delta f_y = 2^{\frac{1}{n}} \cdot \delta f_{y-1} = (2^{\frac{1}{n}})^y \cdot \delta f_{min} \tag{3.6}$$

Where, δf_y is the bandwidth of the y^{th} filter, f_{min} is the lowest filter's central

frequency, and n is the number of filters per octave (Patil et al. 2022; Wang et al. 2019a). The feature extraction in this work is carried out by transforming the input audio recordings using CQT instead of the conventional STFT or FFT technique. Figure 3.7 shows the spectrogram obtained from STFT and CQT algorithms for given input audio recordings chosen from TUT Sound Events 2017 dataset. The circled regions in the Figure 3.7 indicate the sound activity in the audio recordings. Though the Mel spectrogram has captured the sound event activity, CQT-based spectrogram exhibits better visual cues for different sound events. The CQT features capture the lower frequency components in higher resolution. It can be observed from the figure that as compared to Mel spectrogram, CQT spectrogram captures better event information from the input audio signal.

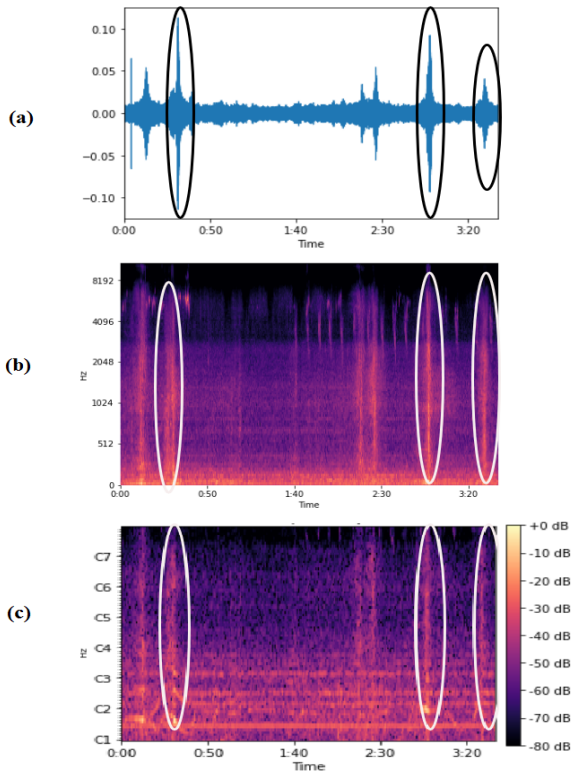

Figure 3.7: Illustration of Mel and CQT spectrograms for an audio recording chosen from TUT Sound Events 2017 development dataset: (a) Input audio, (b) Mel spectrogram, and (c) CQT spectrogram

3.2.2 Event detection using Modified Recurrent Temporal Pyramid Neural Network (MR-TPNN)

MR-TPNN architecture proposed in this work to perform polyphonic SED is shown in Figure 3.8. The network consists of convolutional layers for learning features from CQT spectrograms, an average pooling layer for retaining the relevant information from the feature map, a temporal pyramid layer to convert the input into fixed dimensional vectors, bi-directional LSTM provides the flow of sequence information in both forward and backward direction, at the end, the fully connected layer is used to convert the output matrix into a single dimensional vector.

In the proposed architecture, linear Swish activation function is used in the hidden layers. The Swish activation function formulated using the Equation 3.7:

$$swish(a) = a.sigmoid(\beta a) = \frac{a}{1 + e^{-\beta a}} \tag{3.7}$$

Where, β is a trainable parameter or a constant depending on the requirement of the model. Unlike Rectified Linear Unit (ReLU as an activation function) that thresholds all negative weights to zero, the Swish activation function allows a small number of negative weights to be propagated through. It is necessary to achieve a non-monotonic smooth activation function for better learning in deep neural networks (Prajit Ramachandran 2018). The usage of Swish activation function is done in the hidden layers (convolutional layers).

Temporal Pyramid Pooling (TPP): The output from the convolutional layers is usually varying in size. The working of TPP layer is shown in Figure 3.9 In order to get a fixed-dimensional feature and vector from different temporal scales, we have employed a Temporal Pyramid pooling method inspired by various computer vision solutions (Ma et al. 2020; Yu et al. 2019). The windows used for pooling operation are adaptive as opposed to other pooling operations such as Max pooling or average pooling. A pyramid level consisting of n bins (In Figure 3.9, n=3 i,j, & k), the max-pooling window is traversed across the feature map through time, where i^{th} bin is related to the feature map within $\left[\left\lfloor \frac{i-1}{n} L \right\rfloor, \left\lceil \frac{i}{n} \right\rceil L\right]$, where L is the row dimension of the input feature matrix obtained from the predecessor convolution layer (Yu et al. 2019). The

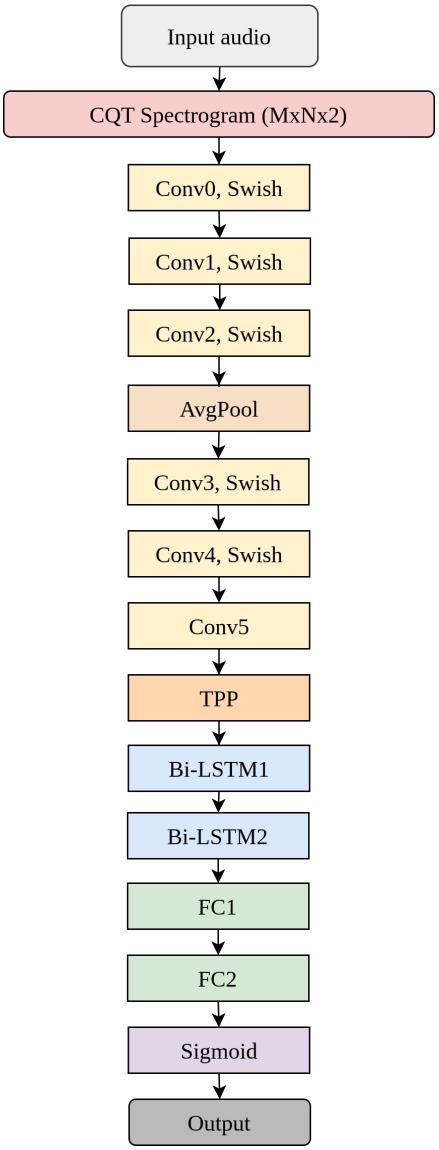

Figure 3.8: Architecture of the proposed MR-TPNN (Legend: Conv-Convolution, AvgPooling-Average Pooling, TPP-Temporal Pyramid Pooling, Bi-LSTM-Bi-directional Long Short Term Memory, FC- Fully-Connected)

number of levels in the TPP layer depends on the input feature map. After obtaining the pooling values of different pyramid levels, the values are concatenated to obtain a fixed-dimensional vector for further processing. The main purpose of TPP is to convert a variable-lengthed feature map into a fixed-length output. This layer helps in capturing and aggregating different sound events spread on different temporal scales.

MR-TPNN consists of different neural network layers. The convolutional layers' kernel sizes range from 3 to 5. The pooling layer's kernel is set to a size of 3. There are

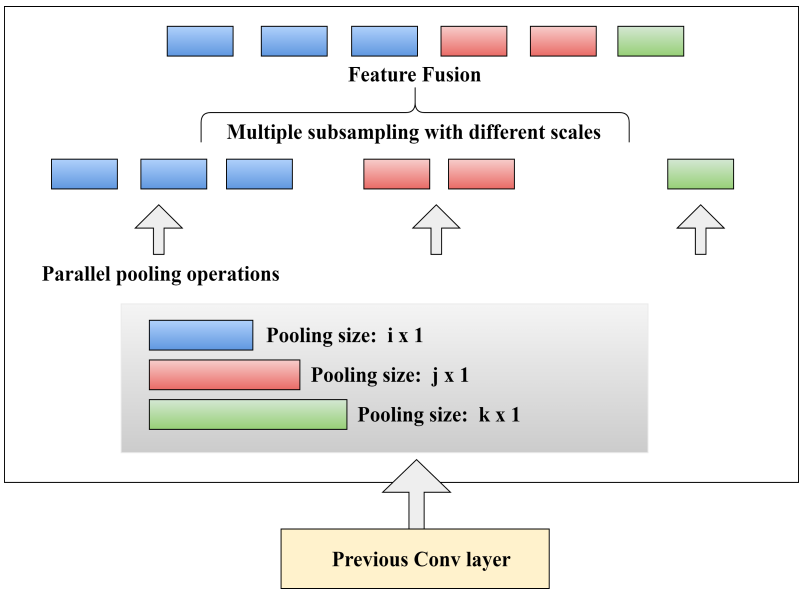

Figure 3.9: Working of Temporal Pyramid Pooling Layer

two Bi-directional LSTM layers in the model. True outputs from the first layer are fed to the second layer, and the true outputs of the second layer form the output of the second LSTM layer of the network. The model has last two layers are Fully connected (FC) layers. The first FC layer is used to convert the previous layer's output into a vector and the second FC layer consists of number of neurons equivalent to the number of output classes. With a learning rate of 0.001, Adam is an optimizer used to train the network. The activation functions used in the hidden layer and final layer are Swish and Softmax, respectively. The number of levels provided for the Temporal Pyramid Pooling is 3. The training and testing datasets are same as the DCASE challenge evaluation setup.

The loss function used is weighted focal loss (Qin et al. 2018) which is suitable for multiclass classification. The computation of the weighted focal loss is as per the Equation 3.8:

$$FL(A_t) = -\sigma(1 - a_t)^{\gamma} log(a_t) \tag{3.8}$$

Where, FL represents weighted focal loss, A_t is the estimated probability of the model for different classes which ranges between [0,1], a tunable modulating factor γ is added to the cross entropy loss function with a value $\gamma > 0$ (Lin et al. 2017), and σ is the weight factor added .

3.2.3 Performance Evaluation

In order to evaluate the performance of the proposed method, the experiments are conducted on a publicly available TUT-SED 2016 and TUT-SED 2017 datasets. The results are compared with the annotated reference labels. The audio recordings in the dataset are of varying lengths. The performance measures used to evaluate the proposed method are Error Rate (ER), and F-measure as per DCASE challenge evaluation (Mesaros et al. 2016a). The results obtained from the log-Mel energies and CQT-based spectrogram are fed as input to the MR-TPNN classifier are presented in Tables 3.4 and 3.5. The significant improvement is observed in the cases of both ER, and F-measure for CQT spectrograms.

Table 3.4: Error rate and F-measure values using log mel energies as features and proposed MR-TPNN as classifier for polyphonic SED

Dataset	ER	F-measure
TUT Sound events 2016 (Dev.)	0.64	48.2%
TUT Sound events 2017 (Dev.)	0.63	47.9%

Table 3.5: Error rate and F-measure values using CQT based spectrograms as features proposed MR-TPNN as classifier for polyphonic SED

Dataset	ER	F-measure
TUT Sound events 2016 (Dev.)	0.60	52.7%
TUT Sound events 2017 (Dev.)	0.58	51.8%

The proposed approach is compared with the existing polyphonic SED systems. The comparative observation is shown in Table 3.6. To evaluate the performance of the proposed approach, state-of-the-art polyphonic SED methods are also considered for comparison, and are mentioned below:

- **Baseline system:** The baseline system using Mel-Frequency Cepstral Coefficients (MFCC) and CNN is developed for DCASE SED 2016 challenge (Mesaros et al. 2016b). This system was the first to introduce deep learning models for SED. The baseline method performed better as compared to traditional GMM models for SED task. However, the ER obtained by this system is 0.91, which is very high.

- **CRNN based system**: The features used for training the CRNN model are logarithmic Mel band energies with 40 mel bands extracted from monoaural audio recordings. The CRNN network has consisted of three convolution layers with 96 filters in each layer, followed by a max-pooling layer. The output of the convolution layers is fed to three Long Short Term Memory (LSTM) layers where each layer consists of 256 hidden units. The output of the LSTM layer is given to the fully connected/dense layer which has the hidden units equal to the total number of classes (Cakır et al. 2017). The combination of convolution and recurrent layers exhibited improved performance as it captured both spatial and temporal information from the input features.

- **Transposed Convolutional Recurrent Neural Network (TCRNN):** The method uses Mel-IFgram (Instantaneous Frequency spectrogram) features to perform polyphonic SED task. The network uses convolution operation where the spatial transformation is reverted and is provided to the RNN layer. Hence, the name Transposed CRNN (Chatterjee et al. 2020). This model uses a Transpose convolution layer which helps in capturing better spatial information from the input. This model exhibited improved F1 score, however, the ER obtained by this model is 0.92, which is high for a reliable polyphonic SED model.

- **DNN-driven feature learning**: In this method, the authors have used a combination of multiple layered DNN to perform polyphonic SED. The features fed to the network are 60 monaural framewise log Mel band energies along with their deltas and acceleration coefficients (Mulimani et al. 2020). In this work, a novel projection layer has been introduced to learn better features in the hidden layers of the neural network which helped in improving the performance of the model.

3.2.4 Contributions and Limitations

In this work, a deep learning model which learns the temporal information from the given input features is proposed to identify the onset and offset times of various sound

Table 3.6: Listing of Error rate and F-measure values obtained from the proposed MR-TPNN and the existing polyphonic SED systems evaluated on DCASE 2016 Sound Events Development Dataset (Legend: ER- Error rate)

Sl.No	Title	Method	ER	F-Measure
1	TUT database for acoustic scene classification and sound event detection (Mesaros et al. 2016b)	Baseline	0.91	23.7%
2	Convolutional recurrent neural networks for polyphonic sound event detection (Cakır et al. 2017)	CRNN	0.95	30.3%
3	A deep neural network-driven feature learning method for polyphonic acoustic event detection from real-life recordings (Mulimani et al. 2020)	DNN-driven feature learning	0.71	44.5%
4	Polyphonic sound event detection using transposed convolutional recurrent neural network (Chatterjee et al. 2020)	TCRNN	0.92	38.7%
5	-	Proposed MR-TPNN	**0.60**	**52.7%**

events along with the event labels. The proposed MR-TPNN used a pyramid pooling approach to get a fixed-dimensional feature vector. The use of a Bi-directional LSTM layer helps in retaining the temporal sequence information from the input. This, in turn, provided improved performance in terms of ER and F-measure for the polyphonic SED task. The results of the experiment on two benchmark datasets demonstrate that the proposed approach performed better than existing other state-of-the-art polyphonic SED systems.

3.3 SUMMARY

In this chapter, two new methods have been proposed to perform detection of polyphonic sound events. The first method introduces a feature extraction method that extracts better discriminative features to capture the event information present in an audio recording. The method uses Pseudo CQT features which work better for sparse audio. The second method proposed for polyphonic SED is a new deep learning method which learns better temporal information from the input feature. The temporal

information is very important for detection of onset and offset of a sound event present in an audio recording. Both approaches exhibited good performance in terms of ER and F1 score. In the next chapter, the localization and detection of the overlapped events will be addressed.

CHAPTER 4

SOUND SOURCE LOCALIZATION AND DETECTION OF ACOUSTIC EVENTS

Sound Event Localization and Detection (SELD) identifies the source of a sound event along with the onset and offset time instances of the same and assigns a semantic label to the detected sound event. The joint task of localization and detection is more beneficial in critical scenarios. Identifying the source of the sound event along the sound event type is an extension of Sound Event Detection (SED) task. This chapter includes two new methods to perform localization and detection of overlapped/polyphonic sound events. The new deep learning architectures are Transpose Sound Event Localization and Detection Network (SELDNet) and Channelwise FusionNet.

4.1 SOUND EVENT LOCALIZATION AND DETECTION USING TRANSPOSE SELD-NET

The tasks of Sound Event Detection (SED) and Sound Source Localization (SSL) are jointly known as Sound Event Localization and Detection (SELD). Figure 4.1 illustrates the schematic architecture of the proposed SELD system. The input audio is fed to the feature extractor, which generates a feature matrix of size *number of rows (M)*$\times$*(64 Log-Mel energies+3 Intensity vectors (I.V)*$\times$ *Number of channels (i.e., 7 in this case).* In this work, two variants of SELDNet are introduced to perform both SED and SSL tasks in the single model. The proposed networks are D-SELDNet (Figure 4.1, b) and T-SELDNet (Figure 4.1, c) inspired by SELDNet (Figure 4.1, a) of DCASE challenge (Adavanne et al. 2018a). In both networks, two types of convolution layers, namely,

depthwise and transpose are used. The subsections below provide a thorough overview of the feature extraction process and the details of the deep learning architecture used in the proposed SELD system.

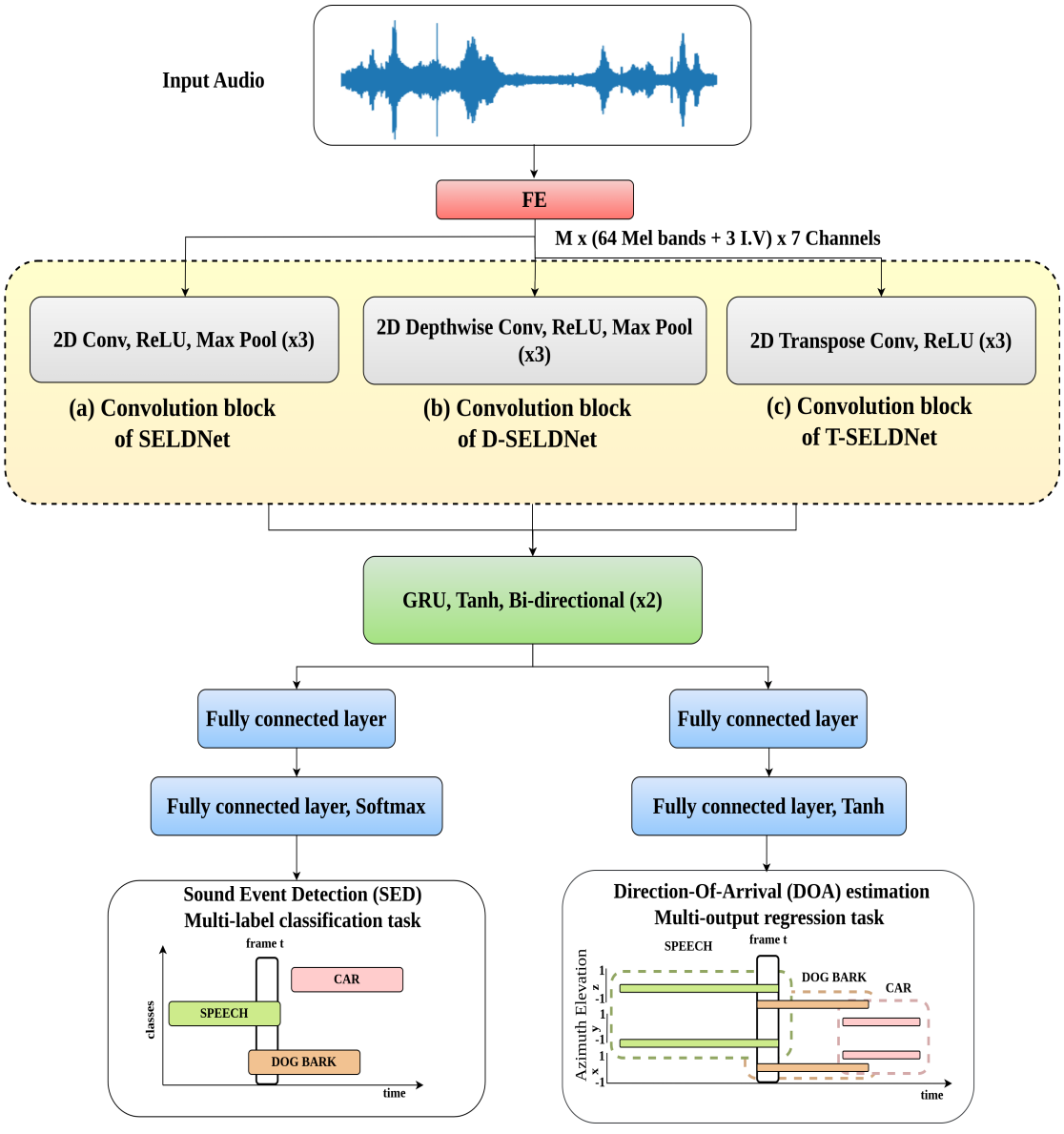

Figure 4.1: Illustration of proposed SELD system. The convolution blocks of the SELD system is different for three deep learning architectures, namely, (a) SELDNet (Baseline), (b) Proposed D-SELDNet, (c) T-SELDNet. Rest of the components in the networks are same for all three models. (Legend: FE-Feature Extractor, I.V- Intensity vectors)

4.1.1 Spectral Features for Sound Event Localization and Detection

A multichannel audio input is used to extract the spectral features. The features extracted for SED task are log-Mel band energies, by setting 40 ms window length and 20 ms as the overlap. 1024-point FFT is considered during spectrogram extraction. The features extracted for DOA estimation are acoustic intensity vectors. The features calculated from each of the 64 Mel-bands, give the signal's net acoustic energy flux.

First Order Ambisonics (FOA) is the type of multichannel input provided in the DCASE Spatial Sound Events dataset. FOA is a four channel signal extracted from four directions, namely, omnidirectional (w), x-directional (x), y-directional (y), and z-directional (z) components. Using the formula $I = pv$, the instantaneous sound intensity vector (I) is calculated. p stands for the sound pressure, which can be calculated using w, and v is the particle vector velocity, which can be calculated using $v = (v_x, v_y, v_z)^T$, calculated using x, y, and z. Intensity vector provides the information of a sound wave's acoustical energy direction, and by reversing the direction, we get DOA. Hence, intensity vectors are used as features to attain DOA on FOA. The log Mel-band energies result in four 64-dimensional channels, and the intensity vector results in three channels totalling to seven channels. The input sequence has a total dimension of 7xTx64, where T is the number of time frames in the sequence.

