National College of Ireland

Speech Emotion Recognition using Deep Learning

MSc Research Project
Data Analytics

Priyanka Prashant Chimthankar
Student ID: x19241721

School of Computing
National College of Ireland

Supervisor: Dr Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Priyanka Prashant Chimthankar
Student ID:	x19241721
Programme:	Data Analytics
Year:	2020-21
Module:	MSc Research Project
Supervisor:	Dr Christian Horn
Submission Due Date:	16/08/2021
Project Title:	Speech Emotion Recognition using Deep Learning
Word Count:	6749
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Priyanka Prashant Chimthankar
Date:	14th August 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Speech Emotion Recognition using Deep Learning

Priyanka Prashant Chimthankar
x19241721

Abstract

Speech Emotion Recognition (SER) has a broad range of applications and there has been a significant amount of research in this fascinating area in recent years. However, the entertainment sector suffers from a lack of study in this research.The Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architectures will be utilized to categorize the emotions in audio recordings captured by actors expressing various emotions. An innovative method will be discussed that combines 2D CNN+LSTM with MFCC features extracted from audio data. Multiple experiments are used to determine the reliability of such systems that use deep learning. The model is based on four widely used datasets in SER: SAVEE, RAVDESS, TESS, and CREMA-D, and has a validation accuracy of 67.58%. Additionally, this model was evaluated on an unknown dataset that included audio samples in the German language and achieved a testing accuracy of 71.28%.

1 Introduction

1.1 Background

Music information retrieval (MIR) and music emotion recognition (MER) research have gained prominence in terms of extracting and suggesting music based on the title, style, artist, and emotions. There are several applications for deriving sentiments from human voice recordings. Emotions in humans may be described in terms of one's actions, behavior, ideas, feelings, gestures, or psychology. Along with the detection and classification of music based on emotions, SER is a fascinating and demanding topic of research (Issa et al.; 2020).

Various researchers have attempted to detect emotions from facial expressions (Florence and Uma; 2020). However, a person's outward expressions may differ from his or her internal emotions. Due to the fact that one might seem cheerful on the outside but be miserable on the inside, algorithms can extract inaccurate emotions in such instances. That's why the focus of this research proposal is on identifying emotions in speech using deep learning techniques, as shown in the Figure 1 [1].

1.2 Motivation

SER is the process of obtaining the speaker's emotions from the speech.Recognizing these emotions provides insight into deeper complexity, which aids in navigating real-world circumstances.The identification of emotions in speech is a significant problem in

[1] https://medium.com/@raihanh93/speech-emotion-recognition-using-deep-neural-network-part-i-68edb5

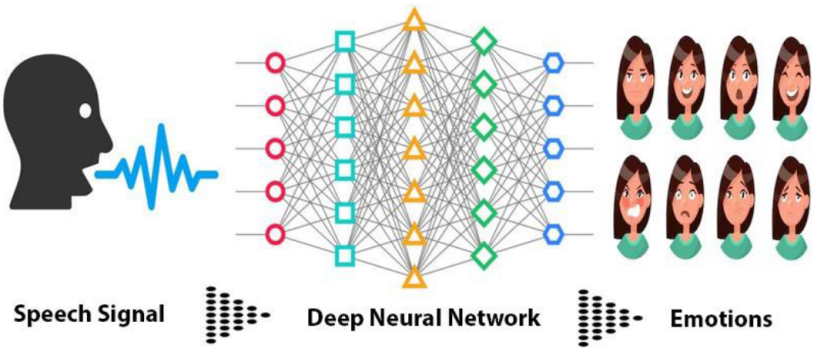


Figure 1: Basic Architecture of Speech Emotion Recognition

the field of human-computer interaction (Patel et al.; 2021). It is thus useful to develop robust emotion detection algorithms that can replicate human perception in such a way that people can identify emotions like happiness, anger, and sadness when conversing.

SER systems have a variety of applications in different domains.The emotions retrieved from a person's speech can be used in a variety of fields, including healthcare, cognitive sciences, psychology, and marketing. Similarly, in the entertainment industry, the emotions that can be detected from an actor's speech can be included in the subtitles as the emotional state of that particular scene. Additionally, it may be used to propose background music depending on the artist's speech emotions. Research into how to synthesize background music for various scenes might aid in the advancement of knowledge in a variety of fields, including human behavior (Wang et al.; 2020). Even though considerable study has been conducted on music synthesis, the challenge of selecting appropriate music based on scene emotion remains unsolved. Background music may be thought of as a reflection of the emotions present in a movie/play scene. In other words, it expresses or communicates human emotions musically.

This research focuses on experimenting with four distinct datasets that comprise voice recordings of several actors in a variety of emotional states. Using these emotion-based datasets, a SER model is built using deep learning that may be utilized for a variety of applications in the entertainment sector.In spite of numerous research studies in this field, the reliability of such systems is still questionable. In order to bring SER models into real world applications, there are multiple theories that need to be analyzed deeply to take this field on the correct path and further provide valuable findings for researchers to help build robust real-time SER systems.

1.3 Research question

Is it reliable to identify emotions in real-time speech using SER models built using deep learning techniques?

1.4 Objective

The major objectives of this research proposal are:

1. To build a SER system using combined deep learning neural networks
2. To build a SER model compatible for multiple data sources

3. To conduct several experiments using four distinct datasets in order to assess a variety of scenarios.

4. To evaluate the proposed model on a completely unseen data.

1.5 Plan of the paper

The rest of the research paper is structured as follows. Section 2 outlines previous research on this subject. Section 3 details the research methodology used in this work, from data collection to the results. Section 4 describes the suggested approach to this research problem in detail. Section 5 contains information about the proposed model's implementation stages, which may be used to replicate this work. Section 6 discusses the main findings that support the research question and the evaluation process. Section 7 summarizes the study and makes recommendations for future work that may be useful for doing further research in this area.

