TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVATIONS

S.No	ABBREVATION	EXPANSION
1	**ASC**	Acoustic Scenes Classification
2	**BLSTM**	Bidirectional Long Short Term Memory
3	**DNN**	Deep Neural Network
4	**AI**	Artificial Intelligence
5	**FSD**	Free Sound Dataset
6	**CNN**	Convolutional Neural Network
7	**GMM**	Gaussian Mixture Model
8	**DNN**	Deep Neural Network
9	**MIR**	Music Information Retrieval

1. INTRODUCTION

Acoustic Scene Classification (ASC) is to determine a test audio recording belongs to which pre-given class of acoustic scenes, e.g. park, home, office. ASC is definitely useful for multimedia retrieval, audio-based surveillance and monitoring. The problem of ASC has been received great attention from the signal processing community with many evaluation campaigns, and is not effectively solved because of various factors, such as heavy background noise, large variations of time-frequency properties within each type of acoustic scenes.

The overall performance of one ASC system mainly depends on two modules: feature extraction and classifier building. Almost all of the previous works focused on these two modules for improving the performance. Most back-end classifiers mainly consist of Gaussian Mixture Model (GMM), Deep Neural Network (DNN), random forest, decision tree, gradient boosting, support vector machine, hidden Markov model. For example, a bag-of-frame method was considered as a baseline system for the challenges, which used MFCCs in combination with GMM. Eghbal-Zadeh et al presented an I-vector extraction scheme for ASC using both left and right audio channels, and used a deep convolutional neural network trained on spectrograms of audio excerpts. Valenti et al used a convolutional neural network fed by the feature of BLSTM to classify acoustic scenes. Another method for ASC applied non-negative matrix factorization to spectro-temporal representation for decomposing features into activations of spectro-temporal elements.

Although many systems were proposed for ASC, to the best of our knowledge, no system combines the feature of framework for acoustic scene representation with the classifier of BLSTM for acoustic scene classification to date. we propose to build a DNN for extracting the framework based on MFCCs, and then feed the framework into a classifier of BLSTM for ASC. Compared to traditional features (e.g., MFCCs), the proposed feature of DAF is a deep transformed and compact representation of the original high-dimensional inputs fed to the input layer of the DNN, and thus can more effectively characterize the property differences among various classes of acoustic scenes.

On the other hand, BLSTM has emerged as a scalable and state-of-the-art classifier for several classification problems related to sequential data. Considering the high context correlation among the acoustic scenes and the advantage of BLSTM in capturing sequence information, we propose to use BLSTM based classifier fed by the proposed feature of the framework to realize the task of ASC. Hence, main contributions of this study are to extract a feature of framework for representing the property of each acoustic scene and propose a method for ASC by combining the framework with the BLSTM network. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

1.1 DOMAIN OVERVIEW

1.1.1. Deep Learning

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain albeit far from matching its ability allowing it to “learn” from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy. Deep learning drives many artificial intelligence (AI) applications and services that improve automation, performing analytical and physical tasks without human intervention. Deep learning technology lies behind everyday products and services (such as digital assistants, voice-enabled TV remotes, and credit card fraud detection) as well as emerging technologies (such as self-driving cars).

If deep learning is a subset of machine learning, how do they differ? Deep learning distinguishes itself from classical machine learning by the type of data that it works with and the methods in which it learns. Machine learning algorithms leverage structured, labeled data to make predictions meaning that specific features are defined from the input data for the model and organized into tables. This doesn’t necessarily mean that it doesn’t use unstructured data; it just means that if it does, it generally goes through some pre-processing to organize it into a structured format.