4.1.2 Event Detection and Direction-of-Arrival Estimation using Transpose SELD-Net

The proposed SELD system involves three different types of convolution layers, namely, convolution, depthwise separable convolution, and transpose convolution layers. The working of these three layers is illustrated in Figure 4.2. The following subsections provide more details on how each of the three convolution layers works:

A. Convolution layer: It is considered as the basic building block of any CNN (Goodfellow et al. 2016). The key components of the convolution layer are kernels/filters, and parameters such as padding, stride, regularizer, activation function, etc. The filter size set to convolve the image is usually much smaller than the original

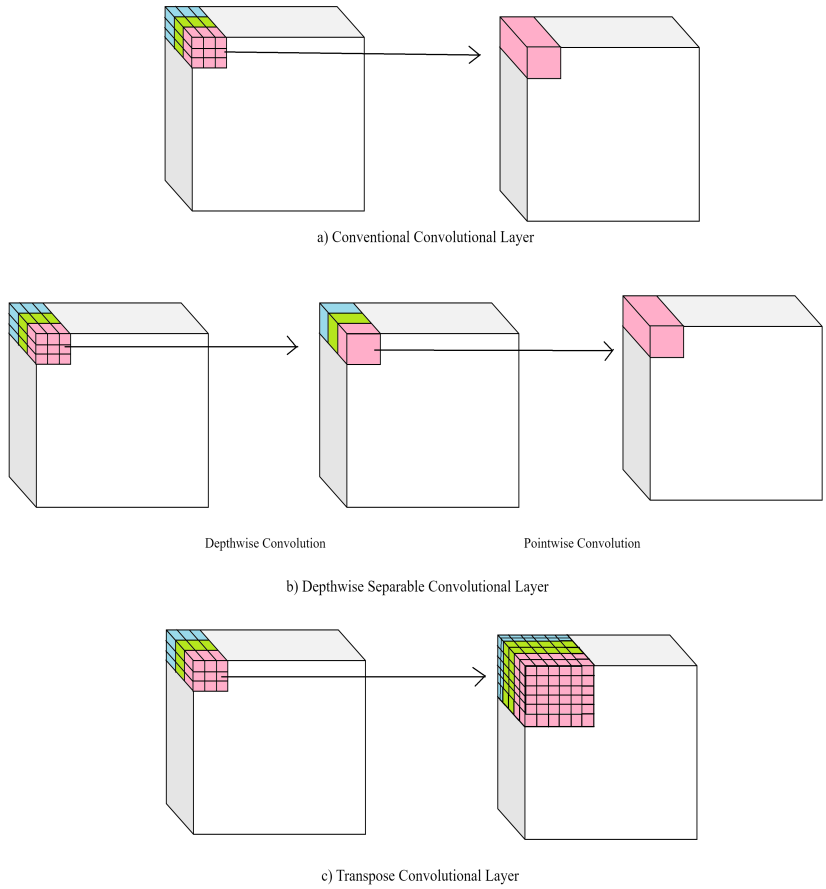

Figure 4.2: Convolution blocks used in the SELD system. (a) Conventional convolution layer, (b) Depthwise separable convolution layer, and (c) Transpose convolution layer

input image. The convolution layer's function is to transform an input image and extract distinguishing features from it. In the convolution operation, using the spatial position of input, a filter is slid across the input image. A dot product is computed between the filter/kernel and the input image. The illustration of a convolution layer output is shown in Figure 4.2, (a). A reduced feature map is obtained after the input is fed to convolution layer.

B. Depthwise separable convolution layer: In contrast to the conventional convolution operation, the depthwise separable convolution layer not only considers spatial dimensions but the depth dimension as well (i.e., the channel information). For instance, an input with 32 channels can be interpreted in 32 different ways. The working of depthwise convolution can be divided into two sets of operations, depthwise convolution, and pointwise convolution as shown in Figure 4.2, (b). The

convolution of an input image in the depthwise convolution is performed with a different kernel, i.e., a depthwise kernel. Following the depthwise convolution on all the channels, pointwise operation is applied. The main advantage of using depthwise separable convolutions in place of standard conventional convolutions in a CNN is that the former results in a lesser number of parameters, and is also faster in terms of computation time (Chollet, 2017).

C. Transpose convolution layer: The transpose convolution is equivalent to the implementation of the convolution layer with no strides and half padding (Shi et al. 2016). This layer is used to perform the up-sampling of the input image from a low-resolution feature space into a high-resolution output. The working of this layer is shown in Figure 4.2, (c). The transposed convolution produces an output larger than the input by broadcasting input elements via the kernel, in contrast to the standard convolution that decreases input elements via the kernel. The learnable parameters present in the transposed convolution layers are used to optimize the up-sampling. The output generated from the 2D transposed convolution layer is equal to or greater than the input fed to the layer.

Convolution layer output is routed to a layer called a Gated Recurrent Unit (GRU). Inherent temporal information present in the features is captured through GRU layers, and then given as input to a fully connected dense layer, which transforms the output of the GRU into a one-dimensional array. Further the output of the dense layer is separately used for the estimation of SED and DOA. Output of the SED block is a class label, where the event that is active in a particular frame gets a higher probability indicating that the respective event is active. The probability is given in a range of [0,1] for each event class in a frame. For each axis of the event class location, the DOA estimation block returns the output in the range [-1,1]. A true class is returned only if the event is active and the value of the DOA estimate is above 0.5 in a frame.

The hyperparameters of the neural networks in the proposed work are as follows: The convolution layers (standard, depthwise separable, transpose) have been experimented with varying filter sizes from 32 to 512 in number. The features fed as input to the deep learning architectures are log-Mel band energies (M$\times$64) along with

three intensity vectors (I) for each x, y, and z axis, where M is the number of rows in the spectrogram, this number varies according to the length of the audio). The number of Mel bands considered is 64. The loss function used for the neural networks is multi-class cross entropy. To avoid overfitting, the dropout layer is added to the network with value set to 0 to 0.5 with an interval of 0.1. The DOA estimation branch uses tanh as the activation function in the last layer because the output of DOA branch is the angular distance estimated in degrees, and tanh activation returns the value in the range of [-1,1]. However, SED branch uses softmax activation funtion as the output of the branch is probabilities of the presence of sound events in a time frame, and softmax activation returns the value in the range [0,1].

4.1.3 Performance Evaluation

The dataset considered for evaluating the proposed approach of SELD is TAU-NIGENS Spatial Sound Events 2020, which, consists of audio recordings with overlapped sound events. The maximum number of overlapped sound events present in any audio in the dataset is three. Table 4.1 presents the results of SELDNet, D-SELDNet, and T-SELDNet models. The metrics considered are Error Rate (ER), F-Score, DOA error, and Frame Recall (FR). The proposed SELD system is compared in two cases, namely, overlap-1, where two sound events are overlapped in a frame, and overlap-2, where three sound events are overlapped in a frame.

Table 4.1: Polyphonic SELD performance of the proposed D-SELDNet and T-SELDNet with baseline system (SELDNet) for overlap-1 and overlap-2 (Legend: Error Rate (ER), F-score (in %), DOA Error (in °), and Frame Recall (FR) (in %))

	Overlap-1				**Overlap-2**			
Method	**ER**	**F-score**	**DOA Error**	**FR**	**ER**	**F-score**	**DOA Error**	**FR**
SELDNet (Baseline)	0.57	53.5	14.2	53.0	0.72	36.9	21.5	53.0
D-SELDNet	0.54	56.3	13.9	68.1	0.70	39.0	20.2	55.2
T-SELDNet	**0.45**	**62.3**	**9.8**	**68.4**	**0.57**	**48.9**	**14.5**	**63.8**

In the audios used for testing, we have overlap-1 (two sound events at a given frame) and overlap-2 (three sound events overlapped at a given frame) types of files.

The performances achieved by the proposed D-SELDNet and T-SELDNet are better compared to the baseline system. The number of computations resulting from depthwise separable convolutions in D-SELDNet is less than the baseline SELDNet. In the T-SELDNet, the number of computations is higher as compared to the SELDNet because of the upsampling operation that is carried out in the convolution layer. Even though the network is computationally little more complex, the model resulted in appreciably better performance. The model retained better discriminative information in the transposed convolution layers as compared to standard convolutions and depthwise separable layers.

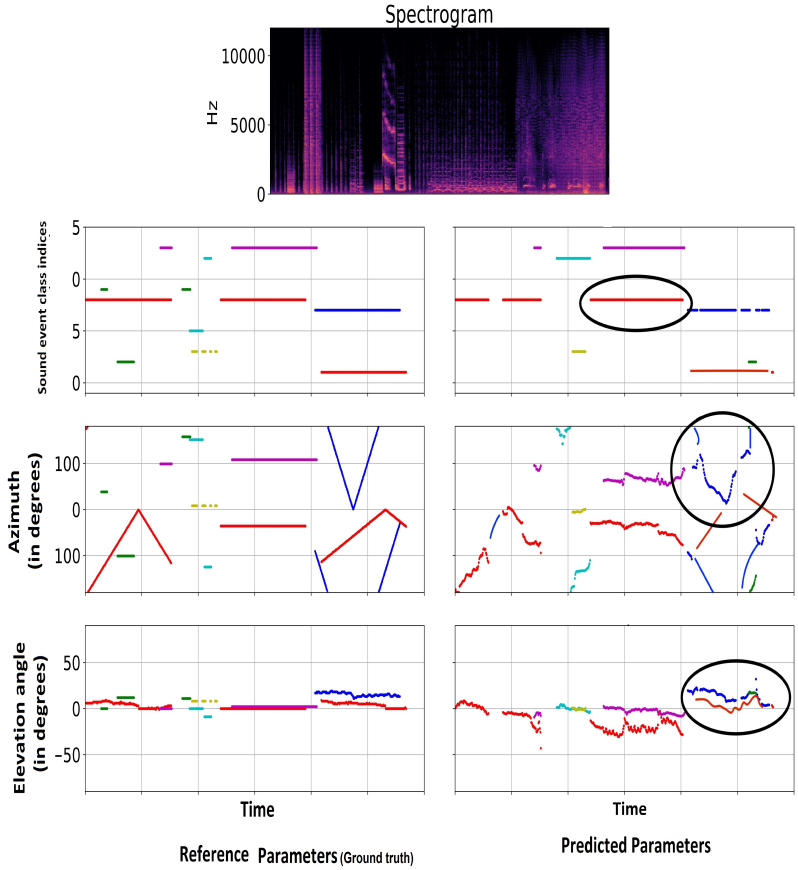

Figure 4.3: Output of the T-SELDNet: the audio recording (fold1_room1_mix001_ov1.wav) chosen from TAU-NIGENS Spatial Sound Events 2020 dataset. The sound events present in the audio are: running engine, female scream, male scream, burning fire, alarm, and crash

In Figure 4.3, the output obtained from the T-SELDNet for the task SED and DOA estimation are presented. The regions encircled in the predicted output are the output regions given by the proposed T-SELDNet. The regions show that the proposed method has successfully detected the onset and offset times of the sound event along with their corresponding sound event label. The figure illustrates the performance of the reference (plots of the left hand side) and the predicted (plots of the right hand side) metrics for multichannel audio input. In the figure, the result of one audio recording is illustrated. The input feature is spectrogram and the final output is given in three parts for two tasks, SED and DOA estimation. For SED estimation, reference SED plot indicates the onset and offset of various sound events at a given time frame. The predicted SED plot indicates the prediction of proposed approach of sound events. The DOA estimation is performed using two angles, Azimuth and Elevation angle. These are measured in degrees. It can be observed from the plots that the proposed approach is predicting sound events for unseen test audio samples as well. The method also provided accurate Azimuth and elevation angle indicating that the DOA estimation is predicted correctly.

The proposed T-SELDNet is also compared with the existing systems proposed for the SELD task. For a fair comparison, the methods that have used the FOA dataset are only considered. Table 4.2 presents the results of existing SELD systems along with our proposed approach. It can be observed that the results achieved from the proposed T-SELDNet are better as compared to the state-of-the-art SELD systems using the FOA dataset. To perform SELD, a CNN and RNN combination is employed. For the SELD task provided in the TAU-NIGENS Spatial Sound Events 2020 challenge, the SELDNet is treated as a baseline (Adavanne et al. 2018a).

To produce SED and DOA predictions, a technique that is frame-level permutation invariant is used (Cao et al. 2021). When detecting event activity, feature embedding information is taken into account to obtain the onset and offset time instants of events. This aided the network in making more accurate predictions. The use of the homogenous MSE loss function throughout the network for SED and DOA estimation in a multi-regression network is proposed (Phan et al. 2020). The performance of this method has increased over the baseline. However, it is still less than the proposed

T-SELDNet system.

4.1.4 Contributions and Limitations

The task of identifying a sound event's onset and offset in audio, along with the source of its origin and assigning a textual label to the sound event, is termed SELD. The task combined is more complex as compared to performing SED and SSL individually. In this work, two variants of the baseline system SELDNet have been proposed. Both the proposed networks performed better as compared to the baseline system. The proposed D-SELDNet resulted to be less complex in terms of computations as compared to SELDNet. The reason that depthwise separable convolution layers have been used in place of conventional convolutional layers is that they use less number of multiplication and addition operations. This resulted in the low-complex deep learning architecture. The proposed T-SELDNet utilizes transpose convolution, which helps in learning spatial features better as compared to SELDNet.

Table 4.2: Performance comparison of the proposed T-SELDNet with different approaches on the TAU-NIGENS Spatial Sound Events 2020 dataset (Legend: ER-Error rate, DOA-Direction-of-Arrival, FR-Frame recall)

Sl. No	Title	Method	ER	F-score (in %)	DOA error (in °)	FR (in %)	Remarks
1	Sound event localization and detection of overlapping sources using convolutional recurrent neural networks (Adavanne et al. 2018a)	Features: Phase and Magnitude spectrogram, Classifier: CRNN with two output branches: SED and DOA estimation	0.72	36.9%	21.5	53.0	This work uses a CNN and RNN combination to perform SELD. For the SELD task given in the DCASE 2020 SELD challenge, this system serves as a baseline system.
2	Audio event detection and localization with multitask regression network (Phan et al. 2020)	Features: Log-Mel magnitude spectrogram and intensity vectors, Classifier: CRNN architecture coupled with self-attention mechanism	0.59	50.8	18.2	64.1	A multi-regression network is proposed in this work, where MSE loss function is applied consistently for both DOA and SED branches. Although this method outperformed the baseline, it still performed poorer than the proposed approach.
3	Event-independent network for polyphonic sound event localization and detection (Cao et al. 2021)	Features: Spectrogram, Classifier: Event-dependent network with three output branches: SED, DOA predictions, and event activity detection (EAD)	0.47	61.5	16.7	75.4	In this work, SED and DOA predictions are obtained using a frame-level permutation invariant method. Event activity detection includes feature embedding information to obtain the on-set and off-set times of events. This improved the network's prediction performance.
4	A Transpose-SELDNet for Polyphonic Sound Event Localization and Detection (Spoorthy and Koolagudi 2023c)	Features: Log-Mel magnitude spectrogram and intensity vectors, Classifier: T-SELDNet	**0.45**	**62.3**	**9.8**	**68.4**	The proposed T-SELDNet method displayed better performance as compared to existing SELD systems. The network utilizes transpose convolution layer which helped in learning better spatial features.

4.2 SOUND EVENT LOCALIZATION AND DETECTION USING CHANNEL-WISE FUSIONNET

In the existing SELD approaches, the cases of a higher number of overlapping events are not considered adequately. The existing works have addressed only one or two overlapped sound events. To get better event detection performance, the input data must contain many sound event labels and better spatial resolution. These methods are not generalizable in terms of features, therefore, input array structure has to be predefined. Input to the model is constrained to the processing of the multichannel input as a whole. However, from our experimentations, it can be observed that there is meaningful information present in individual channels of the input. This information is lost when processed as a whole. Therefore, a novel neural network is designed to utilize channel-wise information of the multichannel input and fuse the learned feature representations hierarchically to enhance the performance of both SED and SSL tasks in the cases of many overlaps.

In contrast to the existing SELD techniques, the proposed method is novel from two aspects. Firstly, a 'Fusion layer' is introduced in the conventional neural network architecture, which blends the output feature maps of all the channels and returns a combined feature map. Secondly, extensive evaluation studies are presented for the proposed method. Both SED and DOA estimation tasks are performed simultaneously by the model.

Figure 4.4 depicts the architecture of the proposed channel-wise FusionNet model. Mel-band energies and intensity vectors are extracted from each audio channel as features. The proposed FusionNet, takes input, a set of features from consecutive spectrogram frames and predicts active sound events for each input frame. Sound source along with a DOA trajectory and temporal activity for each sound event are also predicted.

4.2.1 Feature extraction: Mel-band power spectrogram and Intensity vectors

The proposed FusionNet takes a multichannel audio input. Two sets of features are extracted from the multichannel audio input, namely, Mel-band power spectrogram and

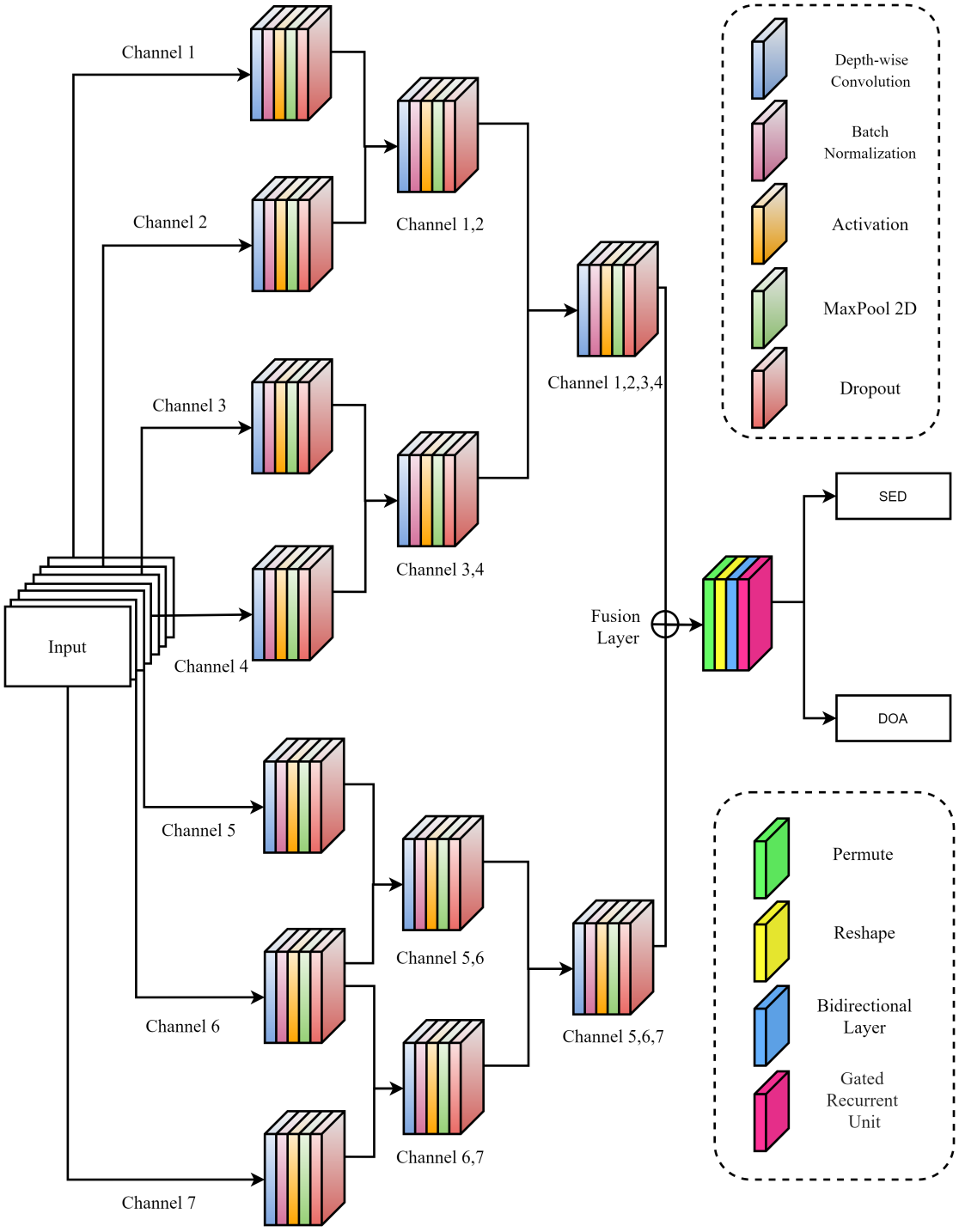

Figure 4.4: Schematic diagram of proposed channel-wise FusionNet for SELD

acoustic intensity vector. Each channel's spectrogram features are calculated as a set of 64 log Mel-band energies with a window length of 40 ms and a 50% overlap. 1024-point FFT is considered for spectrogram generation. The acoustic intensity vectors give the signal's net acoustic energy flux and are computed for each of 64 Mel-bands. This

feature extraction procedure is adapted from SELDNet proposed for the SELD task in TAU-NIGENS Spatial Sound Events 2020 baseline system (Harma et al., 2005).

4.2.2 Channelwise FusionNet architecture details

The proposed neural network model receives the input features from the feature extraction as a HxTxF matrix, where H denotes the number of feature channels, T denotes the temporal length of the modelled sequence (number of time frames), and F denotes the feature dimension. The matrix is split into 1xTxF, 2xTxF, ... , HxTxF individual channels. Each channel's log Mel-band energies and intensity vectors are extracted and fed into the FusionNet separately. Each channel is fed to a separable convolution block, followed by batch normalization, activation function, max-pooling operation, and drop-out layer. In the separable convolution block, one filter is divided into two smaller filters.

The operation of separable convolution block is divided into two steps, namely, depthwise convolution and pointwise convolution. This convolution block is used to reduce the number of computations. The first step, depthwise convolution takes a single channel as input at a time unlike conventional convolution block where all the channels are fed at once. Therefore, the filter size will be $D_k, D_k, 1$, where (D_k, D_k) is the size of the input. Therefore, if there are M channels, then the filter size will be D_k, D_k, M. The cost of this operation will be $D_k \times D_k$ multiplications for a single convolution operation. As the filter is slid across the input with a stride D_p, the number of multiplications of a single convolution operation across all channels will be $M \times D_p \times\ D_p\ \times\ D_k \times D_k$. Therefore, the total number of computations can be given as $M \times D_p^2 \times\ D_k^2$.

The second step, pointwise convolution is a 1×1 convolution operation which is applied on all the input channels. The corresponding filter size is $1 \times 1 \times M$. For N number of filters, the size of the output is D_p, D_p, N. A single convolution operation requires 1 x M multiplications. As the filter is slid across the input on all channels, the number of computations becomes $M \times D_p \times D_p \times$(no. of filters). The total number of multiplications of the pointwise convolution operation is given by $M \times D_p^2 \times N$. The total number of multiplications of the separable convolutions is the addition of

multiplications obtained from depthwise and pointwise convolutions can be given by Equations 4.1 (Chollet 2017):

$$\text{Total number of multiplications} = M*D_k^2*D_p^2+M*D_p^2*N = M*D_p^2*(D_k^2+n) \quad (4.1)$$

Therefore, as compared to conventional convolutions, the number of multiplications performed by the separable convolutions are 100 times lesser (Chollet 2017). For example, if we consider a filter of 3x3, using spatial separable convolution, it can be separated into 3x1 and 1x3. This operation provides near the performance of the conventional convolution block and also reduces the number of multiplications. Thus, it is computationally less expensive, and also network runs faster. The final result of the network gives two outputs, SED and DOA estimation.

In the proposed FusionNet, the outputs of the first consecutive, adjacent channels are fused as shown in Figure 4.4. The process of fusing feature maps or the combination of feature maps is done till all the channels are fused to form a single feature map. This layer is named as "Fusion layer." Once the fusion layer is obtained, the temporal information is maintained through the RNN block, where the feature is reshaped and fed to a bidirectional GRU layer.