2 Related Work

2.1 Speech Emotion Recognition

In comparison to Support Vector Macines(SVM) and K-Nearest Neighbours(KNN) classifiers, the gradient boosting method performed well in recognizing emotions in real-time voice data (Iqbal and Barua; 2019). The authors describe a real-time emotion detection system that extracts 34 characteristics from audio recordings stored in a spoken emotion database. The extraction of features is accomplished through the use of the open-source Python package pyAudioAnalysis. On various datasets, three distinct classification methods are applied: SVM, KNN, and Gradient Boosting. To summarize, various classifiers performed differently on the SAVEE and RAVDESS datasets [2]. SVM is the most often used classifier in databases, but gradient boosting outperformed SVM and KNN on live data.This solution may be enhanced by removing noise from the data and utilizing a huge library of high-quality audio samples. The author suggests to employ deep learning to develop a more accurate system of emotion recognition from speech, as well as to expand the study to include more emotions such as fear, surprise, disgust, etc.

Pandey et al. (2019) introduce several deep learning algorithms and methods for applying them to analyze speech data and identify the emotional states that are being expressed. Algorithms such as CNN and LSTM have attracted attention for evaluating the emotion detection capacity from standard speech representations such as the magnitude spectrogram, mel spectrogram, and Mel-Frequency Cepstral Coefficients (MFCC's) on two publicly available datasets, EMO-DB and IEMOCAP. The results of several studies have been presented, together with the logic behind them, to demonstrate which models and feature combinations are optimal for speech emotion recognition.

Kowtha et al. (2020) applied a LSTM model and a time convolution-LSTM (TC-LSTM) model to extract primitive emotion characteristics such as arousal,valence, and dominance from expression. In comparison to the baseline model, the concordance correlation coefficient (CCC) for valence increased by 30 percent while training on several datasets and utilizing resilient features. The study demonstrated that when compared to earlier work on a publically accessible dataset, TC-LSTM performs significantly better

[2]For description of datasets see section 3.1

in detecting valence. Several findings were made during the research regarding which characteristics were effective at identifying valence and how simple score level fusion may assist in boosting emotion identification efficiency. Ultimately, categorical emotions were extracted from the speech using the emotion primitives scores.

SER is generally viewed as a three-stage statistical pattern recognition issue composed of the following stages: feature extraction, feature selection, and pattern classification. Before conducting SER, it is necessary to grasp a number of ideas.Sönmez and Varol (2020a) discussed these concepts in their research article. For researchers interested in SER, this study provides a single source of information that covers speech development, the human auditory system, emotion terminology, and emotion modelling. Prior to speech signal processing, it is critical to understand verbal communication, the hearing system, the sensations of the brain, and the emotion-generating centers of the brain. This research teaches you how to produce speech, detect emotions, analyze voice, and recognize speech emotions.

One of the main advantages of deep learning approaches is the automatic extraction of features, which may be used to fundamental properties inherent in audio files with a particular emotion, for example, in the part of speech emotion identification. Issa et al. (2020) present a novel architecture for extracting features from audio files such as MFCCs, Tonnetz representation,mel-scale spectrogram, chroma-gram, and spectral contrast and feeding them into a 1-d CNN for emotion detection using three different databases: RAVDESS, EMO-DB, and IEMOCAP. To increase classification accuracy, the prior model is modified by utilizing the incremental method.Rather than using a single audio feature, several audio features are utilized to combine numerous sound properties like pitch, harmony, timbre, etc into a single training utterance. This leads to a more detailed explanation of an audio sample, which improves the performance of speech-related emotion recognition models.

When the suggested approaches' outcomes were compared to those of prior approaches, the best model beat the existing systems on the RAVDESS and IEMOCAP datasets, establishing a new state of the art. While EMO-DB exceeds every previous research except one, it improves it in terms of simplicity, generality, and usability. Further study on this subject is yet to be done. The models' accuracy may still be improved by including additional features or LSTM layers, or by utilizing an auxiliary neural network to obtain high-level features. Additionally, the larger the dataset, the more accurate the model is; hence, increasing the amount of the dataset through data augmentation approaches will increase the model's performance. Further study in this area may potentially reveal the best arrangement of the features to be selected.

Chernykh and Prikhodko (2017) detected emotion in speech using a deep recurrent neural network (RNN) trained on a set of acoustic characteristics recorded across short speech durations. Sequentially, a distinctive probabilistic-nature CTC loss feature permits the consideration of lengthy utterances with both sentimental and emotionless components. In comparison to other contemporary approaches in this field, our methodology was more precise and also aided in evaluating human performance on the same activity. It explains why emotionality may be conveyed in only a few frames of an utterance when the CTC loss feature is used. Additionally, the suggested method forecasts the emotional sequence associated with a single statement. Though the suggested model's features might be enhanced by employing a more powerful feature generating approach. The input audio signals can be in raw format, which may enhance the model's accuracy.

Even when different algorithms for SER are created, success rates vary according

to the language of the speech, emotions, and datasets. A unique text-independent and speaker-independent SER method called 1BTPDN was used to build a lightweight strategy for addressing a nonpolynomial issue by scraping handmade features (Sönmez and Varol; 2020b). The 1BTPDN approach assists in minimizing important data loss while also increasing accuracy.