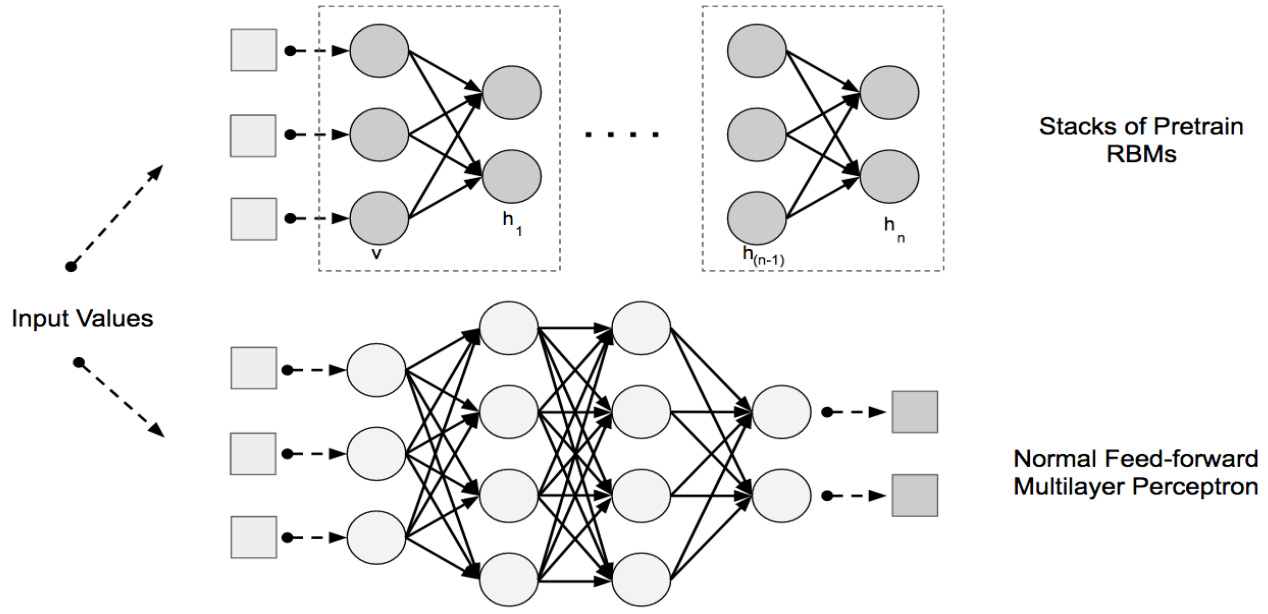

Figure 1.1 Deep Learning Architecture

1.1.2. BLSTM CLASSIFICATION

A Bidirectional LSTM**, or** LSTM**,** is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backwards direction. BLSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm (e.g. knowing what words immediately follow and precede a word in a sentence).