Training procedure: The SED branch consists of a sigmoid activation function, which returns probability values corresponding to the number of sound event classes. The DOA branch returns elevation and azimuth angles as localization outputs. This RNN block's output is used for two separate tasks: SED and DOA estimation. SED, the first output, is a multi-label classification task that predicts the sound events that occur in that frame. The RNN layer's output is fed into a fully connected (FC) layer with a sigmoid activation function. For each sound event, the SED output is in the range of 0 to 1, and this value is thresholded to a binary decision for that sound event. DOA estimation is performed as a multi-output regression task in the second output. This task involves estimating the DOA's 3D Cartesian coordinates X, Y, and Z in 3D space on a unit sphere around the microphone. Each sound event class has three coordinates assigned to it. Finally, the spatial locations of these sound event classes are returned by DOA estimates, ranging from -1 to 1 for each axis of the sound class location, so the

'tanh' activation is used in the regression to obtain the network's output.

Hyperparameters of the neural network are to be tuned differently for different layers. The layers that are considered in our work are the convolutional layer, separable convolutional layer, recurrent layer, and FC layer. Multiple filters varying from 32 to 512 in number are used in convolution and separable convolution layers. In order to achieve best performing network, the output loss weights for SED and DOA branch output are experimented with values set to 1, 5, and 50. The dropout value of the nodes is also experimented with values ranging from 0 to 0.5 with an interval of 0.1. The pooling layers in the network are experimented for different filter size ranging from 2 to 8. The best performing network's performance from all the parameters, i.e., the architecture resulting the lowest DOA and SED error on the validation split, is considered.

4.2.3 Performance Evaluation

The proposed Channelwise FusionNet for the SELD task is evaluated using the TAU-NIGENS Spatial Sound Events 2020 dataset Politis et al. (2020a). The audio recordings present in the dataset consist of different sound events that are recorded in multiple locations and also from different recording positions. Two types of recording formats were used to record the dataset, namely, Microscopic array (MIC) and First-order Ambisonics (FOA). In this work, the results are presented on the FOA dataset. Among six dataset splits, 3, 4, 5, and 6 are used for training, Split 2 is used for validation, and split 1 (unseen data) is used for testing the performance of the proposed FusionNet. The best parameters of the proposed method are chosen from the validation split and then used for testing the unseen data.

The results achieved in the baseline system provided for TAU Spatial Sound Events 2020 challenge, SELDNet with separable convolutions and proposed FusionNet in the cases of overlap 1 and overlap 2 are given in Table 4.3. The performance metrics are the same as those of TAU-NIGENS Spatial Sound Events 2020 metrics, namely, ER, F-score, FR, and DOA error. The proposed method displayed an improved performance for all the metrics for both overlap-1 and overlap-2 cases.

Table 4.3: Polyphonic SELD performance comparison of baseline system, SELDNet with separable convolutions, and proposed FusionNet on overlap 1 (two sound events are overlapped at a given frame) and overlap 2 (three sound events are overlapped at a given frame) TAU Spatial Sound Events 2020 dataset (Legend: Error Rate (ER), F-score (in %), Direction-of-Arrival (DOA) Error (in °), and Frame Recall (FR) (in %))

Method	**Overlap-1**				**Overlap-2**			
	ER	**F-score**	**DOA Error**	**FR**	**ER**	**F-score**	**DOA Error**	**FR**
Baseline	0.57	53.5	14.2	53.0	0.72	36.9	21.5	53.0
SELDNet (Sep.Conv)	0.54	56.3	13.9	68.1	0.70	39.0	20.2	55.2
FusionNet	**0.13**	**89.3**	**8.7**	**90.1**	**0.23**	**81.2**	**9.1**	**86.9**

The visual representation of performances of the baseline system, SELDNet with separable convolutions, and proposed FusionNet is demonstrated in Figure 4.5. From the visual representations of the results obtained by the proposed method, it can be noted that there is an improved F-score and Frame Recall values and minimized errors. Learning of discriminate features from different channels in the Fusion layer of the proposed FusionNet may be justified from the better performance of the SELD task in both SED and DOA estimation. In this figure, the results of two audio recordings are presented. First spectrogram is for one overlap (two sound events are overlapping at a given time frame) and second spectrogram is given for two overlap (three sound events are overlapping at a given time frame). In this figure, SED and DOA estimation results for both the inputs are shown. For SED estimation, reference SED plot indicates the onset and offset of various sound events at a given time frame. The predicted SED plot indicates the prediction of proposed approach of sound events. The DOA estimation is performed using two angles, Azimuth and Elevation angle. These are measured in degrees. The circled regions in the figure indicate the correct detection and localization of various overlapped and non-overlapped sound events for the proposed Channelwise FusionNet.

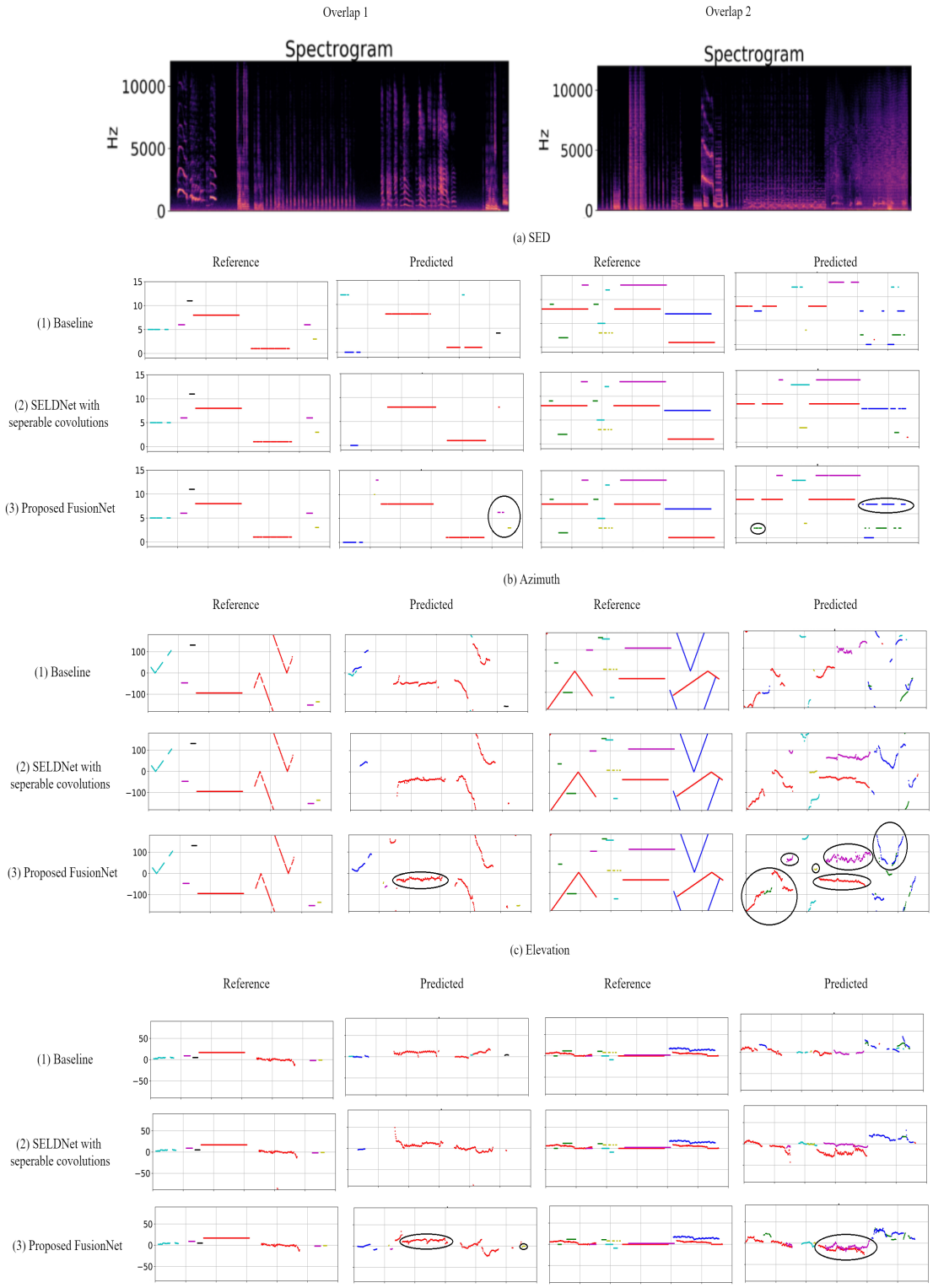

Figure 4.5: Output visualization of SELD task with baseline system, SELDNet with separable convolutions and proposed FusionNet for overlap-1 and overlap-2 events: (a) SED, (b) Azimuth, and (c) Elevation

Table 4.4: Performance comparison of the proposed Channelwise FusionNet with different approaches on the TAU-NIGENS Spatial Sound Events 2020 dataset (Legend: ER-Error rate, DOA- Direction-of-Arrival, FR-Frame Recall)

Sl. No	**Title**	**Method**	**ER**	**F-score (in %)**	**DOA error (in °)**	**FR (in %)**	**Remarks**
1	Sound event localization and detection of overlapping sources using convolutional recurrent neural networks (Adavanne et al. 2018a)	Features: Phase and Magnitude spectrogram, Classifier: CRNN with two output branches: SED and DOA estimation	0.72	36.9	21.5	53.0	A combination of CNN and RNN is used in this work to perform SELD. This system is a baseline system for SELD task given in the DCASE 2020 SELD challenge.
2	Audio event detection and localization with multitask regression network (Phan et al. 2020)	Features: Log-Mel magnitude spectrogram and intensity vectors, Classifier: CRNN architecture coupled with self-attention mechanism	0.59	50.8	18.2	64.1	A multi-regression network is proposed where MSE loss function is used homogeneously throughout the network for both SED and DOA estimation. This technique improved the performance from the baseline, but the performance is less as compared to the proposed system.
3	Event-independent network for polyphonic sound event localization and detection (Cao et al. 2021)	Features: Spectrogram, Classifier: Event-dependent network with three output branches: SED, DOA predictions, and event activity detection (EAD)	0.47	61.5	16.7	75.4	A frame-level permutation invariant technique is used for obtaining SED and DOA predictions. To get the on-set and off-set time of events, feature embedding information is encompassed during event activity detection. This helped the network achieve better predictions.
4	Polyphonic sound event localization and detection using channel-wise FusionNet Spoorthy and Kooolagudi (2024)	Features: Log-Mel magnitude spectrogram and intensity vectors, Classifier: Proposed Channelwise FusionNet	**0.23**	**81.2**	**9.1**	**86.9**	A channel-wise feature learning by using separable convolutions significantly improved the performance of the SELD task.

The systems considered for the comparison are re-implemented on our hardware and outcomes are observed for fair comparison. An FOA multichannel audio input is used as a common dataset during implementation. The best performance achieved in the case of each performance metric is indicated in bold face. In the case of other SELD systems shown in Table 4.4, to improve the training performance, data augmentation techniques are used. However, while training the FusionNet, no such data augmentation is used. From the table, it many be observed that performance of the proposed system is significantly better in terms of error rate, F-score and also frame recall. We observed the higher value of DOA error as compared to the Deep Neural Network based SELD system (Wang et al. 2020).

4.2.4 Contributions and Limitations

In the first step, a channelwise separated multichannel input is fed to the separable convolution block. In the second step, two channels' feature maps are fused to form a new feature map and iteratively fed to a separable convolution block. This step is iterated till a single feature map is obtained. The layer that outputs a single feature map is termed as a "Fusion layer," which is the combination of feature maps of all the channels. The proposed method explores the channel-wise information of the input features. From the experimentation results, a significant performance improvement is observed. The proposed model resulted in better performance than the existing SELD systems and the baseline system with a minimized error rate and DOA error rates. In future, the light-weight convolution blocks may be explored to reduce the number of parameters and hence the computation requirement. A suitable trade-off may also have to be achieved between the complexity of the architecture and performance of the system.

4.3 SUMMARY

In this chapter, localization of overlapped and non-overlapped sound events is performed along with the detection of the sound events. Different deep learning architectures are introduced for SELD task. As compared to the baseline model, the proposed D-SELDNet, T-SELDNet, and also Channel-wise FusionNet outperform in

terms of ER, F-score, DOA error, and FR as well. The D-SELDNet is less computationally complex as compared to SELDNet. The Channelwise FusionNet extracts channelwise information along with spatial and temporal information from the given multichannel input feature. The performance of the proposed approach has displayed better performance as compared to the state-of-the-art SELD systems. In the next chapter, methods proposed to classify acoustic scenes in the case of mismatched recording devices have been discussed in detail.

CHAPTER 5

DEVICE INDEPENDENT ACOUSTIC SCENE CLASSIFICATION

Acoustic Scene Classification (ASC) is a task of assigning a semantic label to an audio based on the surrouding it is recorded in. In this chapter, new methods that are proposed for classifying various acoustic scenes are discussed in detail. In the existing literature, the methods proposed for ASC considered the audio data that are recorded only from a single device. However, the recording device plays some role in the performance of the ASC system when there are audio recordings from different recording devices. Therefore in this chapter, different features and classifiers are introduced to enhance the ASC performance in the case of single and mismatched recording devices. Three methods are proposed to perform ASC in this chapter, namely, Deep Fisher Network, Bi-level Deep Learning model, and Device robust ASC system.

5.1 ACOUSTIC SCENE CLASSIFICATION USING DEEP FISHER NETWORK

Acoustic Scene Classification (ASC) is referred to as the task of identifying the environment in which the scene has been recorded. The scenes comprise the categories such as indoor (residence, restaurant/cafe), outdoor (park, metro station), and transportation (metro, bus, tram) (Barchiesi et al. 2015). The primary objective of this work is to create a system that can categorise different acoustic scenes based on auditory cues found in the input recording. In the existing literature, more focus has been given to the classifier architecture and less effort has been made to investigate the

features that can help the ASC system perform better. Therefore in this work, our aim is to propose a new model that learns better features from the input for classifying the scenes recorded using different devices as well as making use of the deep learning architectures.

The use of Fisher Network in this work to perform ASC is two-fold: Firstly, the feature learning from the input features like spectrogram or Gammatone Time Cepstral Coefficients (GTCCs) are not sufficient to capture the event information in an audio scene recording. The Fisher layer in Fisher network captures the smallest event information as well. Secondly, the Fisher Network function is similar to a feed-forward neural network and also does not require more space or time for computation, as the Fisher vectors computed in the Fisher layers of the network are 1-dimensional arrays and not 2-dimensional matrices.

Feature hierarchy in Fisher network is modeled by feeding an output of a layer as an input to the succeeding layer. Fisher vector encoding method is used to map the frame-wise features into a vector representation, named as the Fisher Vector (FV). Previously, the Fisher vector encoding scheme has been used for performing tasks such as Acoustic Event Classification (AEC) (Mulimani and Koolagudi 2019a), speaker verification (Wan and Renals 2005), and speech recognition (Smith and Gales 2002). In this work, the features such as GTCCs and Mel-band energies are extracted from an audio signal and encoded into an FV. These representations are the most common and conventional representation of an audio recording (Mesaros et al. 2017), (Lehner et al. 2019). The stacking of Fisher layers forms a network, named as "Deep Fisher network". From the literature, an observation is made that majority works are on CNNs, RNNs or a combination of these two architectures (Mesaros et al. 2017), (Lehner et al. 2019), (Zeinali et al. 2019), (NaranjoAlcazar et al. 2020), (Jung et al. 2020). In this work, we aim to explore the conventional audio features with an integrated Fisher encoding scheme using the current neural network architectures. The audio features used are GTCCs and Mel-spectrograms. Fisher vector encoding is a way of transforming a set of low-level framewise features extracted from an audio into a high-level utterance/audio clip level representation. These representations are

considered as intermediate representations that are built using a universal codebook or a dictionary, where variable sized features are transformed into a fixed size feature representation. The universal codebook is learnt in an unsupervised learning, making it more generalized and class independent. This property of Fisher vector encoding scheme makes it suitable later in a supervised task like ASC. Prominent features from all Fisher vectors are considered as the input for the classification of acoustic scenes.

5.1.1 Preliminaries of Fisher Vector Encoding

Fisher vector encoding is applied to spectral features such as GTCCs and Mel-spectrograms. Finally, single FV is generated from an audio signal. For example, an audio signal of 10 seconds, generates a Mel-spectrogram of size 500 x 64, where, window length is 40 ms with 50% overlap and feature dimension is 64. Based on the length of the audio, the number of frames generated in Mel-spectrogram may vary. Due to this variation in the dimensions of feature representations, FVs need to be encoded to a fixed dimensional feature vector. For instance, FV encoded with 4 Guassian components and 64 feature descriptors results in a vector of dimension 1 x 512 (2 x 4 x 64), as each Gaussian is represented by 2 vector parameters namely, mean and standard deviation for a 64 dimensional feature vector.

FVs are derived from Fisher kernel (Simonyan et al. 2013). Consider a feature matrix $X = x_{m,n}, m = 1, 2, ..M, n = 1, 2, ..., N$ be the matrix of M row vectors and N local descriptors. FV generation is a two-step process. The first step is a generative model of local descriptors. The second step is Fisher coding vector computation using likelihood gradients of local descriptors considering the parameters of the model. The Fisher vector encoding is based on a codebook which is also termed as a “dictionary” or “visual vocabulary”. The contents of the dictionary/visual vocabulary have the means and covariances of various acoustic scene classes. The dictionary is computed using local descriptors that are extracted from each audio signal. A clustering is performed on these local descriptors using a Gaussian Mixture Model (GMM) model. The dictionary is defined by the totality of cluster centers (also termed as codewords or visual words) in the codebook (Perronnin et al. 2010). For the proposed method, the

generative model considered is GMM. Fisher vector encoding is based on the mean and covariance deviation vectors generated for each Gaussian component k of the GMM and the dimension of the local descriptors frame-wise. The concatenation of the mean and covariance vectors results in a vector of length 2 x k x D, where k denotes the number of Gaussian components and D denotes the dimension of local descriptors.

From each class, out of several audio samples, randomly five audio samples are chosen for codebook generation as that is sufficient to train the GMM. GMMs with varying number of Gaussian components are applied for the preparation of dictionary and the one with 4 components is finally considered as it has exhibited better performance. The parameters considered for training the GMMs are shown in Equation (5.1) as λ.

$$\lambda = \left\{w_j, \mu_j, \sigma_j\right\}_{j=1}^{K} \tag{5.1}$$

where, w represents weight, μ represents mean vector and σ represents covariance matrix for a Gaussian j. Diagonal covariance values of GMMs capture the average first and second-order differences between the feature vectors and the centers of each GMM. The differences are computed as shown in Equations (5.2) and (5.3) (Simonyan et al. 2013).

$$\phi_j^{(1)} = \frac{1}{N\sqrt{w_j}} \sum_{p=1}^{N} \alpha_j(x_p) \left(\frac{x_p - \mu_j}{\sigma_j}\right) \tag{5.2}$$

$$\phi_j^{(2)} = \frac{1}{N\sqrt{2w_j}} \sum_{p=1}^{N} \alpha_j(x_p) \left(\frac{(x_p - \mu_j)^2}{\sigma_j^2} - 1\right) \tag{5.3}$$

where, $\alpha_j(x_p)$ is the soft assignment weight of the p^{th} feature x_p to the j^{th} Gaussian. The Fisher vectors are obtained by the concatenation of the different values as $\phi = \left[\phi_1^{(1)}, \phi_1^{(2)}, ..., \phi_K^{(1)}, \phi_K^{(2)}\right]$.

5.1.2 Deep Fisher Network for Acoustic Scene Classification

A two-layer architecture of the proposed deep Fisher network is shown in Figure 5.1. Fisher layer is divided into three sub-layers. The first sub-layer computes the Fisher vector from the input features extracted from an audio recording. This process is also known as Fisher vector encoding. This layer can be compared to the convolution layer of the conventional CNN, where additive and multiplicative operations are performed on the input features. The second sub-layer is computation of temporal pyramids from the encoded outcomes of the first layer, to capture the temporal information available in the input audio.

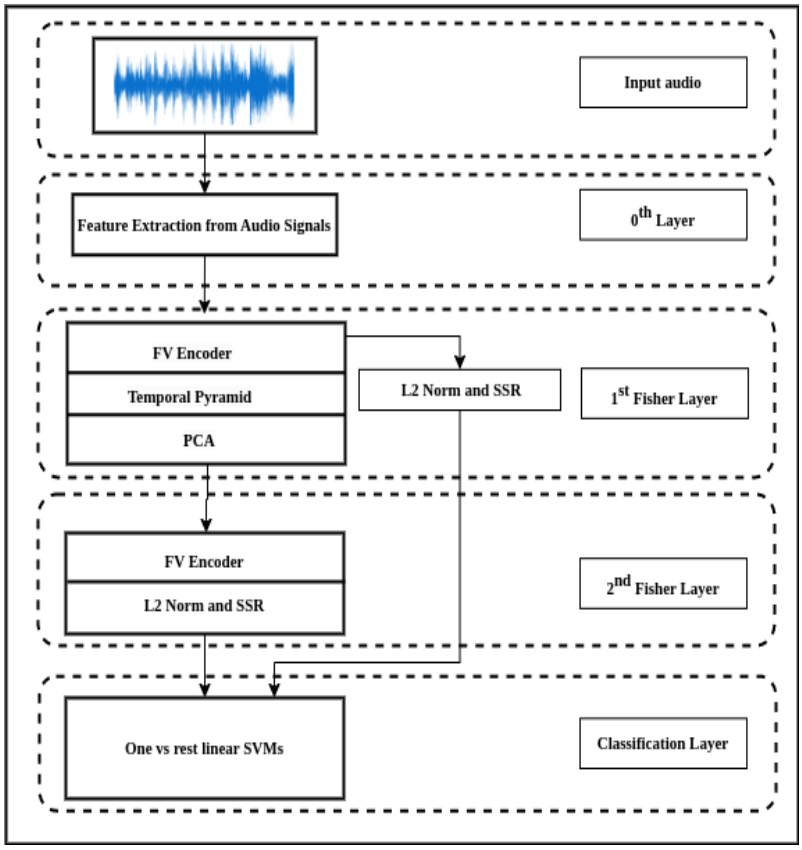

Figure 5.1: Block diagram of proposed Deep Fisher Network

Temporal pyramid is the technique used for processing temporal events by forming declarative representations of the complete sequences. A fixed-dimensional vector is returned as an output. The resultant temporal pyramids are sparse in nature, therefore, to reduce the dimension, a feature reduction technique needs to be applied on each vector. The third sub-layer in the network is L2 Normalization and Principal

Component Analysis (PCA). L2 normalization is the sum of squared vector values. For a vector v, L2 normalization is computed as $L2(v) = \sqrt{\sum v^2}$. For the faster convergence of the model, the normalisation of the values is computed to keep the values smaller. PCA is an unsupervised, non-parametric dimension reduction technique. Feature vectors derived from the frames of audio samples are appended to form a matrix, which is given as an input to PCA for dimension reduction. High dimension of the features usually overfits the model and makes the model less generic. PCA makes the data more diversified by rotating the axes. This step is similar to the "Local Contrast Normalisation" of CNN (Krizhevsky et al. 2012). Finally, the reduced dimensional vectors are classified using the 'one versus rest' Support Vector Machine (SVM). The detailed explanation of each sub-layer is provided in the following sub sections.