Prior to supplying input to the classifier, the most critical processes are feature extraction and feature selection. The features and noise in the raw inputs are identified and eliminated using 1D-DWT. The feature extraction technique makes use of binary patterns, ternary patterns, and DWT. These approaches are used to demonstrate a multileveled function generating methodology. The time complexity of this technique is predicted to be O (nlogn). The 1024 features are extracted from the 7680 features using a version of NCA and DWT. Finally, the selected characteristics are fed into the classifier SVM cubic model, which outperformed all other machine learning models. This approach may also be used to optimize speed when dealing with huge datasets and real-time data.

This work improves the area of MER by bringing new emotionally linked audio features (Panda et al.; 2020). This study is primarily concerned with identifying the auditory features that are most useful for detecting emotional content in music. The current audio features used for MER were examined, and it was discovered that musical ideas lacked computational extractors, necessitating the development of new approaches linked to musical texture expressiveness. The project was assessed through the creation of a public dataset of 900 audio recordings. The extra features proposed in addition to the existing features improved the F1 score by 9 percent as compared to the F1 score obtained with only the existing features (baseline) when using an SVM classifier.

Another recent study,Fu et al. (2021) proposes an attention-dependent CNN-BLSTM model with an end-to-end multitask learning system that depends on the Mel-spectrogram for improving SER. The model was trained to identify emotions, combine emotions and speaker identification, and generate SER using cross-language input. The work contributes to the categorization of emotions and speakers through the use of multitask learning. On the SAVEE dataset, the model was assessed using single task, multitask, and cross-language training.

When compared to their finest data, the suggested method significantly outperforms the baseline PCRN system. The average weighted accuracy improved by 11.13 points when compared to earlier work. On the basis of experimental data, it can be concluded that multitask learning combined with speaker recognition aids in the improvement of SER performance, particularly in cross-language training.

Etienne et al. (2018) developed a neural network for recognizing emotions in speech using the IEMOCAP dataset. In accordance with recent developments in audio synthesis, this article utilizes an architecture that incorporates both convolutional layers for extracting high-level characteristics from raw spectrograms and recurrent layersfor combining long-term dependencies.This article uses the data augmentation approach to alter the duration of vocal tracks. Additionally, the author examined the effect of batch normalization, a technique that is required for nearly all picture identification jobs. Normalization was performed layer by layer and batch by batch in order to keep as much signal structure as necessary. These approaches contributed to the achievement of comparable results of 64.5 percent weighted accuracy and 61.7 percent unweighted accuracy on four fundamental emotions, which did not improve over the baseline, probably due to the small batch size.

2.2 CNN-LSTM

A technique for music emotion identification has been developed using the convolutional long short-term memory deep neural network (CLDNN) architecture (Hizlisoy et al.; 2021). Additionally, a new Turkish emotional music collection has been built, composed of 124 Turkish traditional music snippets lasting 30 seconds each, and the suggested approach has been tested on it. Along with the basic acoustic characteristics, the CNN layers are given log-mel filterbank energies and MFCCs. Feature extraction is performed using tools such as MIRToolbox, but deep learning may also be used to extract features. The greatest results are obtained when combining LSTM + DNNclassifier to integrate the new feature with the conventional features. When LSTM+DNN classifiers are compared to KNN, SVM, and random forest classifiers, the findings indicate that LSTM+DNN classifiers achieve better accuracy. The suggested approach for identifying emotions in music includes four convolutional layers, one LSTM layer, and fully linked layers.

The increasing demand for efficient and accurate real-time SER in human-computer interactions necessitates a study of existing approaches and datasets in SER in order to arrive at the best solution and a thorough understanding of this open-ended issue (Abbaschian et al.; 2021). This article discusses deep learning approaches for SER using publicly available datasets, as well as standard machine learning methodologies for speech emotion detection. Finally, a multi-aspect study of techniques for speech emotion recognition using functional neural networks is provided. The purpose of this research is to provide an overview of differentiated speech emotion detection.

According to numerous other recent research, CNN is superior at resolving emotion detection issues because of their increased low-level and short-term discriminative skills. The incorporation of LSTM networks and the use of deep convolutional LSTM designs have continued to improve the solution's performance and give the network long-term memory, enabling it to identify long-term paralinguistic patterns. Additionally, they demonstrated enhanced speaker-independent emotional processing abilities.

The challenging topic of recognizing emotions in music using a variety of artificial intelligence and machine learning approaches has sparked the interest of several academics. Chen and Li (2020) proposed a multi-feature hybrid network classifier using CNN-LSTM for classification of audio and lyrics in a recent study on music emotion classification to overcome the constraints of the single network classification paradigm. The hybrid model for music classification using two-dimensional and one-dimensional emotional features significantly improved classification accuracy as compared to the single modal classification in this work. This paper offers an introduction to multimodal music emotion detection and highlights the importance of audio and lyrics as key elements for categorizing music based on its emotional content, as well as advocating more study in this field utilizing deep learning.

2.3 SER applications in entertainment sector

Wang et al. (2020) recently published research in which they contributed to a technique for autonomously synthesizing real-time background music while a user navigates around a virtual setting. This technique, which is based on visual sentiment analysis, creates music that corresponds to the emotional states conveyed in the picture while allowing for a seamless transition. This study may be used in a number of contexts, including gaming music, real-world settings, and the incorporation of a user's music tastes. The author employs a cascade method comprised of two components: visual analysis of the

scene and background music synthesis. To begin, a deep-learning-based approach based on neural networks is used to determine the scene's sentiment. The second stage is to optimize a cost function in order to synthesize background music. Additionally, quantitative and qualitative examination of the synthesis findings for several example situations was conducted to confirm the approach's efficacy.