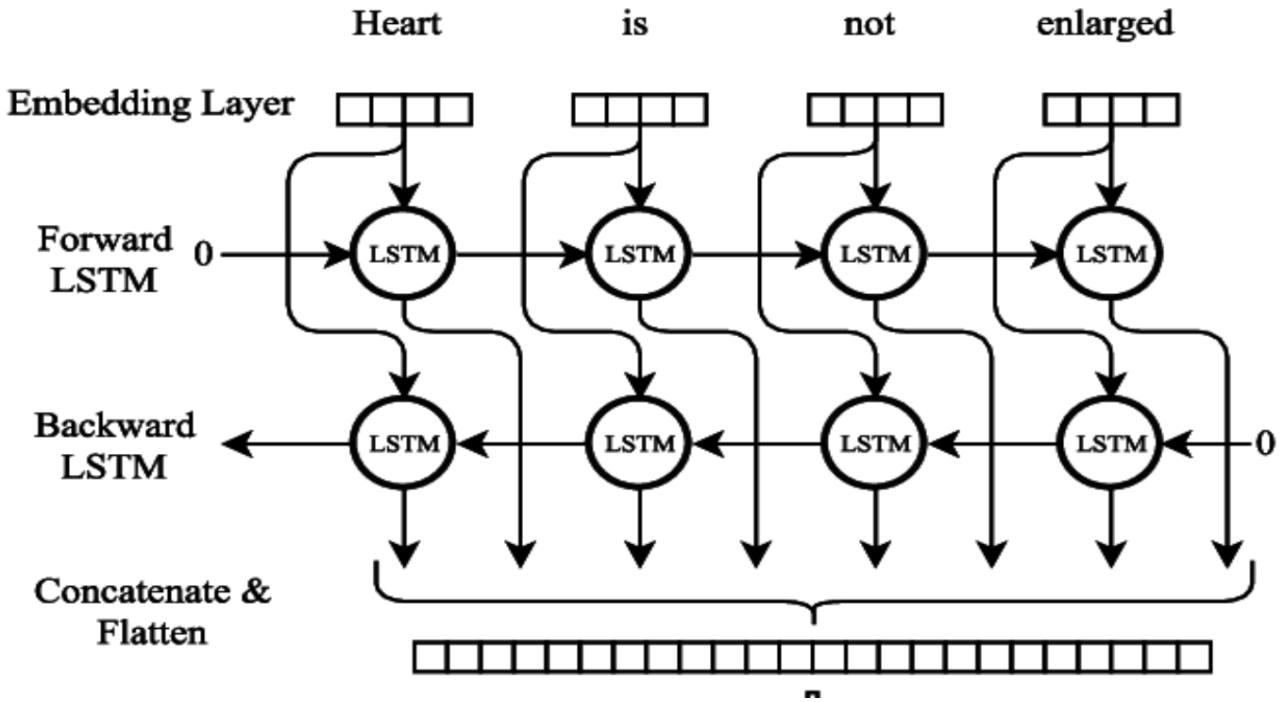

Figure 1.2 BLSTM Architecture

1.1.3. TARSOSDSP

TarsosDSP is a Java library for audio processing. Its aim is to provide an easy-to-use interface to practical music processing algorithms implemented, as simply as possible, in pure Java and without any other external dependencies. This text serves two goals: it serves as a practical introduction into Music Information Retrieval (MIR) techniques for computer science students, and as documentation for the TarsosDSP library. The text assumes familiarity with object oriented programming constructs and a good knowledge of a programming language of the C-family, like Java. The concepts used should be transferable to other programming languages, or platforms as well. TarsosDSP features implementation of the following algorithms. For each algorithm there is an example application1 available. TarsosDSP was originally conceived as a library for pitch estimation, therefore it contains several pitch estimators and an estimator based on dynamic wavelets.

2. LITERATURE SURVEY

A literature review surveys scholarly articles, books, dissertations, conference proceedings and other resources which are relevant to a particular issue, area of research, or theory and provides context for a dissertation by identifying past research. Research tells a story and the existing literature helps us identify where we are in the story currently. It is up to those writing a dissertation to continue that story with new research and new perspectives but they must first be familiar with the story before they can move forward.

2.1 Audio Scene Classification using Deep Learning Architectures.

AUTHOR: Spoorthy.V, Manjunath Mulimani, Shashidhar G.Koolagudi

YEAR: 2021

Enabling devices to make sense of sound is known as Acoustic Scene Classification (ASC). The analysis of various scenes by applying computational algorithms is known as computational auditory scene analysis. The main aim of this paper is to classify audio recordings based on the scenes/environment in which they are recorded. Deep learning is amongst the recent trends in most of the applications. In this paper, two deep learning algorithms are used to perform the classification of acoustic scenes, namely Convolution Neural Network (CNN) and Convolution-Recurrent Neural Network (CRNN). The model is evaluated on three activation functions, namely, ReLU, Leaky ReLU and ELU. The highest recognition accuracy achieved for ASC task is 90.96% from CRNN model. The model performed well on basic convolution architecture with 10.9% improvement from the baseline system of this task.

2.2 Perceptual loss based speech de-noising with an assemble of of Audio Pattern Recognition and Self-Supervised Models.

AUTHOR: Saurabh Kataria, Jesus Villalba, Najim Dehak.

YEAR: 2021

Deep learning based speech de-noising still suffers from the challenge of improving perceptual quality of enhanced signals. We introduce a generalized framework called Perceptual Ensemble Regularization Loss (PERL) built on the idea of perceptual losses. Perceptual loss discourages distortion to certain speech properties and we analyze it using six large-scale pre-trained models: speaker classification, acoustic model, speaker embedding, emotion classification, and two self-supervised speech encoders (PASE+, wav2vec 2.0). We first build a strong baseline (w/o PERL) using Conformer Transformer Networks on the popular enhancement benchmark called

VCTKDEMAND. Using auxiliary models one at a time, we find acoustic event and self-supervised model PASE+ to be most effective. Our best model (PERL-AE) only uses acoustic event model (utilizing Audio Set) to outperform state-of-the-art methods on major perceptual metrics. To explore if de-noising can leverage full framework, we use all networks but find that our seven-loss formulation suffers from the challenges of Multi-Task Learning. Finally, we report a critical observation that state-of-the-art Multi-Task weight learning methods cannot outperform hand tuning, perhaps due to challenges of domain mismatch and weak complementarity of losses.

2.3 Audio Scene Classification with Discriminatively-Trained Segment-Level Features.

AUTHOR: Haichuan Bai, Hangting Chen, Yonghong Yan.

YEAR: 2021

The discriminatively-trained segment-level audio features derived from this network are concatenated to the frame-level features, and fed into the DNN-based back-end classifier. Then the post-processing mechanism is applied. Experiment results demonstrate that the proposed system with the discriminatively-trained segment-level features achieves the classification accuracy of 75.68%, and the absolute improvement of 13.64% is gain in comparison with the referential systems only using the frame-level features.