A. Fisher Layer

The sequence of operations carried out in each Fisher layer is shown in Figure 5.2. A feed-forward feature transformation is performed in three different sub-layers. Initially, a dictionary is built using GMM, a generative model for each class by considering a small training set of 5-6 audio samples.

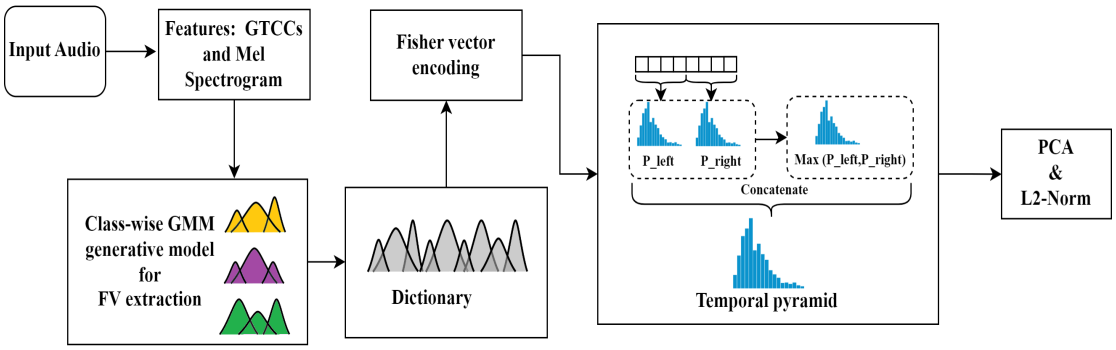

Figure 5.2: Sequence of operations happening in a Fisher layer

Dictionary of code words is a collection of means, covariances, and priors from each Gaussian of a GMM. Using this dictionary, the Fisher vectors are encoded into a single vector, i.e, for one audio recording, one FV is generated. The detailed computation of the FVs using the dictionary is presented in Section 5.1.1. Some of the encoding techniques, such as bag-of-words or sparse coding representations employed

in (Agarwal and Triggs 2006), (Coates et al. 2011), require large sized codebooks to produce discriminative feature representations. This, in turn, makes it challenging to use these techniques on large datasets. On the other hand, FV encoders do not require large codebooks. By adapting dimensionality reduction, even after the projection of FVs into a low-dimensional space, FV's discriminative ability can be preserved (Gordoa et al. 2012).

B. Temporal Pyramid

The second sub-layer is the temporal pyramid. Temporal information in the audio recordings is an important cue for ASC. Once the Fisher vectors are extracted, they are in an order-less representation. Several approaches such as pyramids (Lazebnik et al. 2006), spatial pyramids (Grzeszick et al. 2013), and temporal pyramids (Plinge et al. 2014) are proposed in the literature for time based ordering of these vectors. The pyramid scheme has already been applied for the AED task as well (Plinge et al. 2014). A similar pyramid scheme is applied in our proposed method as a pooling operation in the network in order to extract the order.

The temporal pyramid strategy may be generally implemented in two levels. At the coarse level, the entire Fisher vector is treated as one pooling segment. On the other hand in the finer level, the Fisher vector is hierarchically divided into multiple segments or tiles and pooling is performed on each segment. By concatenating the pooling results from all segments, a complete audio clip-level representation is obtained.

In the proposed method, an FV is sub-divided in a temporal manner using Equations 5.4 and 5.5 as shown in Figure 5.2.

$$p_{left} = \phi_1 = \frac{2}{T} \sum_{t=1}^{\frac{T}{2}} \phi \tag{5.4}$$

$$p_{right} = \phi_2 = \frac{2}{T} \sum_{t=\frac{T}{2}+1}^{T} \phi \tag{5.5}$$

where ϕ_1 and ϕ_2 are defined as the first and second temporal halves/tiles respectively and ϕ represents the values present in the corresponding temporal half. A sub-histogram

is computed for the corresponding values of t in each temporal half.

A max-pooling step is applied to these two temporal halves to obtain a histogram from P_{left} and P_{right} sub-histograms for the whole window using Equation (5.6).

$$\phi_3 = max\left\{\phi_1, \phi_2\right\} \tag{5.6}$$

Final step is the concatenation of sub-histograms of the two temporal halves and the histogram obtained from the max-pooling step into a single feature vector ϕ_{tp} known as term vector that represents complete features as shown in Equation (5.7). The concatenation is performed to obtain the information of the complete window.

$$\phi_{tp} = (\phi_1, \phi_2, \phi_3) \tag{5.7}$$

C. Normalization and Projection

The third sub-layer does the PCA and Normalization. Before passing these features to the FV encoder in the next level, a dimensionality reduction technique PCA is applied. Features are decorrelated in PCA, so that they can be efficiently modelled using diagonal-covariance GMMs of the next layer in the Fisher network. As usual, the use of PCA helps in reducing the computational time. There exists no specific practice of efficiently choosing the 'percentage of cumulative variance'. Maximum variation from the M-dimensional feature vector is retained when the value is set to 99% (Jolliffe, 1986). To improve the invariability properties of features in the vectors, the output obtained from the PCA are L2-normalized (Perronnin et al., 2010). FV encoder of the last layer generates sparse values which negatively impact the dot product of SVM. Therefore, a power normalization technique is applied to avoid biasing and to eliminate the sparseness of the Fisher vectors. This technique involves Signed Square Root (SSR) which is computed as $V = sign(V)\sqrt{|V|}$.

5.1.3 Performance Evaluation

To evaluate the proposed ASC system, a series of experiments were run on the four datasets with varying settings. The experiments are performed on TAU Urban Acoustic Scenes 2019, TAU Urban Acoustic Scenes 2019 Mobile, and TAU Urban Acoustic Scenes 2020 Mobile development datasets (Mesaros et al. 2018). The experimental setup, and results obtained from the deep Fisher network are discussed in the following sections:

A. Experimental Setup

The scene audio recordings are represented using two different feature representations, namely, Mel-spectrogram and GTCCs. The window length is set to 40 milliseconds, and the hop length is set to 20 milliseconds for both features. For Mel-spectrogram, overlapped "hamming asymmetric" window function and 64 number of Mel bands are used. Mel-spectrogram results in 500 x 64 dimensional feature matrix for an audio signal of 10 seconds. This standard setting is used in the baseline system of TAU Urban Acoustic Scenes 2019 challenge (Mesaros et al. 2018). For GTCCs, 13 cepstral coefficients along with velocity and acceleration coefficients are extracted per frame. GTCCs result in a 500 x 42 dimensional feature matrix for an audio signal of 10 seconds. There are 13 coeffients, 13 velocity coefficients, 13 acceleration coefficients along with their log energy components, resulting in a 42 dimensional feature matrix.

FVs are extracted from spectral features of audio recordings. Each feature representation results in one vector. The size of the vector depends on the number of Gaussians and the size of framewise feature descriptors in each feature representation. The Fisher vector encoding was evaluated for various Gaussian components like 2, 4, 8, 16, 64, and 128. The best performance of the Fisher network was observed at 4 Gaussian components. The vector size for GTCC with 42 feature descriptors and 4 Gaussian components is 1×336 ($2\times K\times d$, where K=4, d=42). Similarly, for Mel-spectrogram with 64 feature descriptors the size of the vector is 1×512 (K=4, d=64). FV of Mel-spectrogram and GTCC are horizantally concatenated. The

combination of these FVs were observed to provide better results when fed to SVM classifier. Final ASC is performed using a one-versus-rest SVM classifier. The input to the SVM classifier is the output obtained from Fisher layer. The training and testing datasets for the classifier are the standard splits provided in the TAU Urban Acoustic Scenes 2019 development dataset. FVs are labeled with the scene labels provided in the dataset. The kernels used in the SVM are Linear Kernel (LK) and histogram intersection kernel (HIK). Five-fold cross validation is used and the results are averaged. During testing the model, probability estimates are considered for the classification.

B. Results and Discussion

Average accuracies achieved in the case of two chosen kernels are given in Tables 5.1, 5.2, and 5.3 for TAU Urban Acoustic Scenes 2019, TAU Urban Acoustic Scenes 2019 Mobile and TAU Urban Acoustic Scenes 2020 Mobile development datasets, respectively. The proposed Fisher network is implemented with different number of layers and feature representations. The features obtained from layers of Fisher network are combined. The combination of features is obtained by concatenating outputs of two layers to form a row vector.

Table 5.1: Comparison of Layer-wise ASC accuracies (in %) of different features and their combination on TAU Urban Acoustic Scenes 2019 dataset (Legend: ACC-Accuracy, PRE-Precision, REC-Recall, F1-F1 score, LK-Linear Kernel, HIK-Histogram Intersection Kernel)

Layer Details	**Features**	**LK**				**HIK**			
		ACC	**PRE**	**REC**	**F1**	**ACC**	**PRE**	**REC**	**F1**
Fisher Layer1	GTCC	61	64	70	67	72	73	76	74
Fisher Layer2		70	73	70	71	75	74	76	75
Fisher Layer1 + Fisher Layer2		77	79	77	78	77	81	78	79
Fisher Layer1	Mel-spec	56	59	57	58	63	76	78	77
Fisher Layer2		69	70	72	71	80	81	82	81
Fisher Layer1 + Fisher Layer2		70	73	70	71	84	82	85	83
Fisher Layer1	GTCC + Mel-spec	81	80	83	81	86	89	88	88
Fisher Layer2		82	84	87	85	88	89	91	90
Fisher Layer1 + Fisher Layer2		**83**	**86**	**89**	**87**	**92**	**93**	**92**	**93**

Table 5.2: Comparison of Layer-wise ASC accuracies (in %) of different features and their combination on TAU Urban Acoustic Scenes 2019 Mobile dataset (Legend: ACC-Accuracy, PRE-Precision, REC-Recall, F1-F1 score, LK-Linear Kernel, HIK-Histogram Intersection Kernel)

Layer Details	**Features**	**LK**				**HIK**			
		ACC	**PRE**	**REC**	**F1**	**ACC**	**PRE**	**REC**	**F1**
Fisher Layer1 (device A)	GTCC	61	64	61	62	66	69	67	68
Fisher Layer2 (device A)		68	69	71	70	64	68	66	67
Fisher Layer1 + Fisher Layer2 (device A)		75	73	74	73	75	77	78	77
Fisher Layer1 (devices B & C)		53	54	57	55	55	58	56	57
Fisher Layer2 (devices B & C)		55	59	57	58	61	64	65	64
Fisher Layer1 + Fisher Layer2 (devices B & C)		58	59	61	60	59	59	61	60
Fisher Layer1 (device A)	Mel-spec	55	58	56	57	57	55	56	55
Fisher Layer2 (device A)		66	68	70	69	69	70	71	70
Fisher Layer1 + Fisher Layer2 (device A)		71	72	73	72	71	74	73	73
Fisher Layer1 (devices B & C)		54	56	59	57	57	55	57	56
Fisher Layer2 (devices B & C)		55	58	56	57	64	63	66	64
Fisher Layer1 + Fisher Layer2 (devices B & C)		56	57	56	56	63	66	64	65
Fisher Layer1 (device A)	GTCC + Mel-spec	79	81	84	82	80	79	81	80
Fisher Layer2 (device A)		81	84	82	83	84	88	83	85
Fisher Layer1 + Fisher Layer2 (device A)		**89**	**90**	**91**	**90**	**91**	**88**	**93**	**90**
Fisher Layer1 (devices B & C)		58	61	60	60	61	62	64	63
Fisher Layer2 (devices B & C)		65	67	68	67	66	68	69	68
Fisher Layer1 + Fisher Layer2 (devices B & C)		**67**	**69**	**68**	**68**	**71**	**72**	**69**	**70**

The results obtained from Mel-spectrogram (Mel-Spec) feature representation are comparatively better as compared to the GTCC feature representation considering the case of HIK. Fisher vectors computed for Mel-spectrograms consist of 64 feature descriptors. In comparison, GTCCs contain only 39. The Fisher vectors obtained from

Table 5.3: Comparison of Layer-wise ASC accuracies (in %) of different features and their combination on TAU Urban Acoustic Scenes 2020 Mobile development dataset (Legend: ACC-Accuracy, PRE-Precision, REC-Recall, F1-F1 score, LK-Linear Kernel, HIK-Histogram Intersection Kernel)

Layer Details	**Features**	**LK**				**HIK**			
		ACC	**PRE**	**REC**	**F1**	**ACC**	**PRE**	**REC**	**F1**
Fisher Layer1 (device A)	GTCC	66	64	66	65	67	68	67	67
Fisher Layer2 (device A)		69	69	71	70	70	68	66	67
Fisher Layer1 + Fisher Layer2 (device A)		69	73	69	71	76	78	79	78
Fisher Layer1 (devices B & C)		53	54	57	55	54	58	56	57
Fisher Layer2 (devices B & C)		55	51	54	52	61	62	64	63
Fisher Layer1 + Fisher Layer2 (devices B & C)		58	59	61	60	61	59	61	60
Fisher Layer1 (devices S1, S2 & S3)		60	58	59	58	68	67	66	66
Fisher Layer2 (devices S1, S2 & S3)		68	68	70	69	71	70	71	70
Fisher Layer1 + Fisher Layer2 (devices S1, S2 & S3)		70	72	73	72	74	74	77	75
Fisher Layer1 (device A)	Mel-spec	64	61	64	62	65	68	65	66
Fisher Layer2 (device A)		67	66	64	65	69	70	72	71
Fisher Layer1 + Fisher Layer2 (device A)		71	70	72	71	75	77	76	76
Fisher Layer1 (devices B & C)		56	55	58	56	58	59	60	59
Fisher Layer2 (devices B & C)		56	57	58	57	58	60	61	60
Fisher Layer1 + Fisher Layer2 (devices B & C)		58	60	57	58	65	63	66	64
Fisher Layer1 (devices S1, S2 & S3)		69	70	72	71	72	71	69	70
Fisher Layer2 (devices S1, S2 & S3)		70	68	71	69	71	68	72	70
Fisher Layer1 + Fisher Layer2 (devices S1, S2 & S3)		76	75	74	74	78	79	81	80
Fisher Layer1 (device A)	GTCC + Mel-spec	69	70	71	70	70	68	73	70
Fisher Layer2 (device A)		70	66	69	67	72	74	73	73
Fisher Layer1 + Fisher Layer2 (device A)		**85**	**87**	**84**	**85**	**89**	**90**	**88**	**89**
Fisher Layer1 (devices B & C)		58	55	59	57	60	63	66	64
Fisher Layer2 (devices B & C)		60	62	65	63	62	65	66	65
Fisher Layer1 + Fisher Layer2 (devices B & C)		**68**	**69**	**71**	**70**	**73**	**71**	**74**	**72**
Fisher Layer1 (devices S1, S2 & S3)		77	74	76	75	79	80	81	80
Fisher Layer2 (devices S1, S2 & S3)		79	80	79	79	83	80	78	79
Fisher Layer1 + Fisher Layer2 (devices S1, S2 & S3)		**82**	**83**	**81**	**82**	**85**	**87**	**86**	**86**

mel-spectrograms contain significantly more information as compared to those from GTCCs. The best recognition accuracy is obtained from the combination of these two spectral features.

Table 5.4: Comparison of performance of the proposed Fisher network and other state-of-the-art ASC systems for different datasets (Legend: ACC - Accuracy (in %)

TAU ASC 2018	**ACC**	**TAU ASC Mobile 2018**	**ACC**	**TAU ASC 2019**	**ACC**	**TAU ASC Mobile 2019**	**ACC**
Li et al. (2018)	72.9	Mesaros et al. (2018)	45.6	Salvati et al. (2019)	69.7	Waldekar and Saha (2019)	52.3
Jung et al. (2018)	73.5	Ren et al. (2018a)	58.3	McDonnell and Gao (2020)	82.3	Jiang and Shi (2019)	64.2
Golubkov and Lavrentyev (2018)	80.1	Nguyen and Pernkopf (2019)	66.1	Wang and Liu (2019)	86.9	Song and Yang (2019)	70.3
Xie et al. (2022)	81.2	Xie et al. (2022)	72.3	Xie et al. (2022)	82.2	Wang et al. (2019b)	60.6
Proposed Approach Spoorthy et al. (2023)	**93**		91		92		91

It is observed from the results that there is a considerable improvement in the performance of the system using combination of the layer outputs. It is obvious that in the case of deep networks, more complex information is learnt in the deeper layers. Table 5.4 presents the comparison of the proposed Fisher Network with the state-of-the-art ASC work for various datasets. DCASE 2019 Task 1(a) challenges' best results namely, (McDonnell and Gao 2020) and (Chen et al. 2019) are used for comparing our results. These two are the top entries in the DCASE 2019 Task 1(a) challenge submissions. The comparison is done with respect to features, classifiers and accuracy of the system. Both state-of-the-art systems used time frequency representations such as Mel-spectrograms, log-Mel spectrograms, and scalograms. In our proposed system, we have used Fisher feature vectors which are derived from the time frequency representations namely Mel-spectrograms and GTCCs. Both McDonnell and Gao (2020) and Chen et al. (2019) used deep learning based neural networks whose computation complexity is high compared to the one with 2 layered Fisher network proposed in our work. Compared to the accuracies reported by the

state-of-the-art works, our proposed approach has achieved an average accuracy of 92% which is an improvement of almost 7%. In comparison to all previously proposed ASC systems, the proposed deep Fisher network performed better. The major limitation of the previously proposed ASC systems was the unmanageable depth of networks. Majority of the systems were very deep in nature, which in turn are computationally expensive. The proposed Fisher network encodes the feature representation into a single vector, making the network learn the discriminative information of different acoustic scenes in a minimal amount of time and also being computationally less costly.

5.1.4 Contributions and Limitations

In this work, a deep Fisher network is proposed for ASC. Fisher vectors are encoded from different acoustic feature representations. Encoding of Fisher vectors from the features provides a large single dimensional feature representation with discriminative information. A Fisher layer in the proposed Fisher network consists of three sub-layers, FV encoder, temporal pyramid, and normalization with PCA. Last layer with PCA reduces the dimension of the feature vectors. The proposed Fisher network is built on the idea of CNN's feed-forward feature transmission principle. The output features of one layer are fed as an input to the next layer. The network is evaluated on five different ASC datasets with the highest recognition of 92% on the DCASE 2019 Task 1(a) ASC development dataset.

The proposed Fisher network resulted in better performance for the combination of spectral features. Fisher vector encoding ensures better discriminative information in acoustic scenes and can be used for classifying various acoustic scenes. Out of the sub-layers of Fisher network, more discriminative properties are captured from the outputs of the first and the second layers. In the proposed approach, only two Fisher layers have been explored. In the future, the layered architecture of the Fisher network can be varied, and multiple combinations of Fisher vectors can be tested for development of an ASC. The number of devices in the dataset considered for the proposed approach is only three. However, to get more robust ASC system, the dataset consisting of more number of recordings from different devices can be explored as well.

5.2 BI-LEVEL ACOUSTIC SCENE CLASSIFICATION USING LIGHTWEIGHT DEEP LEARNING MODEL

The task of assigning a label to an audio recording based on the surrounding is termed as Acoustic Scene Classification (ASC). In this work, the scenes are broadly categorized into outdoor (street traffic, public square), indoor (shopping mall, airport), and transportation (bus, metro) types. In this work, the classification of acoustic scenes is performed in two levels. The motivation behind the two-level classification is that the information learnt in the first level of classification is transferred to the second level of classification. This knowledge transfer will be beneficial finer level scene classification.

A bi-level light-weight deep learning model is proposed for ASC as shown in Figure 5.3. The system consists of two independent classifiers connected one after the other and the output of the total assembly is one among ten classes (DCASE Acoustic Scenes dataset consists of 10 acoustic scenes). In this setup, the first classifier (3-class classifier) performs the broader level classification of scenes into 3 broad classes. Then at the second level, with in each broadly classified group of scenes, finer classification into individual scenes is done (Simonyan et al., 2013). After the level-1 classification, the misclassified samples are used in retraining in the next iteration.

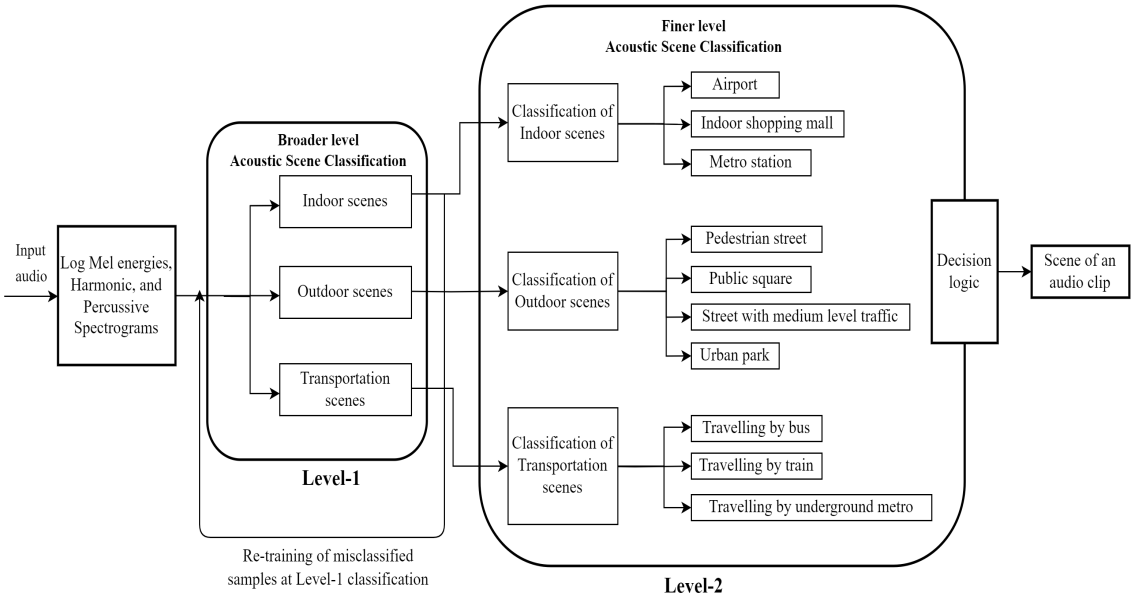

Figure 5.3: Block diagram of the proposed Bi-level ASC model

The misclassified samples are obtained by analysing the predicted output of the level-1 classifier. Decision logic based on the probability, decides the class of an acoustic scene. The input features used to train the deep models are log Mel band energies, harmonic and percussive spectrograms. Decomposition of the spectrograms is performed to capture the characteristics of different sound events.