The suggested technique may cause difficulties in situations with varying lighting, backdrops, and layouts, as the system can not efficiently learn such scenarios and is evaluated only on salient objects. Instead of pictures, videos may be utilized to aid in the understanding of emotions, perhaps enhancing the performance of backing music. Additionally, in future work, certain other aspects, such as storytelling while walking through the scenes, may be explored.

Liu and Chen (2018) present a background music recommendation system for user-generated videos in which the model is trained using latent components of features. The author's objective is to deduce latent variables from a set of data in order to provide new insights. Apart from technological considerations, this essay is mostly concerned with human factors. Along with the suggested algorithm's detailed deviation, SGD is performed to improve the proposed algorithm. By requesting participants to provide feedback, we conducted a comparative analysis of the acquired results, along with quantitative and qualitative research. Additionally, the project intends to identify the data with emotions and make it accessible to any other research investigations that might profit from this information. Furthermore, the author promotes study into topics relevant to developing systems that would accept video as an input and generate corresponding background music. According to the research's future work, deep learning approaches may be utilized to map video and audio characteristics to latent space in order to learn multiple function representations.

Florence and Uma (2020) created a music recommendation system for consumers based on their webcam-captured facial expressions. The pictures recorded are then utilized to determine the user's mood/emotion. After extracting the emotion from the user's facial expressions, a list of music that corresponds to the user's mood is shown on the screen. The user can choose any song from those and the order of the recommended songs changes dependent on their frequency of selection. Experimental analysis is performed on the system's output.

The analysis was divided into two sections: one in which participants were instructed on how to proceed with guessing the emotion conveyed, and another in which they were uninformed of the procedure and were not given any instructions. Both outcomes were logged, and the system occasionally failed when the user's inner mood differed from his or her facial expression. Additionally, the system is unable to accurately capture all emotions since the picture dataset comprises fewer photos of the user and must be shot in a well-lit area in order for the classifier to produce reliable findings.

3 Research Methodology

This research study followed the stages of the KDD approach. From dataset selection through getting useful findings from the research project, each stage is explained below.

3.1 Data Selection and Data Understanding

After gaining a thorough understanding of and accumulating adequate domain knowledge, the four most frequently used datasets for SER were chosen. These key datasets are publicly accessible via kaggle. This research compares the performance of the model created with these datasets and helps in drawing interesting conclusions.

1. Surrey Audio-Visual Expressed Emotion (SAVEE):

   The SAVEE database (Jackson and Haq; 2014) contains 480 British English utterances from four native English male speakers, researchers and postgraduate students at the University of Surrey, aged 27 to 31 years. The captured emotions have been classified as anger, disgust, fear, surprise, sadness, and happiness.This dataset is a male only dataset but can be balanced by the other dataset that contains only female recordings.

2. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):

   The RAVDESS (Livingstone and Russo; 2018) is comprised of 24 professional actors, 12 female and 12 male, each of whom delivers two lexically-related sentences. Speech emotions include expressions of 'calm', 'happy', 'angry','sad', 'fearful','surprise', and 'disgust'. For the sake of our experiment, we chose the 'calm' emotion as a 'neutral' emotion. The dataset comprises audio-only, audio-visual, and video-only files; for our research, we used just the audio speech files.This dataset is very rich in variations and contains a North American English accent, which is present in very few datasets.

3. Toronto Emotional Speech Set (TESS):

   Two female native English speakers picked 200 words to describe each of the seven emotions in this dataset: anger, disgust, pleasant surprise, neutral, happiness, fear, and sorrow (Dupuis and Pichora-Fuller; 2011). Two actresses (aged 26 and 64) delivered a list of target words using the carrier phrase "Say the word." The generated dataset had a total of 2800 samples. The 'pleasant surprise' emotion was determined as the 'surprise' emotion for the experimental study. This dataset is a female-only dataset and has very good quality audio files.

4. Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D):

   CREMA-D is an audio-visual database that was built with the aim of multi-modal emotion detection (Cao et al.; 2014). We conducted our experiment using samples of spoken emotional expressions without visual cues, giving a total of 1440 samples, evenly split between male and female recordings. The recordings were made using actors from a variety of ethnic backgrounds, which enables diverse accents to be addressed during training and validation.

5. Emotional Database (EMO-DB): This dataset is a German dataset that is used as an unseen dataset in this research study.This dataset is described in detail in section 6.4

3.2 Data Preprocessing

The very first step after selecting the datasets was to identify and interpret the audio files. Each dataset had its own distinct naming convention. The emotion label was derived from the file names and then utilized for classification.

The datasets were then analyzed using wave plots displaying randomly picked audio files. These graphs help to highlight the type of data that will be researched. Additionally, the quality of the speech data may be examined by determining if the audio recordings include any background noise and whether the emotions can be easily interpreted by humans.

This study explored how the deep learning model behaves differently when applied to individual datasets and then to all four merged datasets. Due to the fact that the audio files in each dataset have varying durations, the model may not train effectively.The mean duration of all data files was selected, which was 3 seconds for this experiment in order to make them model ready.