2.4 Scene-Dependent Acoustic Event Detection with Scene Conditioning and Fake-Scene-Conditioned Loss.

AUTHOR: Tatsuya Komatsu; Keisuke Imoto; Masahito Togami.

YEAR: 2020

Scene-dependent acoustic event detection (AED) with scene conditioning and fake-scene-conditioned loss. The proposed method employs a multitask network, that has not only AED part but also acoustic scene classification (ASC). The scenes predicted by ASC are employed as an additional feature for scene conditioning of AED to learn the relationship between scenes and events. For efficient training, the proposed method incorporates a new AED loss function, which is the fake-scene-conditioned loss, in addition to the conventional AED loss. Upon training, the AED part is conditioned with fake scenes as well as predicted and true scenes. The fake-scene-conditioned loss is calculated between the fake-scene-conditioned AED results and labels of events that do not exist in

the fake scenes are removed. Whereas training with combinations of true scenes/events, i.e., the conventional AED loss, only reveals that an event is present in a scene, with fake-scene-conditioned loss, the proposed method can learn that an event is absent in a scene. Experimental results show that the proposed method improves the AED performance compared with the baseline; an increase in the f1 score of 23% and a decrease in the false alarm rate of 56% for scenes where no event exists.

2.5 Acoustic Scene Classification Using Deep Audio Features.

AUTHOR: Yanxiong Li, Xianku Li, Yuhan Zhang, Wucheng Wang, Mingle Liu, Xiaohui.

YEAR: 2019

Every mental disorder stems from stress. There are numerous causes of stress, all of which have been shown to have a harmful influence on the human body. As a result of increasing management expectations, time management challenges, and monetary terms, stress has a significant impact on the life of a working professional. When stress is disregarded for an extended period of time, it promotes depression and anxiety risks. The physiological parameters aid in the identification of stress-related disorders. The brain signals were employed to investigate stress in this work. For the purposes of research, the EEG signals are exact, accurate, and dependable. The main interruptions we have obtained by utilising this test are sweating, increased body temperature, and an increase in anxiety level. Despite this, EEG has a high degree of compatibility with stress signals. We conducted a thorough investigation of the various classifiers, and SVM achieved the highest accuracy. The feasibility of employing the EEG for stress detection and clinical intervention and prevention of physical and mental health disorders was determined.

2.6 EXISTING SYSTEM

From the review of literature in existing work, the architecture of some well-known spectral features when fed to Support Vector Machine (SVM) classifier used to classify the audio scenes. Furthermore, it analyzes different methods of combining these features, and also of combining information from two channels when the data is in different format. The existing approach resulted in around 57% accuracy with respect to the baseline system on the development and evaluation dataset. The evaluation dataset used in this project is Rouen Dataset.

2.6.1 DISADVANTAGES OF EXISTING SYSTEM

- Existing audio-based analysis requires significantly High computation time.
- Obviously classification based on these low-level features alone may not be accurate.
- The old dataset doesn’t matches to the modern environment and it only classify the acoustic scenes.
- High complexity algorithm will result in poor performance.

3. PROPOSED SYSTEM

3.1 INTRODUCTION

The scope of this work, increasing the feasibility of the system by making use of Deep Learning Algorithm. The BLSTM is a Deep Learning Algorithm is used to classify the audio scenes and reducing the background noise by using TarsosDSP framework to give a clear output and in this project using a Free Sound Dataset (FSD) which is used to predict the audio which was recorded this dataset has more pre-trained models.

3.2 ARCHITECTURE OF PROPOSED SYSTEM

The aim of the work is to develop a system functionality that will be able to reduce the background noise and classify the acoustic scene that was captured in the system. By using the Deep Learning models and a Computational framework and converting the audio file to .wav format to process the audio file easier and get a clear output.