The sound events occurring in a scene may consist of two different types: continuous and instantaneous. Continuous events are sounds that last longer and are monotonous, like rain, fan, or tap water running in the kitchen, etc. The instantaneous events are sounds that last for a very short time, like a clap or a gunshot. The continuous sound patterns are better captured in the harmonic spectrograms, whereas sudden, sharp events are captured in the percussive spectrograms. Harmonic Percussive Source Separation (HPSS) algorithm is expected to retain the frequent events in each acoustic scene and discards noisy data (Politis et al. 2020b). In the proposed work, we have combined two light-weight CNN architectures, MobileNetV2 and Squeeze-and-Excitation Net (SENet). The proposed SE-MobileNet, which is the combination of both models, incorporates the advantages of both.

5.2.1 Features for Bi-level Acoustic Scene Classification

In this work, three different features are extracted to train the lightweight CNN models. These are log-Mel band energies, harmonic and percussive spectrograms. The reasons for choosing these features is that log Mel-band energies are of the most common and popular TFR representations used for ASC task and the harmonic & percussive components of a spectrogram is expected to work well for sudden events present in an acoustic scene. A two-dimensional matrix representing the log mel band features is fed as an input to the deep learning models. To extract log-Mel band features, an STFT is initially applied to the audio signal with a window-length of 40 ms and 50% overlap. The windowing technique used is ‘Hamming’. The absolute values present in each bin are squared, and a Mel-scale filter bank of 128-bands is applied. The final step is to compute the logarithmic values of the obtained Mel energies.

A median-filtering approach is applied to the spectrograms and later decomposed into harmonic and percussive components. The length of the audio recordings is 10 seconds, and the resultant features' dimension is 500×128. Spectral feature representations using log Mel band energies, the harmonic and percussive spectrograms are shown for two acoustic scenes in Figure 5.4. In the figure, the encircled regions of the spectrogram highlight the distinguishable components of harmonic and percussive spectrograms of different scenes. By observing the circled portions, it can be seen that the information of the event is clearly visible in the harmonic and percussive spectrograms as compared to log Mel band energies.

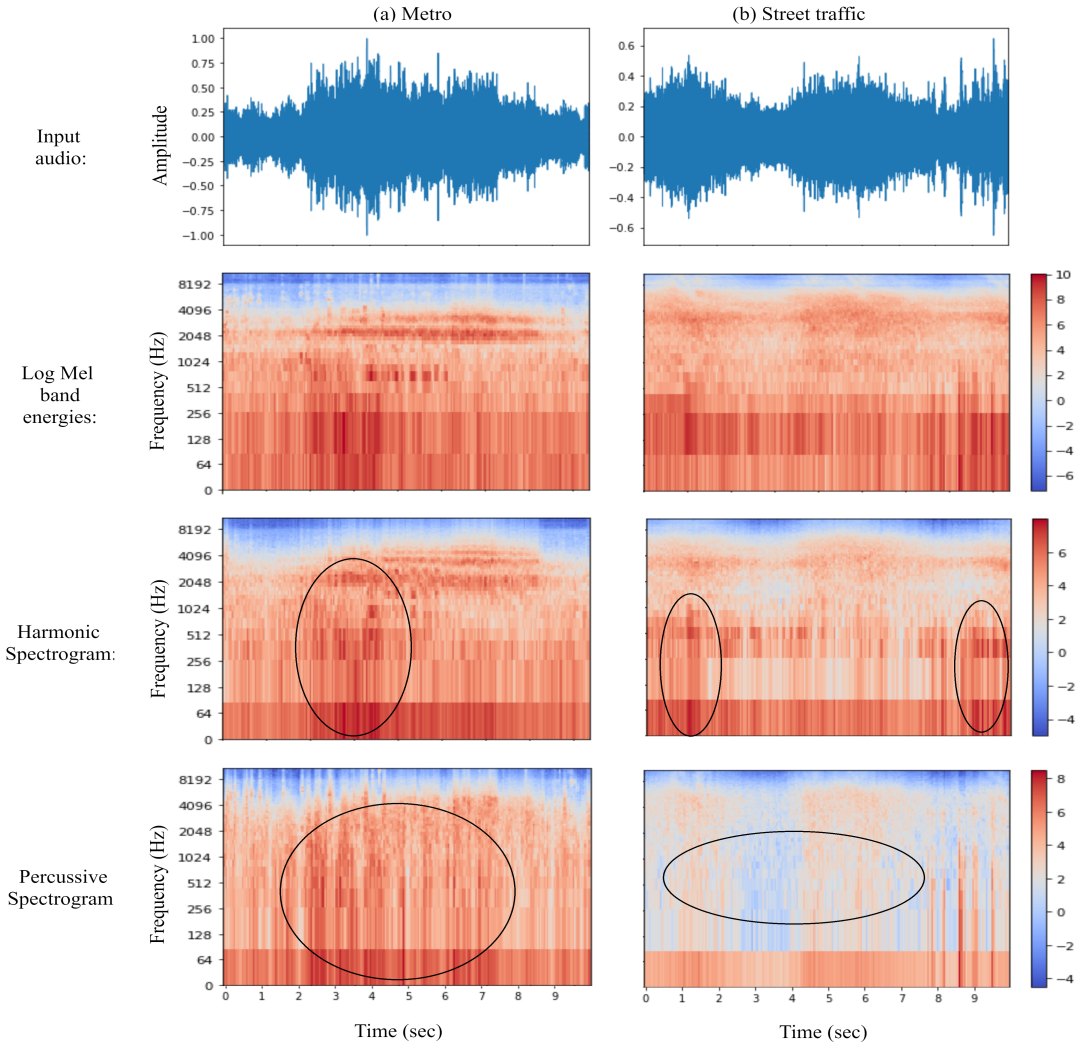

Figure 5.4: Comparison of different feature (spectral) representations of a scene

5.2.2 Bi-level Lightweight Deep Learning Classification Model

In this work, ASC is performed using a bi-level lightweight CNN model. The task is split into two phases. Primarily, in the first level, the acoustic scenes are categorized into three broad classes, namely, indoor, outdoor, and transportation classes. The audio clips are further classified into individual scenes in the second level. The proposed approach is a combination of two popular lightweight CNN models, namely, MobileNetV2 and SENet. The model details of the architecture are given in following sub-sections:

A. MobileNetV2

MobileNets are low-complex, small, low-power (i.e. does not require much parameters while training the model and can deployed on hand-held devices) deep learning models in which the parameters are tuned to work on lesser resources (Eronen et al. 2006). The basic building block for MobileNet architecture is a depth-separable convolution block. The main reason for using depthwise separable convolution over conventional convolution is to speed up the network and reduce the number of trainable network parameters. The illustration of the convolution and depthwise convolution operations is given in Figure 5.5.

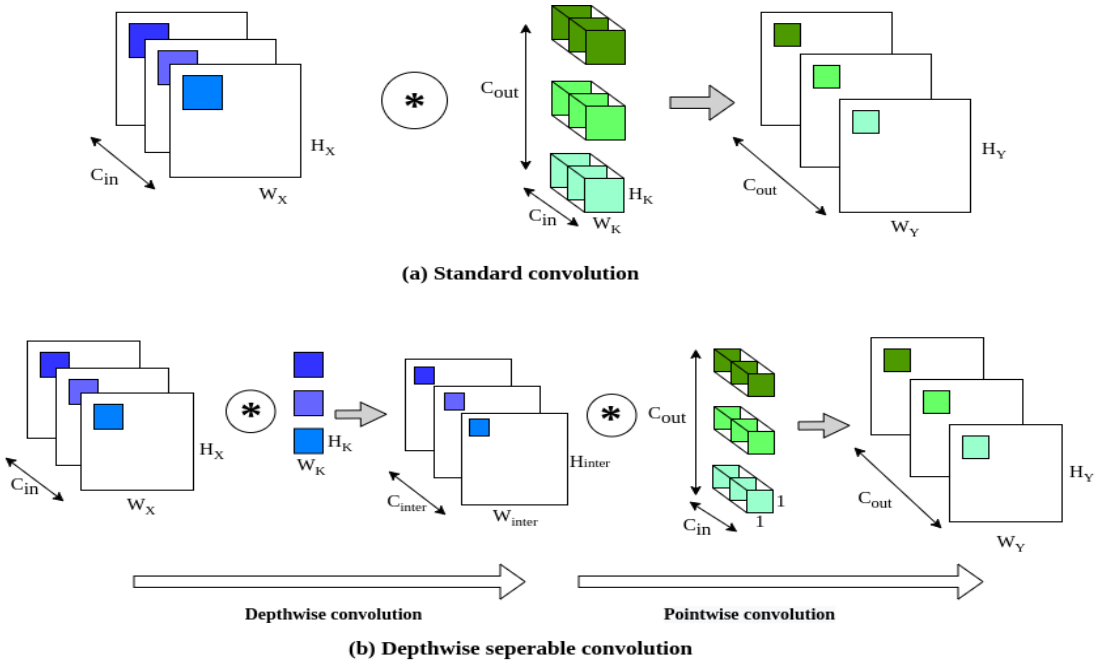

Figure 5.5: Illustration of convolution and depthwise convolution blocks (Roma et al. 2013)

Figure 5.5 (a) shows the standard convolution operation which takes input as a C_{in} number of channels. The output resulting from the convolution layer corresponds to the sum of convolution kernels. Consider the dimension of the input X as W_x, H_x, C_{in}, where, W_x represents width, H_x represents height, and C_{in} represents number of input channels. The dimension of the output Y is given by W_y, H_y, C_{out}, where W_y represents the width, H_y represents height, and C_{out} represents the number of output channels. For example, let us consider a two-dimensional image with dimension 12×12×3 pixels. For computation, let us do a 5×5 convolution on the image with stride of 1 and no padding. The convolution operation results in 8x8 pixel image which undergoes scalar multiplication of kernel 5×5 (25 pixels). The resulting output is of size (12-5+1 = 8) as there is no padding. As there are three channels in the input image, the convolutional kernels are calculated 3 times. Therefore, the total number of multiplications is 5×5×3=75 every time the kernel moves. If we want to increase the depth of the image, the number of kernels may be increased. For example, to get 128 channels, create 128, 8×8×1 images, and then stack them to get an image output of 8×8×128 dimension.

Figure 5.5 (b) illustrates the working principle of the depthwise convolution operation. In depthwise convolution, two operations are performed, namely, depthwise convolution and pointwise convolution. The first operation, depthwise convolution, results in the same output dimension as of convolution block, which is a form of group convolution. Here, the input feature map is $H_x \times W_x \times C_{in}$. The input feature is divided into N group channels, where one group consists of only one channel. The kernel size of the convolution is $H_k \times W_k$, which is the height and width of the kernel. The parameter reduction is attained by extracting spatial features from each channel separately, which reduces the computation cost of the operation. But, using this block may provide poor information flow between channels, and resultant output may not be related to each input channel. To overcome this issue, pointwise convolution is used. The pointwise convolution is a 1x1 convolution block, and it creates a linear combination of the output of the depthwise convolution. This block mixes information in between channels and solves the problem of poor information flow in the depthwise

convolution block. This block is mainly used to change the dimension of the input feature of output channels.

In depthwise separable convolution, the operation takes place in two steps: depthwise and pointwise convolution. In the first step, convolution is applied to input image without changing the depth of the image by using only 3 kernels of shape 5×5×1. Each channel of the image is iterated over 5×5×1 channel resulting in scalar products of every 25 pixel group, giving a 8×8×1 image. For three channels, these 8×8×1 images are stacked forming an image of size 8×8×3. The second step in pointwise convolution which uses 1×1 kernel or a kernel that iterates through every single point. In the standard convolution, we have transformed image of size 12×12×3 to 8×8×256. Whereas, in depthwise convolution, image is transformed to size 8×8×3. In this case, the depth of the input is 3. So, we iterate 1×1×3 kernel through 8×8×3 image which results in 8×8×1 image. Hence, we can create a depth of 128 by performing 1×1×128 on 8×8×1 image to get 8×8×128 image.

The resultant number of operations of convolution and depthwise convolution is as follows: There are 128, 5×5×3 kernels that move 8×8 times. Therefore, the number of multiplications is 128×5×5×3×8×8 = 6,14,400 for convolution. The depthwise convolution, we perform 3, 5×5×1 operations that move 8×8 times resulting in 3×5×5×8×8 = 4,800 multiplications. In pointwise convolution, there are 128 channels. Therefore, we have 128, 1×1×3 kernels that move 8×8 times. This results in 128×1×1×3×8×8=24,576 multiplications. By adding both together, we get 29,376 multiplications. Comparing number of multiplications of convolution and depthwise convolution, it can be observed that 29,376 multiplications are lot less than 614,400 multiplications. Hence, the network will consume less resources and also computation time is faster.

B. SENet

SENet is one of the popular lightweight CNN architectures that has seen high usage in recent times (Gordoa et al. 2012). The SENet architecture is given in Figure 5.6. The network consists of two blocks, namely, squeeze, and excitation. The main aim

of this network is to enhance the interdependency between each channel in the input without additional computational cost.

In the SENet, the initial step is to figure out a method to decompose each feature channel into a single numeric value. In this way, the decomposition would decrease the number of parameters and thus reducing the computational complexity. Then, the output resulting from this operation is a vector of size n, where n represents the number of channels in the convolution layer. The next step is to feed this vector to a neural network consisting of two layers, resulting in a vector of the same size. Finally, the values obtained from this operation can be used as weights on the original feature maps, which are scaled on the basis of the importance of each channel.

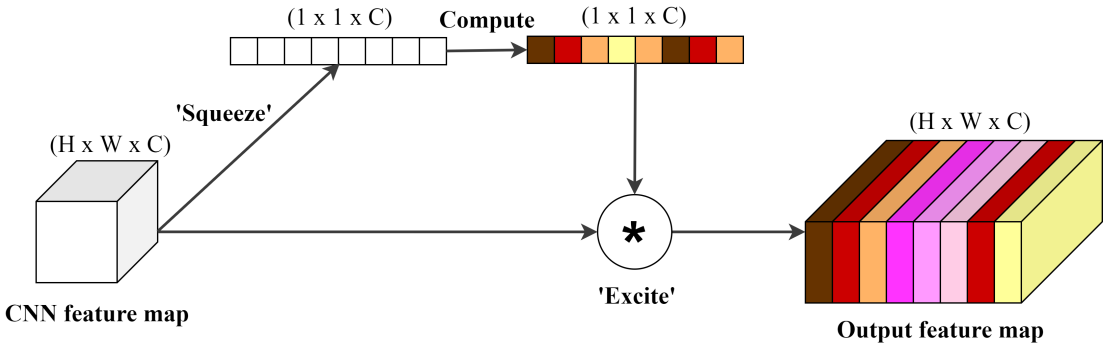

Figure 5.6: Working of SENet architecture (Gordoa et al. 2012)

The SE block mainly consists of 3 operations: squeeze, computation, and excitation. The first operation, squeeze, performs a global average pooling operation on the feature map resulting from a CNN layer. This operation takes the average value over the activation function in spatial dimension (HxW), and it results in one activation per channel. The second operation, computation, takes the vector from the previous squeeze operation through two dense/fully-connected layers. The reason for performing this operation is to capture the channel-wise dependencies present in the feature maps' spatial dimension. Two sets of activation functions are used in the dense layers.

C. Proposed SE-MobileNet

In this work, a combination of MobileNetV2 is used with SENet to classify acoustic scenes in two levels. MobileNet is a low-complexity CNN architecture that uses depthwise separable convolution layers instead of standard convolution layers. The architecture has displayed better processing in both small-scale, and large-scale problems (Jolliffe 1986; Perronnin et al. 2010). In addition to this, the amalgamation of MobileNet and SE blocks into a single network attributes to capture the merits of both networks. SE blocks make use of the channel information from the input and perform feature calibration channelwise dynamically. The combination of these two networks has been utilized for image classification (Zeinali et al. 2019). However, in this work, we stack this network in order to achieve a bi-level classification of acoustic scenes. The SE block is added to the tail of the MobileNetV2, where the final classification layer is discarded. The input features are fed to MobileNetV2 architecture, and high-level features [1] are extracted from the depthwise convolution layers of the architecture. The extracted features are fed to SE-block, followed by the global average pooling layer, fully connected layer, and softmax activation function.

The architecture of SE-MobileNet is shown in Figure 5.7. The input feature (Log Mel Energies/Harmonic/Percussive Spectrogram) is fed to the network, which is provided initially to the depthwise convolution layer followed by the pointwise convolution layer as shown in the MobileNetV2 model. The output of MobileNet architecture is then fed to the SE block's as in SENet, where, first layer is dense/fully connected layer, followed by squeeze and excitation operation. The SE block's output is given to a fully connected layer consisting of 1024 nodes and finally to the classification layer consisting of the number of nodes equivalent to the number of classes (level-1: 3 classes; level-2: 10 classes).

The SE-MobileNet classification model is used in two levels for the proposed ASC task. The first level of classification is to group the audio samples into three classes, indoor, outdoor, and transportation based on their broader level properties. To identify

[1] High-level features are the features that are extracted in the hidden layers of a neural network architecture

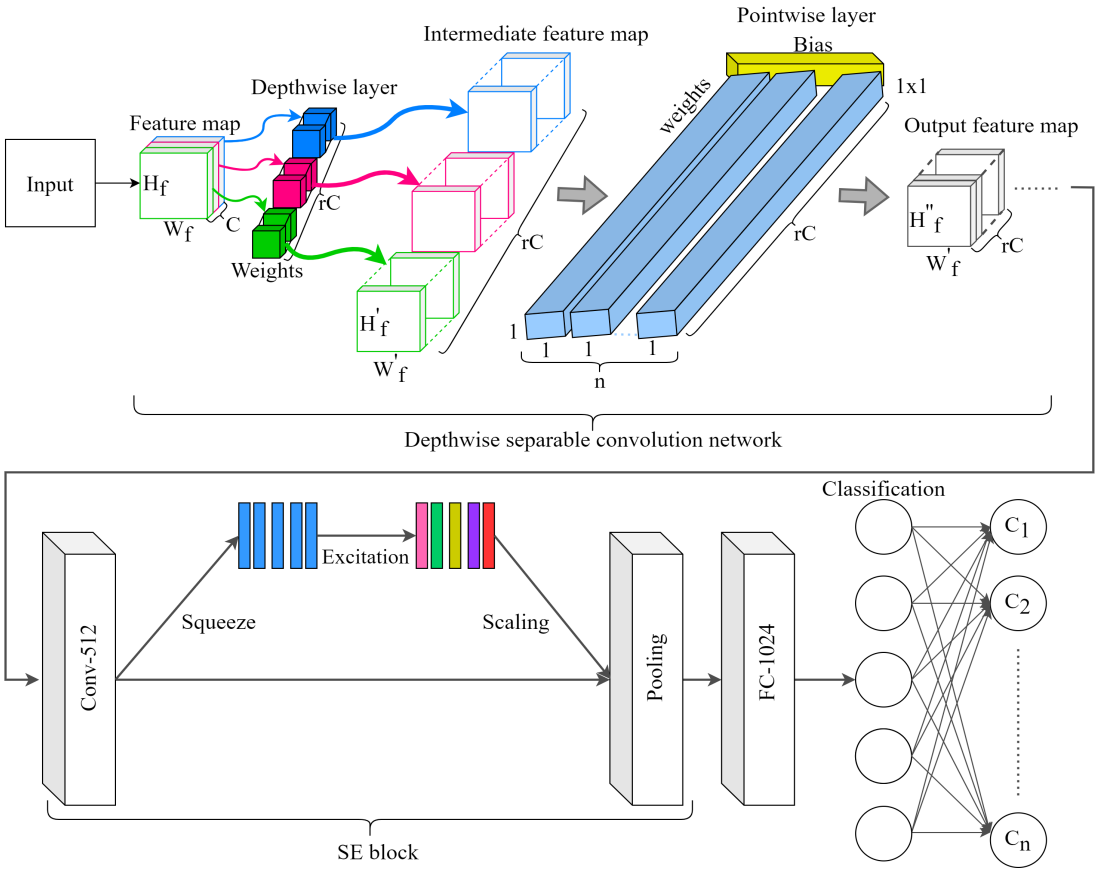

Figure 5.7: Proposed Architecture of SE-MobileNet

these classes, the input features, namely, log Mel-band energies, Harmonic & percussive spectrograms are fed separately to the SE-MobileNet, and the classes are identified. After obtaining the results at the first level, based on the predicted output, the correctly classified samples are passed to the second level of classification. The classes identified in the first level are further classified into ten scenes at the second level. The difference between the first level SE-MobileNet and the second level SE-MobileNet is the number of nodes in the final classification layer which is 3 and 10 nodes respectively.

5.2.3 Performance Evaluation

The following subsections contain information regarding the dataset, input features, and the configurations used in the deep learning architectures. The dataset used in the proposed ASC system is TAU Urban Acoustic Scenes 2020 Mobile (Heittola et al. 2020) one.

A. Experimental setup

The experiments in this work are performed using python packages Tensorflow2.0 and Keras. The network consists of different layers: depthwise convolution, batch normalization, average pooling, dense layer, and an activation layer as shown in Figure 5.8. Adam optimizer is used in the network (Kingma and Ba 2015). The optimizer's learning rate is set at 0.001, and after 10 epochs, if the validation accuracy does not improve, the rate is raised by a factor of 0.5. The loss function used is Focal loss (Qin et al. 2018). Focal loss can be considered as a variant of binary cross entropy loss function which is commonly used for multi-class classification problems. The reason for choosing this loss function is that for the misclassified samples, the function provides more emphasis and forces the system to classify the challenging samples (Soonshin Seo 2021). The focal loss function handles the misclassified samples by adding a hyper-parameter λ which intuitively decreases the penalty of easy samples by extending the range in which they receive low values of loss. If the $\lambda = 0$, then it is a standard cross entropy loss function. The focal loss can be computed as shown in Equation 5.8 (Qin et al. 2018).

$$L_{FL}(x) = -(1-x)^{\lambda} \, . \, log(x) \tag{5.8}$$

Where, L_{FL} is the focal loss and x is the input sample. The model is trained for a maximum of 200 epochs. However, the training ends if no improvement is found in the metric up to 50 epochs as an early stopping criterion.