3.3 Data Transformation

One of the primary stages in SER is the extraction of features and generation of characteristics from the input data. These features and characteristics are often linked to the speech signal's short-term spectrum or to the structure of the vocal tract (Langari et al.; 2020). Because we can not directly manipulate the audio file, it is important to collect the best features. This guarantees that the collection of features we select is descriptive of the underlying data in our models, greatly increasing the role of features in our study. With this in mind, we aim to reduce the number of features that may give us information about the scene's emotions.

The purpose of feature extraction is to concentrate on the information found in the signal, to increase the measure of correlation and differences between various classes, and to decrease the size of the data and computations. While the obtained features from every pattern are useful for classification, this research uses the MFCC feature extraction technique to try to enhance the efficiency of SER.A typical process of feature extraction using MFCC for each raw audio sample will be as shown in Figure 2

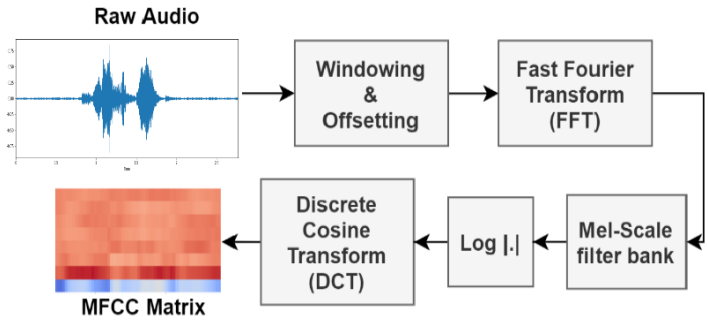


Figure 2: Block diagram of MFCC computation

3.4 Model Building

The baseline emotion recognition model operates as follows: The fixed length MFCC's are derived from the input audio samples. Then the resultant two-dimensional matrix is fed into a CNN classifier with max pooling layers instead of fully connected. In this way, the workable SER system is then built to test the framework and to check the emotion recognition classifier. Further, using the baseline approach, a model is implemented using 2D-CNN and LSTM, giving higher accuracy in detecting emotions.

3.5 Evaluation

In this study, a confusion matrix was utilized to evaluate the model for recognizing emotions from speech.The accuracy of the model was noted for measuring the performance of the SER model and was calculated by using the below equation.Also, the performance of the model was tested by various experiments, which will be discussed in section 6

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{1}$$

where:
TP = True Positive,
TN = True Negative,
FP = False Positive,
FN = False Negative.

4 Design Specification

The study proposal for identifying emotions in recorded speech includes a SER system, as illustrated in Figure 3. A combination of CNN and LSTM models is applied in this system to identify emotions in audio recordings. The proper number of features is determined as an input to the CNN+LSTM model using the feature extraction. The various components that comprise this suggested architecture are described below.

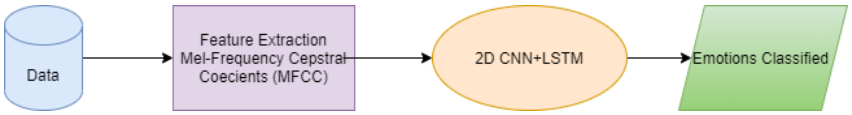


Figure 3: Proposed model for Speech Emotion Recognition

4.1 Data

Every machine learning work demands a training set of samples; SER is no exception. To create a training dataset for SER, human agents must manually classify the samples, because different individuals experience emotions differently.For instance, one person may label an emotional voice as angry, while another may label it as excited. Due to this uncertainty, labeling the samples requires several agents to evaluate each and then a system to reliably select labels for available samples.

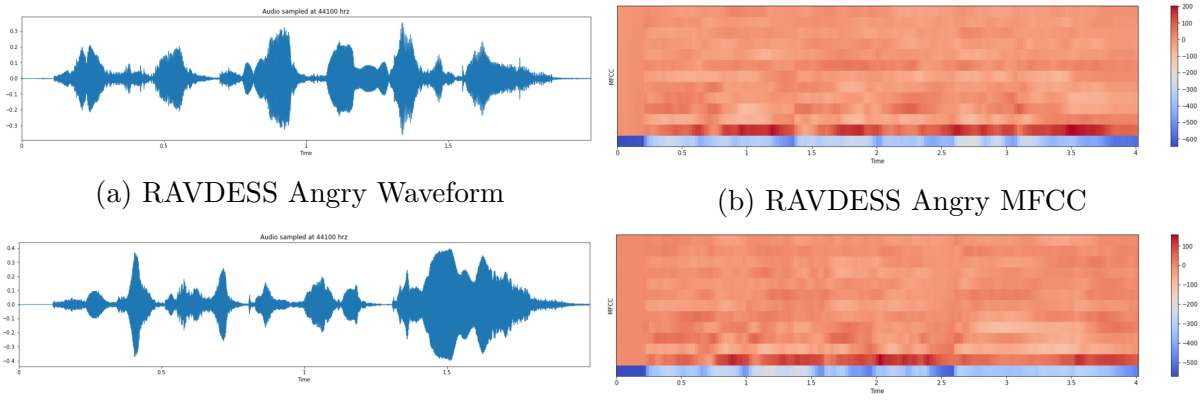
There are three kinds of databases dedicated to the identification of speech emotions: simulated (artificial), semi-natural and natural speech data (Abbaschian et al.; 2021). The

simulated datasets are generated by qualified speakers reciting the same sentence with various levels of emotional intensity.Semi-natural collections are created by requesting that individuals or actors read a scenario including a range of emotions. Additionally, natural datasets are collected from television programs, YouTube videos, etc., and the emotions are classified by human listeners.