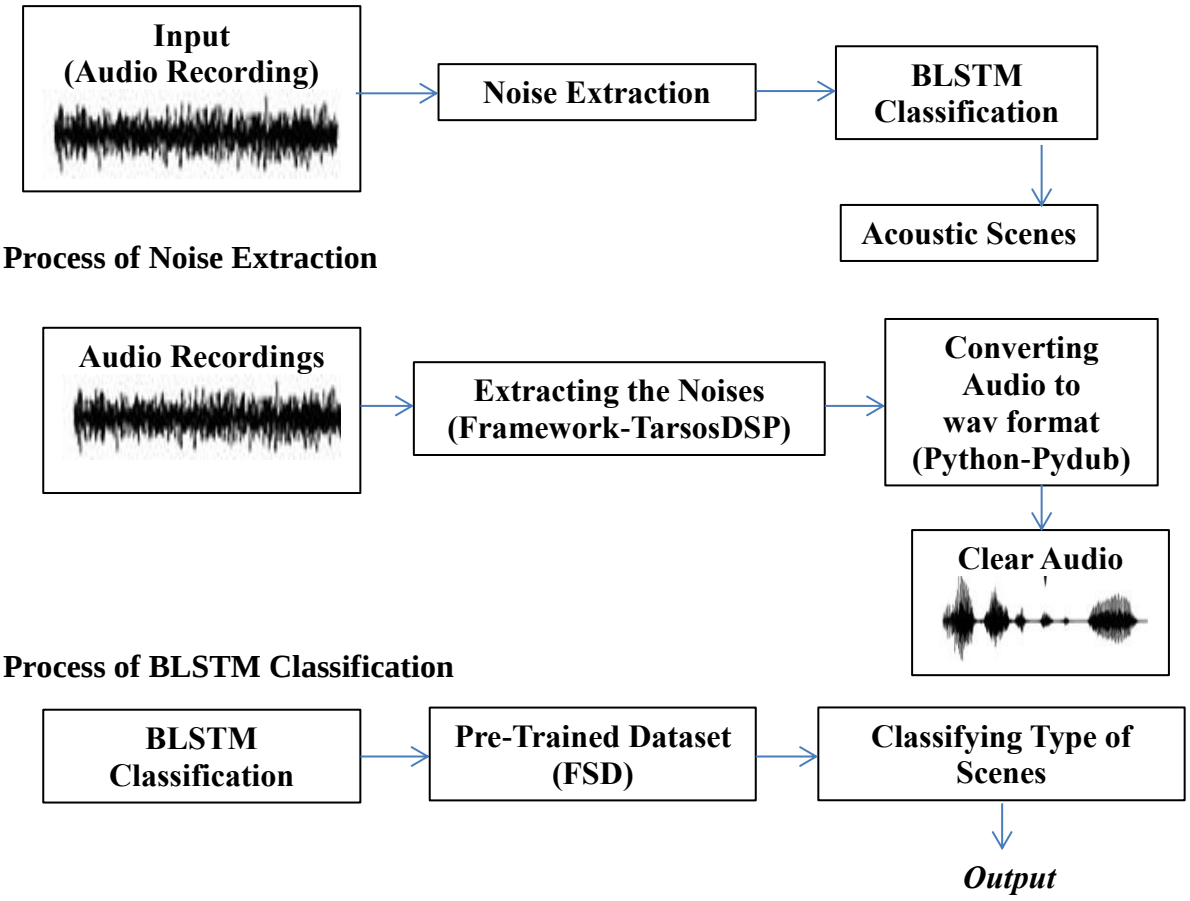

Figure 3.1 Architecture of Proposed System

3.2.1 TarsosDSP

TarsosDSP is a Java library for audio processing. Its aim is to provide an easy-to-use interface to practical music processing algorithms implemented, as simply as possible, in pure Java and without any other external dependencies. This text serves two goals: it serves as a practical introduction into Music Information Retrieval (MIR) techniques for computer science students, and as documentation for the TarsosDSP library. The text assumes familiarity with object oriented programming constructs and a good knowledge of a programming language of the C-family, like Java. The concepts used should be transferable to other programming languages, or platforms as well.

3.2.2 Free Sound Dataset

The Audio Set Ontology is a hierarchical collection of over 600 sound classes and we have filled them with 297,159 audio samples from Free Sound. This process generated 678,511 candidate annotations that express the potential presence of sound sources in audio clips. FSD includes a variety of everyday sounds, from human and animal, sounds to music and sounds made by things, all under Creative Commons licenses. By creating this dataset, we seek to promote research that will enable machines to hear and interpret sound similarly to humans. Free sound is a platform for the collaborative creation of audio collections labelled by humans and based on Free Sound content.