The ASC system developed in this work is a setup of two independent classifiers used for classifying scenes into broader (3-class) and finer (10-class) level scenes. The broader level classification is from the prior knowledge that the audios can be roughly categorised into three classes, namely, indoor, outdoor, and transportation. The main classifier is the 10-class classifier. The input audio must belong to one of ten scenes, namely, airport, shopping mall, metro station, pedestrian street, public square, street traffic, tram, bus, metro, and park. The final class label is chosen by fusion of scores obtained by two classifiers (Simonyan et al. 2013). Let the two classifiers be $C1$ and

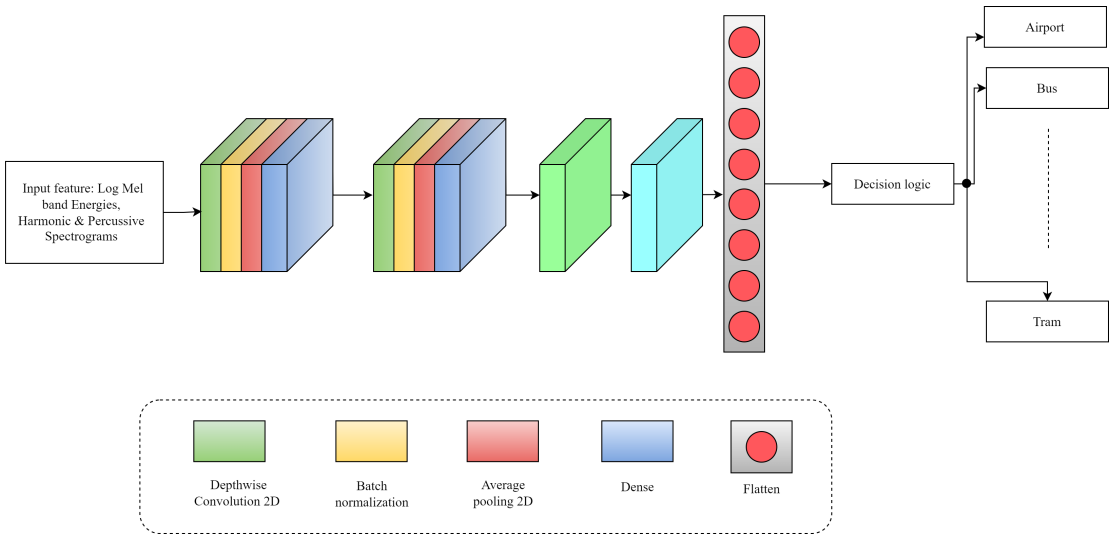

Figure 5.8: Architecture of proposed SE-MobileNet model

$C2$ for 3-class and 10-class respectively. The outputs of the two classifiers be (O_a^1) and (O_b^2) respectively. For a given audio input i, the final output scene is given by Equation 5.9.

$$Class(i) = \underset{b,(a \in C1, b \in C2, a \supset b)}{argmax} O_a^1(i) * O_b^2(i) \tag{5.9}$$

Where, $a \supset b$ denotes that a can be considered as a superset of b. For say, the scenes such as public square, shopping mall, and park can be considered as a subset of outdoor class. Therefore, the final class (I) is obtained by computing the product of the outdoor class (O_a^1) and the public square class (O_b^2).

The steps involved in the implementation of proposed bi-level ASC system are given below:

1. For the input audio, log mel band energies, harmonic and percussive spectrograms are extracted.

2. The input features, namely, log Mel-band energies, harmonic & percussive spectrograms are fed to classifier C1 for training the model at level-1 classification.

3. The correctly classified samples are provided to classifier C2 as input for training

the model for level-2 classification. The misclassified samples in C1 classified are used for re-training the C1. For this re-training, Focal loss function is used.

4. After the level-2 classification, the final class label of 10-class classifier is obtained using Equation 5.9.

B. Results and Discussion

In this section, the performance of the proposed ASC system is compared with the state-of-the-art CNN architectures and existing ASC systems. The proposed system is also analyzed for the number of trainable network parameters for the three architectures, MobileNet, SENet, and SE-MobileNet for bi-level classification of acoustic scenes. The performance resulted from the two levels are given in Table 5.5 for TAU Urban Acoustic Scenes 2020 Mobile dataset. The evaluation metrics considered for the evaluation of the system are macro-average accuracy (average of the class-wise accuracies) and multiclass cross-entropy (Log loss). The results are given for three different architectures and features. The results obtained for individual feature representations are given in Table 5.5.

The results show that the highest average accuracy is achieved for percussive spectrogram features trained using the SE-MobileNet network with the accuracies of 96.9% and 86.6% in level-1 and level-2 respectively using TAU Urban Acoustic

Table 5.5: Performance of the ASC system for level-1 and level-2 classification on MobileNet, SENet, and SE-MobileNet architectures using 3 different features on DCASE 2020 Development dataset (Legend: ACC- Accuracy)

Features	**Model**	**Level-1**		**Level-2**	
		Accuracy	**Log-Loss**	**Accuracy**	**Log-Loss**
Log-Mel band Energies	MobileNetV2	95.3	0.105	81.8	0.198
	SENet	94.1	0.117	81.5	0.390
	SE-MobileNet	96.8	0.103	82.4	0.269
Harmonic Spectrogram	MobileNetV2	92.4	0.138	81.2	0.368
	SENet	91.6	0.129	82.8	0.294
	SE-MobileNet	95.4	0.124	84.1	0.203
Percussive Spectrogram	MobileNetV2	93.6	0.221	84.6	0.224
	SENet	94.1	0.118	84.8	0.215
	SE-MobileNet	**96.9**	**0.101**	**86.6**	**0.211**

Scenes 2020 Mobile development dataset, respectively. The results achieved by the proposed system in level-1 classification are improved by 9.6% in terms of average accuracy and 0.336 in terms of loss value as compared to the baseline system developed using TAU Urban Acoustic Scenes 2020 Mobile dataset (Heittola et al. 2020). The reason behind the separation of the spectrogram into harmonic and percussive components is to analyze the behavior of the events in different scenes. The percussive component of the spectrogram captures the events that are sudden in nature. Generally, the acoustic scene recorded in a particular surrounding consists of two kinds of events, continuous events like rain, the humming of any electrical appliance, etc., and instantaneous events such as phone ring, dog bark, and the keys' clinking, etc. To label a scene, instantaneous events play a major role.

The performance of the proposed ASC system is compared with the current CNN-based architectures such as CNN, MobileNetV1, MobileNetV2, MobileNetV3, SqueezeNet, and SENet. The results obtained for these architectures for the level-1 and level-2 classifications are presented in Table 5.6. The main purpose of choosing these architectures for comparison is they are all popular widely accepted architectures, and except CNN and DenseNet, all other architectures are low-complexity models.

Table 5.6: Performance comparison of the proposed ASC with existing CNN architectures for level-1 and level-2 ASC (Legend: ACC- Accuracy, M-Millions)

Model	**Level-1 ACC (in %)**	**Level-2 ACC (in %)**	**# of parameters (in M)**
CNN (6-layer) (Heittola et al. 2020)	87.3	79.4	8.9
MobileNetV1 (Howard et al. 2017)	92.3	81.7	5.1
MobileNetV2 (Sandler et al. 2018)	93.6	84.6	4.8
MobileNetV3 (Howard et al. 2019)	96.1	83.9	4.9
SqueezeNet (Iandola et al. 2016)	89.8	82.3	5.4
SENet (Hu et al. 2020)	94.1	84.8	10.4
SE-MobileNet (Proposed)	**96.9**	**86.6**	**4.7**

The main aim of this experimentation is to show that good ASC performance is also possible using low-complex networks. The performance reported in Table 5.6 is in terms of average accuracies in the level-1 and level-2 classifications along with the number of trainable parameters. The CNN model is the baseline model provided by the DCASE challenge organizers for the low-complexity ASC task (Heittola et al. 2020). The results are obtained using log Mel band energies as features and deep learning models as classifiers. The performances resulting from the low-complexity models such as MobileNetV1, MobileNetV2, SENet, and SE-MobileNet are comparatively better than conventional CNN models. It is also observed that number (#) of parameters for the SE-MobileNet is the lowest as compared to other lightweight CNN models. The highest number of parameters are reported for 6-layer CNN model. The main aim of choosing low-complex CNNs is that they yield in better performance along with lesser computational complexity.

The performance of the proposed system is compared with the state-of-the-art, low-complexity ASC systems and are presented in Table 5.7. The dataset chosen is TAU Urban Acoustic Scenes 2020 Mobile. Here, the aim of this challenge was to develop low-complex/lightweight deep learning models to perform ASC. The scenes are broadly categorized into three classes. Therefore to have a fair comparison, only level-1 classification performance is considered. The metrics chosen for comparison are average accuracy and log loss. The table provides the methods for classifying acoustic scenes, the performance metrics, and the remarks on the work.

Table 5.7: Comparison of level-1 average accuracy of proposed approach with the baseline and state-of-the-art methods on the TAU Urban Acoustic Scenes 2020 Mobile dataset. (Legend: ACC-Accuracy)

Sl no.	Title	Method	ACC	Loss	Remarks
1.	Acoustic scene classification in DCASE 2020 Challenge: generalization across devices and low complexity solutions (Heittola et al. 2020)	Features: Log Mel band energies, Classifier: CNN	87.3%	0.437	The model is a baseline system provided by the DCASE challenge organizers. The CNN model is trained on log Mel-band energy features. The model resulted in higher misclassification for 3-class classification.
2.	CP-JKU Submissions to DCASE'20: Low-Complexity Cross-Device Acoustic Scene Classification with RF-Regularized CNNs (Koutini et al. 2020)	Features: Log Mel band energies, Classifier: Regularized CNN	96.8%	0.101	The method uses perceptually weighted log-Mel energies trained for Receptive Field (RF)-regulaized CNNs. This method resulted in rank 1 in terms of average accuracy in the TAU Urban Acoustic Scenes 2020 Mobile challenge. The use of weighted log-Mel energies resulted in better performance in terms of both accuracy and loss values.
3.	Low-Memory Convolutional Neural Networks for Acoustic Scene Classification (LopezMeyer et al. 2020)	Features: GCC-grams, Classifier: CNN	91.2%	0.677	The method resulted in good performance for Generalized Cross-Correlation (GCC) with phase transformation features fed as input to CNN. The method resulted in fewer parameters. However, the performance can still be improved.
4.	Lightweight Convolutional Neural Networks on Binaural Waveforms for Low Complexity Acoustic Scene Classification (Pajusco et al. 2020)	Features: Log Mel band energies, Classifier: 1D-CNN	91.2%	0.269	The method uses Mel filter banks trained on 1D CNNs to achieve a good performance. However, the audio signals need to be preprocessed using pruning techniques to achieve the reported results. This adds up as an additional step increasing the complexity of the ASC system.
5.	Bi-level Acoustic Scene Classification Using Lightweight Deep Learning Model (Spoorthy and Koolagudi 2023a)	Features: Percussive component of spectrogram, Classifier: **SE-MobileNet** (Proposed system)	**96.9%**	**0.101**	The proposed SE-MobileNet is combination of two low-complexity CNNs providing both better performance and less trainable parameters. The proposed model does not need any additional augmentation method for the audio files.

5.2.4 Contributions and Limitations

In this work, a lightweight CNN model is used to perform ASC. The features used for training the lightweight CNN networks are log Mel band energies, harmonic and percussive spectrograms. The main reason for decomposing the spectrogram into harmonic and percussive spectrogram is to perform the ASC based on sound events in a scene. The three deep learning networks, namely, MobileNet, SqueezeNet, and the proposed Fusion models were proposed in this work for ASC. All the models are lightweight CNN architectures.

The first observation that can be made from this work is that the decomposition of the spectrogram features resulted in the improvement of the performance of the ASC. The reason behind the improvement is that the percussive spectrograms capture the information about instantaneous events. The second observation of this work is the use of low-complexity or lightweight CNN models to perform an ASC. The use of lightweight CNN models resulted in less trainable parameters than the conventional deep CNN architectures. In the future, other time-frequency representations may be explored.

5.3 DEVICE ROBUST ACOUSTIC SCENE CLASSIFICATION USING ADAPTIVE NOISE REDUCTION AND CONVOLUTIONAL RECURRENT ATTENTION NEURAL NETWORK

In the recent trends, device robust ASC is more in focus. The most common challenge in the ASC is the mismatch in the recording devices. DCASE 2019 challenge organizers have released a dataset consisting of audio samples recorded from three different devices, and the focus is to draw attention to device invariant ASC without leveraging any device information. However, a huge data skewness is observed in the number of samples recorded with device A (main recording device), devices B and C. Also, another device D was included for the evaluation dataset, which made the problem similar to real-world conditions.

In this work, the device distortion information in the audio samples is minimized by preprocessing the samples before performing the feature extraction. Here, the audio

samples recorded by mobile devices consist of higher distortion, and also their quality is low. The device distortion is minimized by applying an adaptive noise reduction technique on these samples. Further, log-Mel band energies are extracted from these preprocessed audio samples. In order to reduce any more variability in the features, they are normalized using a zero mean and unit variance normalization technique which helps in getting rid of any biased values in the spectrogram. The next step is to perform ASC, where the normalized features are fed to the Light weight Convolutional Recurrent Attention Neural Network (LW-CRANN). In the LW-CRANN architecture, the number of operations is reduced by making the network less complex. In the CRANN architecture, the attention layer is introduced to the CRNN to enforce more weightage on misclassified samples.

5.3.1 Device Distortion Analysis

In the TAU Urban Acoustic Scenes 2019 Mobile (Mesaros et al. 2018) dataset, the acoustic scenes are recorded using high-quality recording devices, smartphones, and cameras. The audio signal consists of an original signal $s(n)$ which is convolved with an impulse response (IR) $h(n)$. The device distortion is present in the $h(n)$ of the signal (Song and Yang 2019). The presence of device distortion in the audio signal may be illustrated using spectral analysis of the signal in the frequency domain. The audio samples of an acoustic scene 'Tram' are chosen from the three different recording devices. The samples picked are recorded from the same location, 'Stockholm'. Three different spectral analysis characteristics are extracted from audio samples, namely, spectral flux, spectral rolloff, and spectral entropy. They are shown in Figure 5.9 for perception.

The spectral flux is the measure of the variability of the spectrogram over time (Scheirer and Slaney 1997). In Figure 5.9(a), it can be observed that the flux values of devices B and C have higher peaks than that of device A. The spectral rolloff points give the audio signal bandwidth by determining the frequency bin in which a certain percentage of the overall energy is present (Scheirer and Slaney 1997). The bandwidth of device B is higher as compared to devices A and C, as seen in Figure 5.9(b). The final measure used is spectral entropy which measures the peakiness of the spectrogram

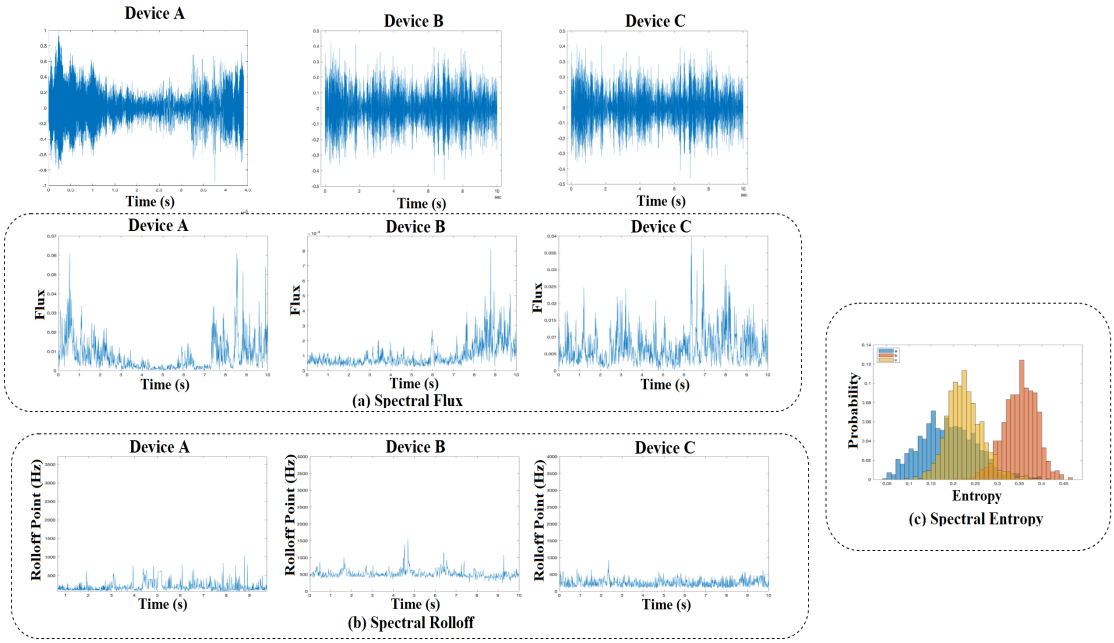

Figure 5.9: Spectral feature analysis for multiple device recordings of DCASE 2019 Task 1a dataset

or also known as the measure of disorder (Misra et al. 2004). From Figure 5.9(c), the entropy is very high for devices B and C. This measure gives a clear understanding of the distortion present in the audio signal, which is one of the significant factors affecting the performance of the ASC system.

5.3.2 Proposed Device Robust Acoustic Scene Classification Method

Block diagram of a device robust ASC system is shown in Figure 5.10. The audio recordings are from different recording devices, such as, high-quality recording devices, smartphones, cameras, etc. However, the recording quality can vary from device to device. In high-quality recording devices, noise cancellation techniques are built in, but that is not the case for smartphones or cameras. Hence, in these cases, distortion may be present in the audio samples affecting the system's performance. Therefore, the input audio samples recorded with different devices need to be preprocessed to minimize device distortion.

In this work, an adaptive noise reduction technique is applied to the audio samples before extracting features. In this technique, an audio signal is passed to a High Pass Filter (HPF) to attenuate the lower frequency components such as wind or idle noise.

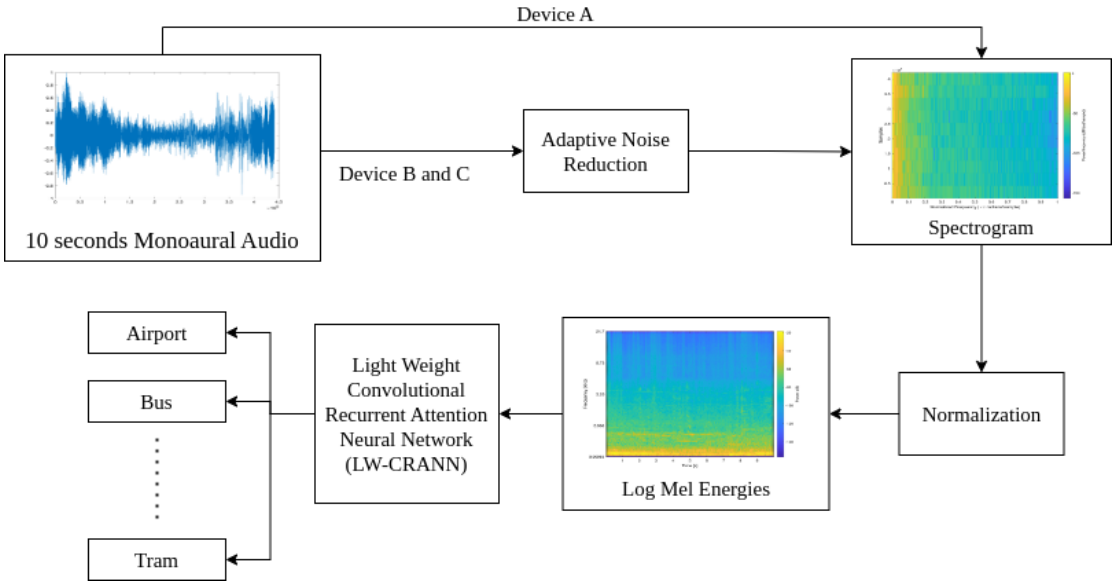

Figure 5.10: Block diagram of the proposed device robust ASC system

The additional noise that is not removed in the recording stage is discarded. In all acoustic scenes, the wind and idle noise may be considered as the additional noises and be safely attenuated without discarding the useful information. Similarly, the high-frequency components that are generated due to certain mics need to be attenuated. This is performed using the Infinite Impulse Response (IIR) Filter. The use of IIR filters is a common and effective approach to reducing device distortion, enhancing audio quality in various electronic devices. IIR filters are designed to mimic the behavior of analog filters, utilizing feedback mechanisms to achieve a desired frequency response with fewer computational resources compared to Finite Impulse Response (FIR) filters. When applied to audio signals, IIR filters can effectively attenuate unwanted frequencies that contribute to distortion, such as harmonics generated by nonlinearities in the device's signal path. By carefully designing the IIR filter coefficients, engineers can target and reduce specific distortion components, thereby improving the clarity and fidelity of the audio output. The recursive nature of IIR filters allows them to achieve a sharp cutoff and a smooth transition between passband and stopband, which is essential for maintaining the integrity of the desired audio signal while minimizing the presence of distortion. The last step is to characterize the type of audio recording. This is performed using Dynamic Range

Control (DRC). In DRC, the input signal is modified at four levels, namely, makeup gain, compressor, noise gate, and limiter. Makeup gain is applied for middle-level signals [1] to achieve a total gain. The compressor is used to attenuate the higher-level signals [2] or to lessen the amplification of the signal. The noise gate is used when the signal is very low. The limiter is used only in a few cases where the compressor cannot catch some specific audio. These four components help remove noise and distortion in the audio samples of devices B and C in the TAU Urban Acoustic Scenes 2019 Mobile dataset.

Once the audio samples are preprocessed, the spectrogram features are extracted. In order to avoid the biased values in the spectrogram, the zero mean and unit variance normalization method is applied to the spectrograms. For the classification of the acoustic scenes, LW-CRANN is proposed and shown in Figure 5.11. The architecture is made light-weight/less complex to minimize the number of additions and multiplications during the computation of weights. The proposed method is efficient in terms of computation cost as the number of computations is reduced by 35% as compared to conventional CRANN. This is achieved by replacing the convolutions in the network with depthwise separable convolutions. Also, the addition of the attention layer in the architecture gives global attention to the features, and the information that is usually not given priority is also utilized. This reduces information loss in the network.

[1] Low-level signals have extremely low output levels and need a lot of amplification, Middle-level signals have middle output levels and more amplification than low-level signals

[2] High-level signals high output levels and need less amplification

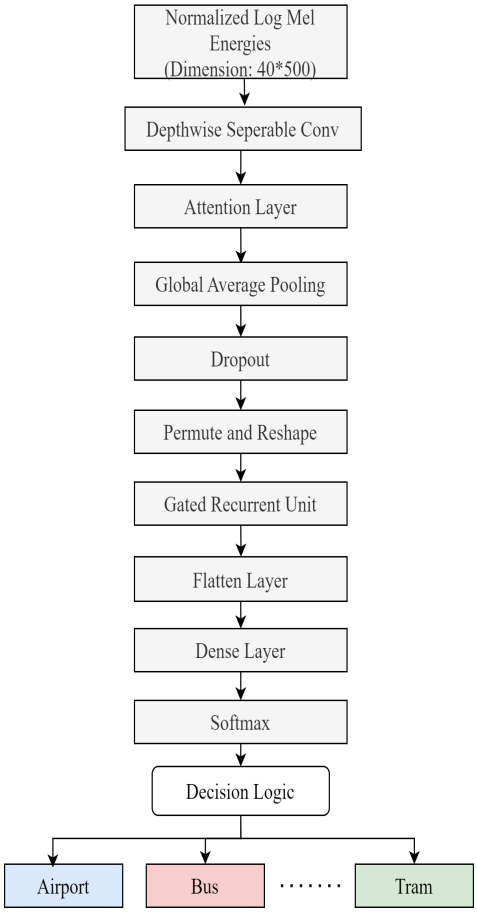

Figure 5.11: Light weight Convolutional Recurrent Attention Neural Network (LW-CRANN) architecture

The attention mechanism is added in the conventional CRNN architecture to handle the input with long and/or complex sequences. If there are variable input length, then a dimensionality issue will be raised, where in these type of inputs will be forced to have same dimension (Bahdanau et al. 2015). In the cases of same acoustic scenes, the event lasts for only shorter span of time. To capture this information, adding attention layer to the network is advantageous. In this work, an attention layer is introduced in the CRANN architecture. The input fed to the CRANN are Normalized log Mel energies with a dimension of 40 * 500. The first layer of the network is depthwise convolution network. This layer learns the spatial information of the input features and the depthwise separable convolution layer uses less number of computations as compared to conventional convolution layer. This is followed by the attention layer. The attention layer uses a dropout score value of 0.2, i.e, 20% of the units are randomly

dropped while training, as it reduces the overfitting. The features that are learnt by the convolution layers are fed to embed the sequences and followed by the attention layer. The next layer is the pooling layer. The global average pooling layer captures the global information of the input feature as opposed to other pooling layers that pools small section of the input. The next layer includes the reshaping of the feature to feed into Gated Recurrent Layer (GRU). The features resulted by the GRU layer are flattened and fed to dense layer with softmax activation function for final acoustic scene label.