In this research four different simulated datasets are used namely RAVDESS, SAVEE, TESS and CREMA-D as mentioned in the section 3.1. The raw audio files contained in these datasets are labelled into seven different emotions.

4.2 Feature Extraction

MFCC is one method for feature extraction that is be used to analyze speech by extracting critical data and features from subsets of the speech data.The MFCCs of an audio signal are a minimal number of features (often 10–20) that succinctly reflect the pattern of a spectral envelope.MFCC features are computed using linearly spaced frequency filters at low frequencies and logarithmically spaced frequency filters at high frequencies. This architecture is similar to the human ear, which is sensitive to low frequency (Mishra and Sharma; 2020). As shown in Figure 4, the MFCC matrix for every audio sample is calculated.



(a) RAVDESS Angry Waveform

(b) RAVDESS Angry MFCC

(c) RAVDESS Happy Waveform

(d) RAVDESS Happy MFCC

Figure 4: Sample audio waveform and MFCC from RAVDESS-female

4.3 Model

4.3.1 Baseline Model: CNN

Deep learning algorithms extract higher-level characteristics from voice signals than low-level features do. The capacity of CNN to describe two-dimensional signals is well-known. Due to the adaptive feature extraction properties of the CNN model, it is utilized for emotion detection in audio files (Zhang et al.; 2016). Throughout the initial computation, the CNN network retains and recognizes the speech signal's spatial structure, while disregarding sequential cues. Numerous additional recent results indicate that convolutional neural networks are efficient at addressing problems related to emotion detection because they possess enhanced low-level and short-term discriminative capabilities(Abbaschian et al.; 2021)

4.3.2 Proposed Model: 2D CNN-LSTM

In this study, we integrated LFLB and LSTM to learn local and global features from the raw audio samples and MFCCs, respectively. It is capable of learning a sequence feature in which each component is a function of a limited number of the input's adjacent members. LSTM, on the other hand, is suited for processing a series of values. Every member of the trained feature is a function of the output's preceding members (Zhao et al.; 2019).

The combination of CNN and LSTM may be used to learn high-level characteristics that include both local and long-term contextual dependencies.Rather than employing a conventional CNN, the convolutional LSTM (ConvLSTM) is utilized in this context to preserve both spatial and temporal information, such as the spatiotemporal context of successive voice samples (Kwon et al.; 2020). Thus, by combining the CNN and LSTM architectures, hidden patterns are extracted, long-term spatial relationships in speech segments are recognized, and temporal cues are discovered. Thus, the model integrates various CNNs and LSTM networks to extract emotional characteristics by stacking four specified LFLBs and other building layers. Speech is a time-varying signal that requires specific processing to accurately represent its time-varying characteristics. As a result, a layer of LSTM is added to extract long-term contextual dependencies.

The purpose of the 1D CNN LSTM network is to identify speech emotions from audio clips; the purpose of the 2D CNN LSTM network is to learn global contextual information from handcrafted features. To detect speech emotions, the proposed 1D and 2D CNN LSTM models acquire hierarchical local and global parameters. Whereas the majority of data models can extract only low-level features for emotion classification, and the majority of the existing DBN-based and CNN-based algorithmic models can only learn one kind of emotion-related feature for emotion recognition.

5 Implementation

5.1 Baseline Approach

Our baseline SER system consists of the following: The fixed-length MFCCs are calculated from the raw audio sample input. Then, the resultant two-dimensional matrix is sent to a CNN classifier as an input. The maximum utterance length was adjusted to 2.5 seconds with a 0.5-second offset based on the mean duration of all samples having a sampling rate of 44.1 KHz. Thus, the obtained MFCC feature from these raw audio samples was given as an input to the classifier as a 2d matrix. By using the spatial information included in images, CNNs have been utilized to extract high-level features. CNNs are typically constructed by performing multiple convolutions backed by pooling. The convolution layer extracts many local patterns from the input space and returns feature maps using fixed-size kernels.Convolution is accompanied by batch normalization in our model. By normalizing the last layer's output, batch normalization minimizes the internal covariance transition in a feature map. Additionally, this procedure has a regularization impact, which helps to minimize overfitting. Following the batch normalization layer, a max-pooling layer is used to down sample the input feature vectors while retaining their data. By obtaining the largest value in the given window, Max Pooling minimizes the set of parameters for future layers.The block's last layer is a dense layer using SoftMax as the activation function and containing an equal number of nodes as

classes.

5.2 Novel Approach

In this approach, four local feature learning blocks (LFLB) are used. They consist of one convolutional layer, a batch normalization (BN) layer, a max-pooling layer and one activation layer. These LFLBs are used for extracting local features. Also, the additional LSTM layer is used to learn from the sequence of local features and to have long-term dependencies. A combination of LFLB and LSTM are used to determine the local as well as global features from the raw audio files and MFCCs respectively. The Convolution layer, the main layer of the LFLB, is specially used for processing the sequence of grid values. Whereas, the LSTM layer helps in processing the sequence of values. LFLB is used to extract emotional features from audio files. The BN layers are used to normalize the output of each batch of the convolution layer and provide stability along with improved performance. The relu activation layer explains the output of the BN layer.Whereas Max-Pooling layers help to make the features extracted robust against the noise from the audio clips. The 2D CNN-LSTM model is developed by linking these four LFLBs to the LSTM layer and a dense layer at the end. Each LFLB's convolution kernel is the same size, stride one, and same padding. The first and second LFLBs (LFLB1 and LFLB2) have 64 convolution kernels, whereas the third and fourth LFLBs (LFLB3 and LFLB4) have 128 convolution kernels. Each LFLB has a kernel size of four and a stride of four for max-pooling. The top layer of this design is a softmax classifier, which is used to classify emotions based on their learnt characteristics.