3.2.3 BLSTM Classification

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was designed by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

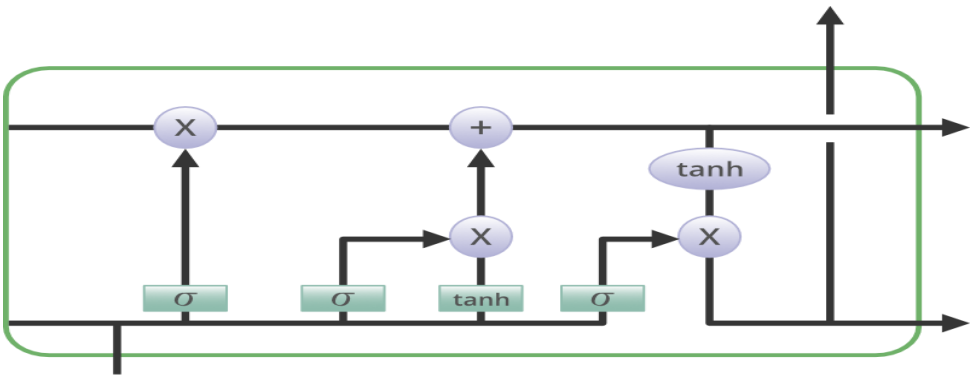

Figure 3.2 LSTM Classifier Architecture

3.2.4 PyDub (Python Library)

Python provides a module called Pydub to work with audio files. Pydub is a Python library to work with only .wav files. By using this library we can play, split, merge, edit our .wav audio files.

3.3 ADVANTAGES OF PROPOSED SYSTEM

- TarsosDSP is a framework and it is used for audio file processing like extracting the noise from the background to give a clear output.
- It uses Deep Learning Algorithm called BLSTM (Bidirectional Long Short-Term Memory Networks) which is used to classify the audio file using the dataset.
- Proposed System uses Free Sound Dataset (FSD) which has more audio samples.

3.4 APPLICATIONS

- Used in various successful applications in both Computer Vision (CV) and Speech Recognition.
- Applications of this system can be in context-aware and intelligent wearable devices, hearing-aids, robotic navigation systems, and audio archive management systems.
- Airport , Indoor shopping mall, Metro station, Pedestrian street, Public square, Street with medium level of traffic, Travelling by a tram, Travelling by a bus, Travelling by an underground metro, Urban park.

4. SYSTEM REQUIREMENTS

4.1 SOFTWARE REQUIREMENTS

Operating system	:	Windows 10
Languages	:	Python, Java
Dataset	:	Rouen Dataset, Free Sound Datasets
Algorithm	:	BLSTM Algorithm
Framework	:	TarsosDSP

4.2 HARDWARE REQUIREMENTS

Monitor	:	Any Monitor
Processor	:	Intel i3 and above
RAM	:	4GB and above
Devices	:	Mic, Receiver

5. CONCLUSION

A system of ASC for the challenge by using a feature of framework and an effective classifier of BLSTM. The proposed system outperformed the baseline system on both development and evaluation datasets. The results have shown that the proposed system is effective for solving the problem of ASC. The future work includes exploring other Deep-transformed features for more effectively representing the properties of different acoustic scenes studying on the design and parameters settings of classification models for further improving the performance investigating the technique of data augmentation to generate additional data for training deep neural network, since participants of the challenge are constrained to use a relatively small training dataset as a rule. There is an assumption that increase in the depth of the neural network provides better performance of the system. That is true to only one extent. The model's performance depends on the activation function chosen in the hidden layers of the network. The highest recognition accuracy is achieved is 90.96% from the CRNN model In future work, different deep learning architectures and feature representations can be used to improve the performance of the ASC system.

REFERENCES

[1] Y. Li, Q. Wang, X. Zhang, W. Li, X. Li, J. Yang, X. Feng, Q. Huang, and Qianhua He, "Unsupervised classification of speaker roles in multiparticipant conversational speech," Computer Speech and Language, vol. 42, pp. 81-99, 2017.

[2] Mesaros, A., Heittola, T. and Virtanen, T., 2018. A multi-device dataset for urban acoustic scene classification. arXiv preprint arXiv:1807.09840.

[3] Waldekar, Shefali & Saha, Goutam. (2018). Classification of audio scenes with novel features in a fused system framework. Digital Signal Processing. 75. 10.1016/j.dsp.2017.12.012.

[4] Q. Kong, Y. Xu, W. Wang and M. D. Plumbley, "A joint detectionclassification model for audio tagging of weakly labelled data," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 641-645, doi: 10.1109/ICASSP.2017.7952234.

[5] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," in Proc. of Detection and Classification of Acoustic Scenes and Events 2019, Sep. 2019.

[6] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. of IEEE ICASSP, pp. 6645- 6649, 2018..

[7] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proc. of the Conf. on Empirical Methods in Natural Lang. Process., pp. 1724-1734, 2014.

[8] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in Proc. of IEEE ICASSP, pp. 4869-4873, Apr. 2015.

[9] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: a systematic review," ACM Computing Surveys, vol. 48, no. 4, pp. 1-46, 2018.