5.3.3 Performance Evaluation

The dataset considered to evaluate the proposed device robust ASC system is TAU Urban Acoustic Scenes 2019 Mobile (Mesaros et al. 2018) dataset. The audio samples in the dataset are recorded with three different devices. The dataset consists of a total of 10 acoustic scenes, and the data recordings from different devices are captured simultaneously. The devices the samples are recorded in are Soundman OKM II Klassik/studio A3, electret binaural microphone and a Zoom F8 audio recorder (Device A), Samsung Galaxy S7 (Device B), and iPhone SE (Device C). The number of audio samples present for each device is given in Table 5.8. The training and testing set consist of different audio samples and testing set's data are unseen data samples.

Table 5.8: Number of audio samples for three devices for train and test sets

Device	Device A	Device B	Device C
Training	9185	540	540
Testing	4185	540	540

A. Signal Preprocessing

The adaptive noise reduction is applied on audio samples of devices B and C. The attenuation of idle noise and distortion is performed by HPF and an IIR filters. These filters remove the lower frequency components in the signal, such as wind, and the cutoff frequency is set to 50 Hz. IIR is necessary because certain microphones consist of high peak level frequencies that need to be attenuated to get a lesser distorted signal.

B. Feature Extraction

After the signal preprocessing step, feature extraction is performed on the audio

samples. The audio segments are 10 seconds in length. Spectrogram features are extracted through STFT using "hanning" window. The window length is set to 2048, and the hop size is set to 512 samples. To increase the frequency resolution, the hop size is set smaller (Nguyen et al. 2020). After extracting spectrogram features from the audio segments, the features are processed in the log-mel domain with 40 mel filters. Followed by this, the log-mel features are normalized using the zero mean and unit variance normalization techniques in order to eliminate biased values in the feature matrix.

C. Neural Network Configuration

In this work, the ASC is performed with four different neural network architectures, namely, CRNN, Light-weight Convolutional Recurrent Neural Network (LW-CRNN), CRANN, and LW-CRANN. All architectures are implemented using Keras 2.2.4 and Tensorflow 1.13.1 libraries in Python. The common parameters set for these architectures are as follows: The learning rate of the Adam optimizer is set to 0.001, the loss function used for multi-class classification is Multi-Class Cross-Entropy Loss, the number of epochs the models are trained is set to 200 with early stopping criterion if the validation loss does not improve after ten epochs, and the batch size of samples is set to 64. The CRNN architecture consists of both convolutional and recurrent layers, which capture the input's spatial and temporal information. In the LW-CRNN, the CRNN model is made less complex by replacing the convolution layers with depthwise separable convolution layers. In CRANN, the CRNN network is added with the Global Attention Layer. The model LW-CRNN is also made light-weight by replacing convolution layers with depthwise separable convolutions. All models consist of GRU layer after convolution layers. The last layers are dense.

D. Results and Discussion

The proposed device robust ASC system is evaluated for two sets of features, log mel energies, and normalized log mel energies. The features are fed to four neural networks to analyze the performance of different deep learning models. The systems

developed with these techniques are generalized ones and can be used in real-world applications. The results achieved by proposed ASC system are given in Table 5.9.

Table 5.9: Acoustic Scene Classification by the CRNN, LW-CRNN, CRANN, and the proposed LW-CRANN classifiers on TAU Urban Acoustic Scenes 2019 Mobile dataset uing normalized log Mel energies (Legend: ACC-Accuracy)

Model	**Noise Reduction**	**ACC (in %)**		
		A	B	C
CRNN	No	81.4	76.5	75.8
	Yes	82.0	76.9	76.1
LW-CRNN	No	81.8	77.3	75.6
	Yes	82.7	77.8	75.9
CRANN	No	85.1	78.7	77.9
	Yes	85.8	79.4	78.1
LW-CRANN	No	85.4	79.3	78.0
	Yes	**86.1**	**80.6**	**79.2**

The results obtained by the LW-CRANN with normalized log-mel energies resulted in the highest ASC accuracy in all the cases of devices A, B, and C. By relatively comparing the accuracies, it may be observed that the gap between the performance of the three devices is reduced. The baseline system performance of ASC is 61.9%, 39.6%, and 43.1% for devices A, B, and C, respectively. In comparison with the baseline system, the performance of the proposed systems is better in terms of accuracy and the performance gap between the devices A, B, and C. Additionally, the normalization of the features has enhanced network performances on an average by at least +1%. From the obtained results, it can be stated that elimination of distortion from the audio signal can significantly improve the performance of the ASC system.

The class-wise (separately 10 classes of scenes) performance analysis of the LW-CRANN model is presented in Table 5.10. The results indicate that the performance obtained for the different classes is balanced across all devices. The accuracy values of the different classes for devices A, B, and C are given in the table. The LW-CRANN with normalization step resulted in highest accuracy of ASC task for all three devices. Therefore, only LW-CRANN results are provided in Table 5.10. The performance of outdoor classes such as public square and street pedestrian are

observed to be less accurate. The reason for this may be the interclass similarity between the events present in these acoustic scenes. The highest results are achieved for the park and bus scenes.

The proposed ASC system is compared with the state-of-the-art systems in Table 5.11. DCASE 2019 Task 1(a) challenges' best results namely, McDonnell and Gao

Table 5.10: Class-wise Acoustic Scene Classification by log Mel energies as features and LW-CRANN as classifier of Devices A, B, and C (Legend: ACC-Accuracy)

Classes	**ACC (in %)**		
	A	**B**	**C**
Airport	78.5	71.2	70.3
Bus	96.2	86.4	88.7
Metro	89.9	80.3	84.2
Metro station	81.4	66.8	77.5
Park	96.4	97.5	95.5
Public square	68.0	61.2	66.2
Shopping mall	81.2	87.3	79.4
Street pedestrian	79.4	59.4	55.2
Street traffic	93.4	97.6	93.9
Tram	96.6	98.3	81.1
Average	**86.1**	**80.6**	**79.2**

Table 5.11: Comparison of Accuracy (ACC.) of the proposed LW-CRANN network and other state-of-the-art ASC systems on TAU Urban Acoustic Scenes 2019 Mobile dataset

Sl.No	**Title**	**Method**	**ACC. (in %)**
1.	Integrating the Data Augmentation Scheme with Various Classifiers for Acoustic Scene Modeling (Chen et al. 2019)	Features: Scalogram and Mel Filter Bank features, Classifier	85
2.	Acoustic Scene Classification Using Deep Residual Networks with Late Fusion of Separated High and Low Frequency Paths (McDonnell and Gao 2020)	Features: Log Mel Spectrograms, Classifier: Deep Residual Networks	82.3
3.	Device robust acoustic scene classification using adaptive noise reduction and convolutional recurrent attention neural network (Spoorthy and Koolagudi 2022)	Features: Log Mel Spectrograms, Classifier: **LW-CRANN** (Proposed system)	**86.1**

(2020) and Chen et al. (2019) are used for comparing our results. These two are the top entries in the TAU Urban Acoustic Scenes 2019 challenge submissions. The comparison is done with respect to features, classifiers and accuracy of the system. Both state-of-the-art systems used time frequency representations such as Mel-spectrograms, log-Mel spectrograms, and scalograms. Both McDonnell and Gao (2020) and Chen et al. (2019) used deep learning based neural networks whose computation complexity is high compared to the LW-CRANN (proposed). Based on the accuracies reported by the state-of-the-art works, our proposed approach has achieved an average accuracy of 86.1% which is an improvement of almost 1.1%.

5.3.4 Contributions and Limitations

This work proposed a device robust ASC system using adaptive noise reduction and low-complex deep learning methods. In a real-world scenario, the classification of acoustic scenes must be robust, with respect to the recording device. The recording device characteristics should not affect the performance of the system. Main reason for significant performance decrease in the case of real-time systems reported by the existing ASC systems is embedded noise. An adaptive noise reduction technique is introduced in our work to reduce the device distortion in the case of audio samples recorded from mobile devices. Because of this approach, a significant improvement in the performance of ASC is observed on the mobile devices dataset. The primary reason for improvement in the performance of ASC for mismatched devices is the elimination of the distortion and additional noises in the audio samples. The method for removal of the distortion has to be chosen carefully as in some cases critical audio information can also be discarded in the process. Once the preprocessing is performed, feature extraction is carried out. Spectrograms are extracted from the audio samples, and to remove the biased values in the feature matrix, the normalization is applied. Later, the features are fed to four deep learning architectures. The key contributions of this work are the preprocessing of audio samples of mobile devices to attenuate device distortion and the use of LW-CRANN deep learning architecture to perform the classification of acoustic scenes. From the results, it may be observed that eliminating distortion in the audio samples increases the accuracy and decreases the loss of the

ASC system without using any additional data augmentation. Therefore in future work, more distortion removal techniques may be explored to improve the ASC performance in mismatched ASC audio recording datasets.

5.4 SUMMARY

In this chapter, three methods proposed to perform ASC are discussed. The first method introduces a deep Fisher Network, which is a fusion of two traditional algorithms. In the second method, the classification of acoustic scenes is performed in two levels: a broader and a finer level. This method also proposes a lightweight complex deep learning technique that outperforms existing methods in terms of accuracy and computational complexity (trainable parameters). The third method focuses on making the ASC system more robust by eliminating device distortion in low-quality recordings. A new Adaptive Noise Reduction technique is proposed to deal with device distortions. The results obtained from the proposed ASC methods outperform the state-of-the-art ASC systems reported before in the literature. The next chapter presents the conclusion of the present research work and potential future research directions.

CHAPTER 6

SUMMARY, CONCLUSIONS AND SCOPE FOR FUTURE WORK

The thesis is organized into 6 chapters. The first chapter introduces different auditory scene analysis (ASA) tasks such as Sound Event detection (SED), Sound Source Localization (SSL), joint Sound Event Localization and Detection (SELD) and Acoustic Scene Classification (ASC) with their applications and challenges in brief. The second chapter critically reviews the research work done in the area of polyphonic SED, SELD, and ASC concerning different preprocessing methods, features, augmentation techniques, and classifiers. At the end of this chapter, research gaps are analyzed and problem statement is identified. In the third chapter, Mel-Pseudo CQT technique and a modified recurrent temporal pyramid neural network are proposed for polyphonic SED. The chapter four, presents two deep learning models, namely, Transpose SELDNet and Channel-wise FusionNet for localizing and detetcting non-overlapped and overlapped sound events. Chapter five explores different methods to classify acoustic scenes that are recorded using mismatched recording devices. Chapter six concludes the present work and opens up the path for further research.

6.1 SUMMARY OF THE PRESENT WORK

In this thesis, different ASA tasks such as polyphonic SED, SSL of overlapped and non-overlapped sound events, and ASC have been investigated. Frame-based speech features were preliminarily designed and extracted for speech/speaker recognition tasks.

In this research, we have proposed acoustic event specific features as well as new deep learning methods to perform various ASA tasks. This section gives summary of each contribution in brief.

6.1.1 Characterization and detection of overlapped sound events

A new feature extraction method has been proposed to detect the onset and offset of sound events. Two datasets namely, TUT SED 2016 and TUT SED 2017 have been considered for performing experiments. The dataset consists of different sound events recorded in various scenarios. To perform polyphonic SED, two new methods have been proposed. In the first method, a new feature extraction technique named Mel Pseudo-Constant Q-Transform (MP-CQT) is proposed. These features are fed as input to the CRNN deep learning model. In the second method, a deep learning model named Modified Recurrent Temporal Pyramid Neural Network (MR-TPNN) is proposed for performing SED task. In the network, a temporal pyramid layer is used which extracts temporal information about the events from the given input features. Both proposed methods outperform the state-of-the-art polyphonic SED methods.

6.1.2 Sound Source Localization of different acoustic events

After detection of events, it is necessary to identify the origin of the sound event. This task is termed as Sound Source Localization (SSL). To perform SSL, TAU Nigens Spatial Sound Events 2020 dataset is utilized. The dataset consists of different sound events' onset and offset with the event label. Also, the spatial locations of the sound events are provided. Two new deep learning architectures are proposed to perform localization of sound events. The Transpose Sound Event Localization and Detection Network (SELDNet) utilizes transpose convolution layer which helps in learning spatial information better. Also, a Channel-wise FusionNet is introduced which learns channel information that is present in the multichannel input. The methods displayed an enhanced performance in terms of various metrics in the cases of both detection and localization tasks.

6.1.3 Device Robust Acoustic Scene Classification

Though the performance of the existing ASC systems are high, one of the important factors that can affect the performance is the quality of the audio recordings. The quality of the recordings tends to vary with the devices they have been recorded in. However, in real-life scenario, an ideal ASC system must be able to identify the scene irrespective of the mismatch in recording devices. In this work, we have presented three methods to perform ASC that handle the audio recordings recorded from different recorders. The main datasets considered for this work are TAU Urban Acoustic Scenes 2019 and TAU Urban Acoustic Scenes 2019 Mobile datasets, recorded from three devices. The first method proposes a new feature vector named Fisher vector, which encodes the scene information from the audio recording. These features are extracted using GMM model and then fed to a pre-built dictionary. A deep Fisher Network is developed using fusion of traditional machine learning algorithms. The second method proposes a bi-level light weight deep learning model to perform ASC. In this method, the scene classification is performed at two levels, a broader level, and a finer level. This method has reduced the number of misclassifications by a significant amount as compared to classifying scenes at one level. In the third method, a device robust ASC method is proposed. The method uses an adaptive noise reduction technique to remove noise in the case of the audio samples which are recorded in low-quality recording devices. The use of this noise reduction technique has leveraged the performance in the case of the data recorded using different recording devices. The results obtained from these three methods have outperformed both baseline system and the state-of-the-art ASC systems.

6.2 CONCLUSIONS:

- Mel Pseudo CQT spectrograms captured better event information and provided enhanced performance in the case of polyphonic SED with both clean and noisy audio recordings. Hence, the proposed Mel Pseudo CQT features are robust in nature.

- The combination of CNN and RNN provides better discrimination of sound events and their locations. Therefore, the proposed MR-TPNN, TSELDNet, and

Channelwise FusionNet models helped in achieving better event detection and localization performances.

- The use of Temporal Pyramid Pooling layer in MR-TPNN converts multiple frame-level features into a fixed-lengthed scene-level representation. Therefore, the proposed MR-TPNN works well for variable lengthed audio recordings as well.

- The information present in the different input recording channels is essential to identify the source of any sound event. The proposed channel-wise FusionNet learns the channel-wise information present in the multichannel input, and the new fusion layer combines by concatenating the information obtained from different channels. This type of learning enhanced the localization performance.

- The proposed Deep Fisher Network is a feed-forward feature transformation model which has used Fisher Vector encoding technique to learn discriminative information from the input audio files. The combination of two Fisher layers and the combined GTCC and Mel spectrogram features achieved the highest accuracy for the ASC task.

- Performing ASC task in two levels, i.e., a broader level classification and a finer level classification, resulted in better ASC performance than a single level classification. Also, the misclassified samples after broader level classification are retrained to improve the ASC performance.

- The proposed device robust ASC system in which an Adaptive Noise Reduction (ANR) method is applied to the data recorded using low-quality recording devices helped remove the device distortion present in the audio recordings and provided better ASC performance.

6.3 FUTURE RESEARCH POINTERS:

- ASA systems have a wide range of applications, such as surveillance systems, context-aware devices (mobile phones), rare-sound event detection, and so on. In order to develop a state-of-the-art ASA system for deploying in surveillance systems especially on low resource devices, the ASA systems must be of low complexity. This is one of the major future challenges that needs to be addressed.
- There is a demand for larger datasets to achieve better performance in ASA systems. In the current trends, the DCASE community has released numerous datasets for research purposes. However, the dataset provides a limited number of acoustic event and scene classes. There is a need for a larger number of classes to solve the event detection and scene classification problems in the real-world scenarios.

- There is a need to develop more multimodal standard universal datasets which include information of different sound events' onset and offset, event origin and also the scene type to perform ASA tasks. Autonomous agents may be developed to adapt to classify new sound events while retaining knowledge of the previously learned acoustic events and scenes in real-life scenarios where they continuously keep sensing the surrounding environment.

- One important domain that needs more exploration is the joint sound event and scene analysis. There is a need to develop systems that can identify events present in a scene along with the scene label. In the existing systems, the research on this domain is very limited and the systems perform the event and scene detection tasks independently.

- Mostly the existing systems are concentrated on spectral features in the current ASA systems. It is need of the hour to develop and use deep and image-based features so that advanced machine learning algorithms such as deep neural networks may be explored.

- Most ASA systems proposed in the existing literature are on deep learning architectures. Some of the ensemble methods worked well in improving the

recognition accuracy of the system. More neural network ensembles may be explored in order to get the better performance on different ASA tasks. Also, fusion systems may be developed to leverage the advantages of two or more features or classifiers.

- Many times real-time audio surveillance systems need to investigate regarding the rare sound events in the presence of various background noises and multisource situations. In this case, identifying rare events like gun shots from the acoustic scene helps in forensic applications making the system more robust.

REFERENCES

AbdelHamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G. and Yu, D. (2014). "Convolutional neural networks for speech recognition." *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10), 1533–1545.

Abidin, S., Togneri, R. and Sohel, F. (2017). "Enhanced LBP texture features from time frequency representations for acoustic scene classification." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 626–630.

Adavanne, S., Parascandolo, G., Pertilä, P., Heittola, T. and Virtanen, T. (2017). "Sound event detection in multichannel audio using spatial and harmonic features." In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events*, 6–10.

Adavanne, S., Politis, A., Nikunen, J. and Virtanen, T. (2018a). "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks." *IEEE Journal of Selected Topics in Signal Processing*, 13(1), 34–48.

Adavanne, S., Politis, A. and Virtanen, T. (2018b). "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network." In *26th European Signal Processing Conference (EUSIPCO)*, IEEE, 1462–1466.

Agarwal, A. and Triggs, B. (2006). "Hyperfeatures – Multilevel local coding for visual recognition." In *Computer Vision*, 30–43.

Aziz, S., Awais, M., Akram, T., Khan, M. U., Khursheed, K. and Alhussein, M. (2019). "Automatic scene recognition through acoustic classification for behavioral robotics." *Electronics*, 8, 1–17.

Bahdanau, D., Cho, K. and Bengio, Y. (2015). "Neural machine translation by

jointly learning to align and translate." In *3rd International Conference on Learning Representations*, 1–15.

Bai, X., Du, J., Wang, Z. and Lee, C. (2019). "A hybrid approach to acoustic scene classification based on universal acoustic models." In *Interspeech*, 3619–3623.

Barchiesi, D., Giannoulis, D., Stowell, D. and Plumbley, M. D. (2015). "Acoustic scene classification: Classifying environments from the sounds they produce." *IEEE Signal Processing Magazine*, 32(3), 16–34.

Bear, H. L., Nolasco, I. and Benetos, E. (2019). "Towards joint sound scene and polyphonic sound event recognition." In *Interspeech*, 4595–4598.

Beltrán, J., Chávez, E. and Favela, J. (2015). "Scalable identification of mixed environmental sounds, recorded from heterogeneous sources." *Pattern Recognition Letters*, 68, 153–160.

Bisot, V., Essid, S. and Richard, G. (2017a). "Overlapping sound event detection with supervised nonnegative matrix factorization." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 31–35.

Bisot, V., Serizel, R., Essid, S. and Richard, G. (2017b). "Feature learning with matrix factorization applied to acoustic scene classification." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6), 1216–1229.

Cakir, E., Heittola, T., Huttunen, H. and Virtanen, T. (2015). "Polyphonic sound event detection using multi label deep neural networks." In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 1–7.

Cakır, E., Parascandolo, G., Heittola, T., Huttunen, H. and Virtanen, T. (2017). "Convolutional recurrent neural networks for polyphonic sound event detection." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6), 1291–1303.

Cao, Y., Iqbal, T., Kong, Q., Zhong, Y., Wang, W. and Plumbley, M. D. (2021). "Event-independent network for polyphonic sound event localization and detection." In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 885–889.

Cao, Y., Kong, Q., Iqbal, T., An, F., Wang, W. and Plumbley, M. D. (2019). "Polyphonic sound event detection and localization using a two-stage strategy." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 30–34.

Chandrakala, S. and Jayalakshmi, S. L. (2019). "Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies." *ACM Computational Survey*, 52(3), 1–34.

Chatterjee, C. C., Mulimani, M. and Koolagudi, S. G. (2020). "Polyphonic sound event detection using transposed convolutional recurrent neural network." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 661–665.

Chen, H., Liu, Z., Liu, Z., Zhang, P. and Yan, Y. (2019). "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 1–4.

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions." In *IEEE conference on computer vision and pattern recognition*, 1251–1258.

Chollet, F. et al. (2018). "Keras: The python deep learning library." *Astrophysics source code library*, 1–23.

Coates, A., Ng, A. and Lee, H. (2011). "An analysis of single-layer networks in unsupervised feature learning." *Journal of Machine Learning Research*, 15, 215–223.

Crocco, M., Cristani, M., Trucco, A. and Murino, V. (2016). "Audio surveillance: A systematic review." *ACM Computing Surveys (CSUR)*, 48(4), 1–46.

Dennis, J. W. (2014). *Sound event recognition in unstructured environments using spectrogram image processing*. PhD thesis, Nanyang Technological University, Singapore.

Eronen, A. J., Peltonen, V. T., Tuomi, J. T., Klapuri, A. P., Fagerlund, S., Sorsa, T., Lorho, G. and Huopaniemi, J. (2006). "Audio-based context recognition." *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), 321–329.

Fitzgerald, D., Cranitch, M. and Cychowski, M. T. (2006). "Towards an inverse constant

Q transform." In *120th Audio Engineering Society (AES) Convention*, volume 1, 1–5.