6 Evaluation

This section summarizes the major results that support the research question. Numerous tests have been conducted to determine the reliability of the SER models and their suitability for use in real-time applications. The proposed deep learning model is evaluated by running it on different combinations of the four datasets.

6.1 Experiment 1: Baseline Model applied to four datasets (SAVEE, TESS, RAVDESS,CREMA)

The baseline model using CNN was applied to each individual dataset and the accuracy was noted.The baseline model is the simplest model that can be thought of and is built mostly to test the framework and get the path which can be used to build the best possible classifier. The baseline model is the one that can be considered as the basis or support for building the proposed model using that base. Here, the model was run against individual datasets one by one and it was observed that the CREMA dataset performed less as compared to all the three datasets as shown in Figure 6. The TESS dataset model was 96% percent accurate whereas the SAVEE dataset obtained accuracy of 38%.The Figure 5 shows the results obtained over TESS dataset. There was a possibility that this difference in accuracies was due to the differentiation in the gender of the speakers who recorded the speeches as SAVEE dataset consisted recordings by only male speakers and TESS dataset consisdted by only female speakers. This hypothesis was then evaluated in another experiment, which is explained later in section 6.2.

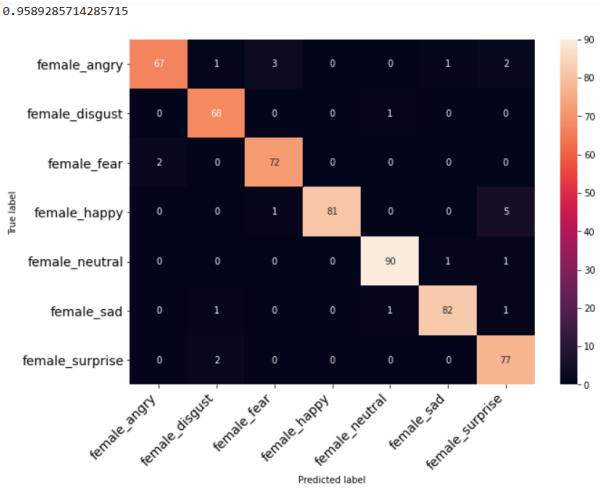


Figure 5: Heatmap showing results for TESS dataset

Also, one more finding from this experiment was that the size of the dataset might the affect the performance of the model.The SAVEE dataset with 480 sample size showed less accuracy as compared to the others. Again, this was another hypothesis which was further tested by experimenting the SER system with unseen dataset having small size of data.The results of this experiment can be found in section 6.4.

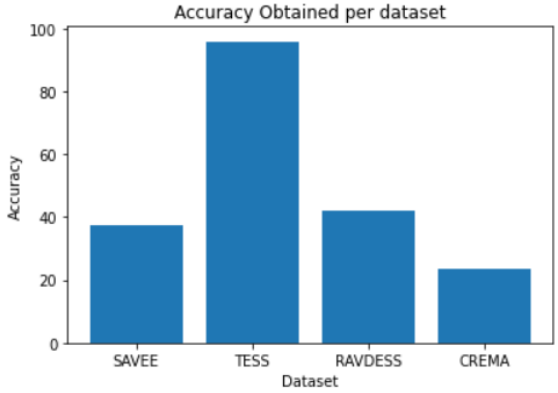
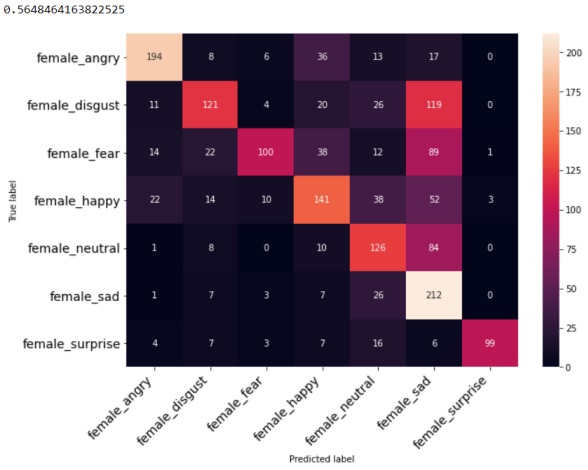


Figure 6: Comparison of baseline model accuracy for each dataset

6.2 Experiment 2: Gender wise Model performance

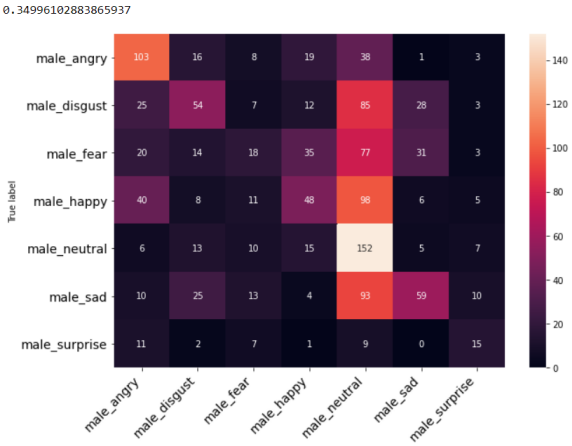
This experiment was conducted in order to verify the findings of the previous experiment where the model performed differently on individual datasets. After evaluating the findings obtained for each dataset, it was determined that the TESS dataset performed well in comparison to the other datasets. This may be explained by the fact that the TESS dataset included only female speaker recordings, while the others included both female and male speakers. To test this hypothesis, we conducted an experiment in which the model was run independently using the recordings of male and female speakers.