Fonseca, E., Gong, R. and Serra, X. (2018). "A simple fusion of deep and shallow learning for acoustic scene classification." In *15th Sound and Music Computing Conference (SMC)*, 1–8.

Foster, P., Sigtia, S., Krstulovic, S., Barker, J. and Plumbley, M. D. (2015). "Chime-home: A dataset for sound source recognition in a domestic environment." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1–5.

Gemmeke, J. F., Vuegen, L., Karsmakers, P., Vanrumste, B. and Vanhamme, H. (2013). "An exemplar-based NMF approach to audio event detection." In *IEEE workshop on applications of signal processing to audio and acoustics*, 1–4.

Gharib, S., Drossos, K., Emre, C., Serdyuk, D. and Virtanen, T. (2018). "Unsupervised adversarial domain adaptation for acoustic scene classification." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 138–142.

Golubkov, A. and Lavrentyev, A. (2018). "Acoustic scene classification using convolutional neural networks and different channels representations and its fusion." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 1–4.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep learning*, MIT press.

Gordoa, A., Rodríguez-Serrano, J. A., Perronnin, F. and Valveny, E. (2012). "Leveraging category-level labels for instance-level image retrieval." In *IEEE Conference on Computer Vision and Pattern Recognition*, 3045–3052.

Grzeszick, R., Rothacker, L. and Fink, G. A. (2013). "Bag-of-features representations using spatial visual vocabularies for object classification." In *IEEE International Conference on Image Processing*, 2867–2871.

Han, Y. and Lee, K. (2016). "Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification." *IEEE AASP challenge on detection and classification of acoustic scenes and events*, 1–11.

Harma, A., McKinney, M. F. and Skowronek, J. (2005). "Automatic surveillance of

the acoustic activity in our living environment." In *IEEE International conference on multimedia and expo*, 1–4.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." In *IEEE international conference on computer vision*, 1026–1034.

Heittola, T., Mesaros, A. and Virtanen, T. (2020). "Acoustic scene classification in DCASE 2020 challenge: generalization across devices and low complexity solutions." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 56–60.

Howard, A., Sandler, M., Chu, G., Chen, L., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Zhu and Yukun (2019). "Searching for MobilenetV3." In *IEEE/CVF International Conference on Computer Vision*, 1314–1324.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017). "MobileNets: Efficient convolutional neural networks for mobile vision applications." *Computing Research Repository (CoRR)*, 1–9.

Hu, H., Yang, C. H., Xia, X., Bai, X., Tang, X., Wang, Y., Niu, S., Chai, L., Li, J., Zhu, H., Bao, F., Zhao, Y., MarcoSiniscalchi, S., Wang, Y., Du, J. and Lee, C. (2021). "A two-stage approach to device-robust acoustic scene classification." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 845–849.

Hu, J., Shen, L., Albanie, S., Sun, G. and Wu, E. (2020). "Squeeze-and-excitation networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8), 2011–2023.

Iandola, F. N., Moskewicz, M. W., Ashraf, K., Han, S., Dally, W. J. and Keutzer, K. (2016). "SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and 1MB model size." *Computing Research Repository (CoRR)*, 1–13.

Jiang, S. and Shi, C. (2019). "Acoustic scene classification using ensembles of convolutional neural networks and spectrogram decompositions." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 1–5.

Jolliffe, I. T. (1986). "Choosing a subset of principal components or variables." *Principal Component Analysis*, 92–114.

Jung, J., Heo, H., Shim, H. and Yu, H. (2018). "DNN based multi-level feature ensemble for acoustic scene classification.." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 118–122.

Jung, J., Heo, H., Shim, H. and Yu, H. (2020). "Knowledge distillation in acoustic scene classification." *IEEE Access*, 8, 166870–166879.

Kapka, S. and Lewandowski, M. (2019). "Sound source detection, localization and classification using consecutive ensemble of CRNN models." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 119–123.

Kingma, D. P. and Ba, J. (2015). "Adam: A method for stochastic optimization." In *International Conference on Learning Representations (ICLR)*, 1–15.

Koutini, K., Henkel, F., Eghbal-zadeh, H. and Widmer, G. (2020). "CP-JKU submissions to DCASE'20: Low-complexity cross-device acoustic scene classification with RF-regularized CNNs." In *Detection and Classification of Acoustic Scenes and Events DCASE Challenge*, 1–5.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems*, 1097–1105.

Lafay, G., Benetos, E. and Lagrange, M. (2017). "Sound event detection in synthetic audio: Analysis of the DCASE 2016 task results." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 11–15.

Lazebnik, S., Schmid, C. and Ponce, J. (2006). "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2169–2178.

Lehner, B., Koutini, K., Schwarzlmüller, C., Gallien, T. and Widmer, G. (2019). "Acoustic scene classification with reject option based on resnets." In *Proceedings of the Workshop on Detection Classification of Acoustic Scenes and Events*, 1–4.

Leng, Y., Zhao, W., Lin, C., Sun, C., Wang, R., Yuan, Q. and Li, D. (2020). "LDA-based data augmentation algorithm for acoustic scene classification." *Knowledge-Based Systems*, 195, 1–9.

Li, Y., Li, X. and Zhang, Y. (2018). "The SEIE-SCUT systems for challenge on DCASE 2018: Deep learning techniques for audio representation and classification." In *Detection and Classification of Acoustic Scenes and Events Workshop*, 1–3.

Lin, T., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017). "Focal loss for dense object detection." In *IEEE International Conference on Computer Vision*, 2980–2988.

LopezMeyer, P., Del Hoyo Ontiveros, J. A., Lu, H., Cordourier Maruri, H. A., Stemmer, G., Nachman, L. and Huang, J. (2020). "Low-memory convolutional neural networks for acoustic scene classification." In *the Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 1–5.

Ma, L., Smith, D. and Milner, B. (2003). "Environmental noise classification for context-aware applications." In *International Conference on Database and Expert Systems Applications*, 360–370.

Ma, Q., Lin, Z., Chen, E. and Cottrell, G. (2020). "Temporal pyramid recurrent neural network." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5061–5068.

McAdams, S. E. and Bigand, E. E. (1993). "Thinking in sound: The cognitive psychology of human audition." In *4th Workshop in the Tutorial Workshop series organized by the Hearing Group of the French Acoustical Society*, Oxford University Press.

McDonnell, M. D. and Gao, W. (2020). "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 141–145.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. and Nieto, O. (2015). "Librosa: Audio and music signal analysis in python." In *14th Python in Science Conference*, volume 8, 18–25.

Mesaros, A., Adavanne, S., Politis, A., Heittola, T. and Virtanen, T. (2019). "Joint

measurement of localization and detection of sound events." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 333–337.

Mesaros, A., Heittola, T., Dikmen, O. and Virtanen, T. (2015). "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 151–155.

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B. and Virtanen, T. (2017). "DCASE 2017 challenge setup: Tasks, datasets and baseline system." In *Detection and Classification of Acoustic Scenes and Events Workshop*.

Mesaros, A., Heittola, T. and Virtanen, T. (2016a). "Metrics for polyphonic sound event detection." *Applied Sciences*, 6(6), 1–17.

Mesaros, A., Heittola, T. and Virtanen, T. (2016b). "TUT database for acoustic scene classification and sound event detection." In *24th IEEE European Signal Processing Conference (EUSIPCO)*, 1128–1132.

Mesaros, A., Heittola, T. and Virtanen, T. (2018). "A multi-device dataset for urban acoustic scene classification." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 9–13.

Mesaros, A. and Serizel, R. (2013). "Dcase community.").

Misra, H., Ikbal, S., Bourlard, H. and Hermansky, H. (2004). "Spectral entropy based feature for robust ASR." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 193–196.

Mulimani, M., Kademani, A. B. and Koolagudi, S. G. (2020). "A deep neural network-driven feature learning method for polyphonic acoustic event detection from real-life recordings." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 291–295.

Mulimani, M. and Koolagudi, S. G. (2019a). "Robust acoustic event classification using fusion Fisher vector features." *Applied Acoustics*, 155, 130–138.

Mulimani, M. and Koolagudi, S. G. (2019b). "Segmentation and characterization of acoustic event spectrograms using singular value decomposition." *Expert Systems*

with Applications, 120, 413–425.

NaranjoAlcazar, J., Perez-Castanos, S., Zuccarello, P. and Cobos, M. (2020). "Acoustic scene classification with squeeze-excitation residual networks." *IEEE Access*, 8, 2287–2296.

Nguyen, T. and Pernkopf, F. (2019). "Acoustic scene classification with mismatched recording devices using mixture of experts layer." In *IEEE International Conference on Multimedia and Expo (ICME)*, 1666–1671.

Nguyen, T., Pernkopf, F. and Kosmider, M. (2020). "Acoustic scene classification for mismatched recording devices using heated-up softmax and spectrum correction." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 126–130.

Ozer, I., Ozer, Z. and Findik, O. (2018). "Noise robust sound event classification with convolutional neural network." *Neurocomputing*, 272, 505–512.

Pajusco, N., Huang, R. and Farrugia, N. (2020). "Lightweight convolutional neural networks on binaural waveforms for low complexity acoustic scene classification." In *Detection and Classification of Acoustic Scenes and Events Workshop*, 135–139.

Patel, D. and Patel, B. (2023). "Low cost and robust solar tracking system based on data of daily and seasonal variation in sun position regard to specific location on earth." *International Journal of Innovative Research in Science, Engineering and Technology*, 3, 15888–15893.

Patil, A. T., Khoria, K. and Patil, H. A. (2022). "Voice liveness detection using constant-Q transform-based features." In *30th IEEE European Signal Processing Conference (EUSIPCO)*, 110–114.

Peltonen, V., Eronen, A., Parviainen, M. and Klapuri, A. (2001). "Recognition of everyday auditory scenes: Potentials, latencies and cues." In *Proceedings of Audio Engineering Society*, 1–5.

Perronnin, F., Sánchez, J. and Mensink, T. (2010). "Improving the Fisher kernel for large-scale image classification." In *Proceedings of Computer Vision*, 143–156.

Pham, L., Phan, H., Nguyen, T., Palaniappan, R., Mertins, A. and McLoughlin, I.

(2021). "Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework." *Digital Signal Processing*, 110, 1–10.

Phan, H., Koch, P., Hertel, L., Maass, M., Mazur, R. and Mertins, A. (2017). "CNN-LTE: a class of 1-X pooling convolutional neural networks on label tree embeddings for audio scene classification." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 136–140.

Phan, H., Pham, L., Koch, P., Duong, N. Q., McLoughlin, I. and Mertins, A. (2020). "Audio event detection and localization with multitask regression network." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 1–4.

Plinge, A., Grzeszick, R. and Fink, G. A. (2014). "A bag-of-features approach to acoustic event detection." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3704–3708.

Politis, A., Adavanne, S. and Virtanen, T. (2020a). "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 165–169.

Politis, A., Mesaros, A., Adavanne, S., Heittola, T. and Virtanen, T. (2020b). "Overview and evaluation of sound event localization and detection in DCASE 2019." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 684–698.

Prajit Ramachandran, Barret Zoph, L. Q. (2018). "Searching for activation functions." In *6th International Conference on Learning Representations (ICLR)*, 1–13.

Qin, R., Qiao, K., Wang, L., Zeng, L., Chen, J. and Yan, B. (2018). "Weighted focal loss: An effective loss function to overcome unbalance problem of chest X-ray14." In *IOP Conference Series: Materials Science and Engineering*, 1–8.

Rajapakse, M. and Wyse, L. (2005). "Generic audio classification using a hybrid model based on GMMs and HMMs." In *11th IEEE International Multimedia Modelling Conference*, 53–58.

Rakotomamonjy, A. and Gasso, G. (2015). "Histogram of gradients of time–frequency

representations for audio scene classification." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1), 142–153.

Ranjan, R., Jayabalan, S., Nguyen, T. N. T. and Gan, W. S. (2019). "Sound event detection and direction of arrival estimation using residual net and recurrent neural networks." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 214–218.

Ren, Z., Kong, Q., Qian, K., Plumbley, M. D. et al. (2018a). "Attention-based convolutional neural networks for acoustic scene classification." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 39–43.

Ren, Z., Qian, K., Zhang, Z., Pandit, V., Baird, A. and Schuller, B. (2018b). "Deep scalogram representations for acoustic scene classification." *IEEE/CAA Journal of Automatica Sinica*, 5(3), 662–669.

Roma, G., Nogueira, W. and Herrera, P. (2013). "Recurrence quantification analysis features for environmental sound recognition." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1–4.

Salamon, J. and Bello, J. P. (2015). "Unsupervised feature learning for urban sound classification." In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 171–175.

Salvati, D., Drioli, C. and Foresti, G. L. (2019). "Urban acoustic scene classification using raw waveform convolutional neural networks." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 1–4.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks." In *IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Scheirer, E. and Slaney, M. (1997). "Construction and evaluation of a robust multifeature speech/music discriminator." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1331–1334.

Schilit, B., Adams, N. and Want, R. (1994). "Context-aware computing applications." In *IEEE Workshop on Mobile Computing Systems and Applications*, 85–90.

Schörkhuber, C. and Klapuri, A. (2010). "Constant-Q transform toolbox for music processing." In *7th Sound and Music Computing conference*, 3–64.

Sehili, A., Lecouteux, B., Vacher, M., Portet, F., Istrate, D., Dorizzi, B. and Boudy, J. (2012). "Sound environment analysis in smart home." *Lecture Notes in Computer Science*, 7683, 208–223.

Serizel, R., Bisot, V., Essid, S. and Richard, G. (2016). "Machine listening techniques as a complement to video image analysis in forensics." In *IEEE International Conference on Image Processing (ICIP)*, 948–952.

Shi, W., Caballero, J., Theis, L., Huszar, F., Aitken, A., Ledig, C. and Wang, Z. (2016). "Is the deconvolution layer the same as a convolutional layer?." *arXiv preprint arXiv:1609.07009*, 1–7.

Simonyan, K., Vedaldi, A. and Zisserman, A. (2013). "Deep fisher networks for large-scale image classification." In *Advances in Neural Information Processing Systems 26*, 163–171.

Singh, A., Thakur, A., Rajan, P. and Bhavsar, A. (2018). "A layer-wise score level ensemble framework for acoustic scene classification." In *26th IEEE European Signal Processing Conference (EUSIPCO)*, 837–841.

Smith, N. and Gales, M. J. (2002). "Using SVMs and discriminative models for speech recognition." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 77–80.

Song, H. and Yang, H. (2019). "Feature enhancement for robust acoustic scene classification with device mismatch." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 1–5.

Soonshin Seo, J. K. (2021). "MobileNet using coordinate attention and fusions for low-complexity acoustic scene classification with multiple devices." In *Proceedings of the Workshop on Detection Classification Acoustic Scenes and Events*, 1–5.

Spoorthy, V. and Koolagudi, S. G. (2022). "Device robust acoustic scene classification using adaptive noise reduction and convolutional recurrent attention neural network." In *International Conference On Speech And Computer*, Springer, 688–699.

Spoorthy, V. and Koolagudi, S. G. (2023a). "Bi-level acoustic scene classification using lightweight deep learning model." *Circuits, Systems, and Signal Processing*, 1–20.

Spoorthy, V. and Koolagudi, S. G. (2023b). "Polyphonic sound event detection using Mel-Pseudo Constant Q-Transform and Deep Neural Network." *IETE Journal of Research*, 1–13.

Spoorthy, V. and Koolagudi, S. G. (2023c). "A Transpose-SELDNet for polyphonic sound event localization and detection." In *IEEE 8th International Conference for Convergence in Technology (I2CT)*, 1–6.

Spoorthy, V. and Kooolagudi, S. G. (2024). "Polyphonic sound event localization and detection using channel-wise fusionnet." *Applied Intelligence*, 54(6), 5015–5026.

Spoorthy, V., Mulimani, M. and Koolagudi, S. G. (2023). "Acoustic scene classification using deep fisher network." *Digital Signal Processing*, 139, 1–13.

Team, T. T. D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A. et al. (2016). "Theano: A python framework for fast computation of mathematical expressions." *arXiv preprint arXiv:1605.02688*, 1–19.

Vesperini, F., Gabrielli, L., Principi, E. and Squartini, S. (2019). "Polyphonic sound event detection by using capsule neural networks." *IEEE Journal of Selected Topics in Signal Processing*, 13(2), 310–322.

Vuegen, L., Broeck, B., Karsmakers, P., Gemmeke, J. F., Vanrumste, B. and Hamme, H. (2013). "An MFCC-GMM approach for event detection and classification." In *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, 1–3.

Waldekar, S. and Saha, G. (2019). "Wavelet based Mel-Scaled features for DCASE 2019 task 1a and task 1b." In *the Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events*, 1–5.

Wan, V. and Renals, S. (2005). "Speaker verification using sequence discriminant support vector machines." *IEEE transactions on speech and audio processing,*, 13(2), 203–210.

Wang, D. and Brown, G. J. (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press.

Wang, M., Wang, R., Zhang, X. and Rahardja, S. (2019a). "Hybrid constant-Q transform based CNN ensemble for acoustic scene classification." In *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1511–1516.

Wang, Q., Wu, H., Jing, Z., Ma, F., Fang, Y., Wang, Y., Chen, T., Pan, J., Du, J. and Lee, C. H. (2020). "The USTC-IFLYTEK system for sound event localization and detection of DCASE 2020 challenge." In *the Proceedings of the Workshop on Detection of Acoustic Scenes and Events*, 1–4.

Wang, R., Wang, M., Zhang, X. and Rahardja, S. (2019b). "Domain adaptation neural network for acoustic scene classification in mismatched conditions." In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 1501–1505.

Wang, W. and Liu, M. (2019). "The SEIE-SCUT systems for acoustic scene classification using CNN ensemble." In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events*, 1–3.

Xia, X., Togneri, R., Sohel, F. and Huang, D. (2018). "Confidence based acoustic event detection." In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 306–310.

Xie, W., He, Q., Yu, Z. and Li, Y. (2022). "Deep mutual attention network for acoustic scene classification." *Digital Signal Processing*, 123, 1–13.

Xu, Y., Huang, Q., Wang, W. and Plumbley, M. D. (2016). "Hierachical learning for DNN-based acoustic scene classification." *Detection and Classification of Acoustic Scenes and Events Workshop*, 1–5.

Yang, W. and Krishnan, S. (2017). "Combining temporal features by local binary pattern for acoustic scene classification." *IEEE/ACM Transactions on Audio, Speech, and Language Processing,*, 25(6), 1315–1321.

Ye, J., Kobayashi, T., Murakawa, M. and Higuchi, T. (2015). "Acoustic scene

classification based on sound textures and events." In *23rd ACM international conference on Multimedia*, 1291–1294.

Yu, Z., Xu, X., Chen, X. and Yang, D. (2019). "Temporal pyramid pooling convolutional neural network for cover song identification.." In *28th International Joint Conference on Artificial Intelligence (IJCAI)*, 4846–4852.

Zeinali, H., Burget, L. and Černocký, J. (2019). "Acoustic scene classification using fusion of attentive convolutional neural networks for DCASE 2019 challenge." In *Proceedings of Workshop on Detection and Classification of Acoustic Scenes and Events*, 1–5.

Zhao, J., Kong, Q., Song, X., Feng, Z. and Wu, X. (2022). "Feature alignment for robust acoustic scene classification across devices." *IEEE signal processing letters*, 29, 578–582.

Zieliński, S. and Lee, H. (2019). "Automatic spatial audio scene classification in binaural recordings of music." *Applied Sciences*, 9, 17–24.

PUBLICATIONS

JOURNAL PAPERS

1. Spoorthy. V, Manjunath Mulimani, and Shashidhar G. Koolagudi. (2023). Acoustic Scene Classification using Deep Fisher Network, Digital Signal Processing (Elsevier), vol. 139, pp. 1–13, doi: 10.1016/j.dsp. 2023.104062

2. Spoorthy. V, Shashidhar G. Koolagudi (2023). Bi-level Acoustic Scene Classification using Lightweight Deep Learning Model, Circuits, Systems, and Signal Processing (Springer), Vol. 42, pp. 1–20, doi: https://doi.org/10.1007/s000 34-023-02478-0

3. Spoorthy. V, Shashidhar G. Koolagudi. (2023). Polyphonic Sound Event Detection using Mel-Pseudo Constant Q-Transform and Deep Neural Network, IETE Journal of Research (Taylor & Francis), Vol. 69, pp. 1–13, doi: https://doi.org/10.1080/03772063.2023.2253768

4. Spoorthy. V, Shashidhar G. Koolagudi. (2023). Polyphonic Sound Event Localization and Detection using Channel-Wise FusionNet, Applied Intelligence (Springer), Vol. 54, pp. 1–12, doi: https://doi.org/10.1007/s10489-024-05438-6

5. Manjunath Mulimani, Spoorthy. V, and Shashidhar G. Koolagudi. (2023). Acoustic Event and Scene Classification: A Review, SN Computer Science (Springer). **(Major Revision Submitted)**

CONFERENCE PAPERS

1. Spoorthy. V, Manjunath. Mulimani and Shashidhar. G. Koolagudi, "Acoustic Scene Classification using Deep Learning Architectures," 2021 6th International

Conference for Convergence in Technology (I2CT), 2021, pp. 1-6, doi: 10.1109/I2CT51068.2021.9418177.

2. Spoorthy. V and Shashidhar. G. Koolagudi, "Device Robust Acoustic Scene Classification using Adaptive Noise Reduction and Convolutional Recurrent Attention Neural Network," 24th Speech and Computer (SPECOM), 2022, pp. 688-699, doi: 10.1007/978-3-031-20980-2_58

3. Spoorthy. V and Shashidhar. G. Koolagudi, "A Transpose-SELDNet for Polyphonic Sound Event Localization and Detection", 8th International Conference for Convergence in Technology (I2CT), 2023, pp. 1–6, doi: 10.1109/I2CT57861.2023.10126251

4. Spoorthy. V and Shashidhar. G. Koolagudi, "Polyphonic Sound Event Detection using Modified Recurrent Temporal Pyramid Network", 8th International Conference on Computer Vision and Image Processing (CVIP), Communications in Computer and Information Science, Vol. 2009. Springer, Cham. doi:10.1007/978-3-031-58181-6_47

BIODATA

Name:	SPOORTHY. V
Date of Birth:	20th September, 1994
Gender:	Female
Marital Status:	Single
Father's Name:	Venkatesh. C
Mother's Name:	Vasantha. T
Address:	Srinivasa Krupa, Opposite Dr. Indudhar Clinic, Behind T. B, Chitradurga-577501 Karnataka, India
E-mail:	vspoorthy036@gmail.com
Mobile:	+91-8310726037
Qualification:	B.E in Information Science and Engineering, GMIT Davangere, Visvesvaraya Technological University, 2016 M.Tech in Computer Science & Engineering, NIT Goa, Speech Processing, 2018
Areas of Interest:	Audio Processing, Deep Learning, Machine Learning