Recognizing emotions from human speech is challenging because of the lack of distinct temporal boundaries in human emotions, and each person expresses their feelings differently. Additionally, there are acoustical differences between male and female speakers. Adult males often have a lower pitch in their voices, while females typically have a high-pitched voice (Mishra and Sharma; 2020). Due to this pitch variation, both genders convey the very same emotion in markedly different ways. This can be stated by observing the outcomes of this experiment as shown in Figure 7 and Figure 8, which demonstrated that the same model behaved differently when gender-specific speaker data was used. As can be observed, female speakers convey emotions in a more diverse manner, and the model may quickly acquire the ability to recognize emotions when spoken by females.If a recognition system fails to take this distinction into account, misinterpretation of the audio-emotion may result. For example, the speech of happy men may be mistaken for that of neutral women.



	precision	recall	f1-score	support
female_angry	0.79	0.71	0.74	274
female_disgust	0.65	0.40	0.50	301
female_fear	0.79	0.36	0.50	276
female_happy	0.54	0.50	0.52	280
female_neutral	0.49	0.55	0.52	229
female_sad	0.37	0.83	0.51	256
female_surprise	0.96	0.70	0.81	142
accuracy			0.56	1758
macro avg	0.66	0.58	0.59	1758
weighted avg	0.64	0.56	0.57	1758

Figure 7: Results obtained for dataset with only female speakers



	precision	recall	f1-score	support
male_angry	0.48	0.55	0.51	188
male_disgust	0.41	0.25	0.31	214
male_fear	0.24	0.09	0.13	198
male_happy	0.36	0.22	0.27	216
male_neutral	0.28	0.73	0.40	208
male_sad	0.45	0.28	0.34	214
male_surprise	0.33	0.33	0.33	45
accuracy			0.35	1283
macro avg	0.36	0.35	0.33	1283
weighted avg	0.37	0.35	0.33	1283

Figure 8: Results obtained for dataset with only male speakers

6.3 Experiment 3: 2D CNN-LSTM Model performance on merged four datasets

The baseline model helped to develop a working SER system and had about 43% validation accuracy. This model was built to get an idea of how deep learning techniques work and not for evaluating the performance against speech data. So the need for a better performing SER model led to this experiment of developing a robust emotion detection system. Thus, after studying the literature work in this area and conducting several experiments, a novel SER system was proposed using a hybrid model of 2D CNN and LSTM. This hybrid model showed an improvement over the baseline model of more than 25% in validation accuracy and achieved 95% training accuracy. The evaluated results are plotted using a heatmap as shown in Figure 9

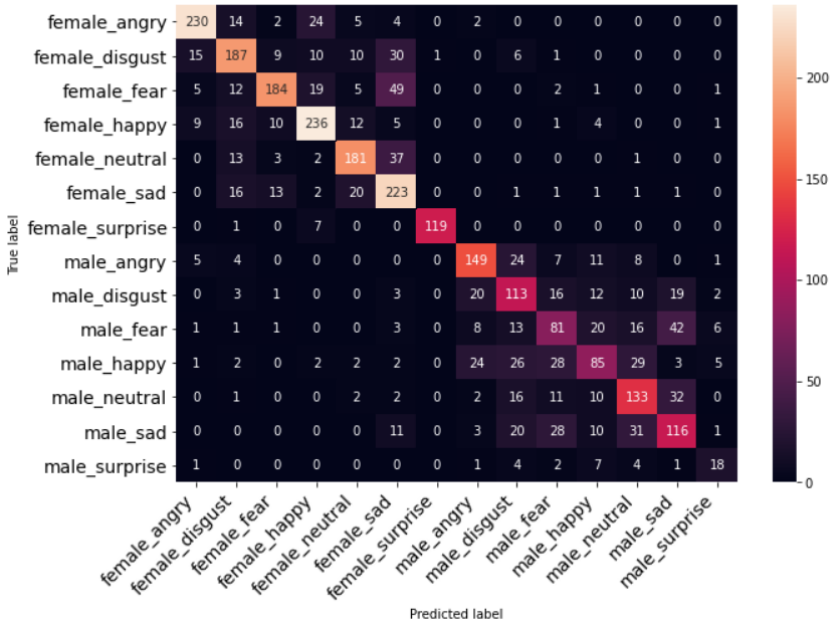


Figure 9: Evaluated results of SER model using 2D CNN-LSTM

6.4 Experiment 4: Model Performance on Unseen dataset

The proposed model was trained on data from four different datasets. Many researchers have trained SER models on different datasets individually and have achieved higher accuracies, but when such models are given an input of an unseen dataset, they have not provided good results. In order to overcome this issue,the SER model was provided with an unseen dataset (EMO-DB) in this experiment.The EMO-DB dataset is a publicly accessible emotional database in German. The Institute of Communication Science at Berlin's Technical University developed the database. Ten professional speakers (five men and five women) took part in data collection. There are a total of 535 utterances in this database distributed in seven emotions. This German database against the proposed model was able to achieve 100% training accuracy and 71.28% validation accuracy.The results obtained are shown in Figure 10

The final evaluation of the model was done by using a speech recorded by myself and was provided as an input to the saved model.The wave plot for the same is shown

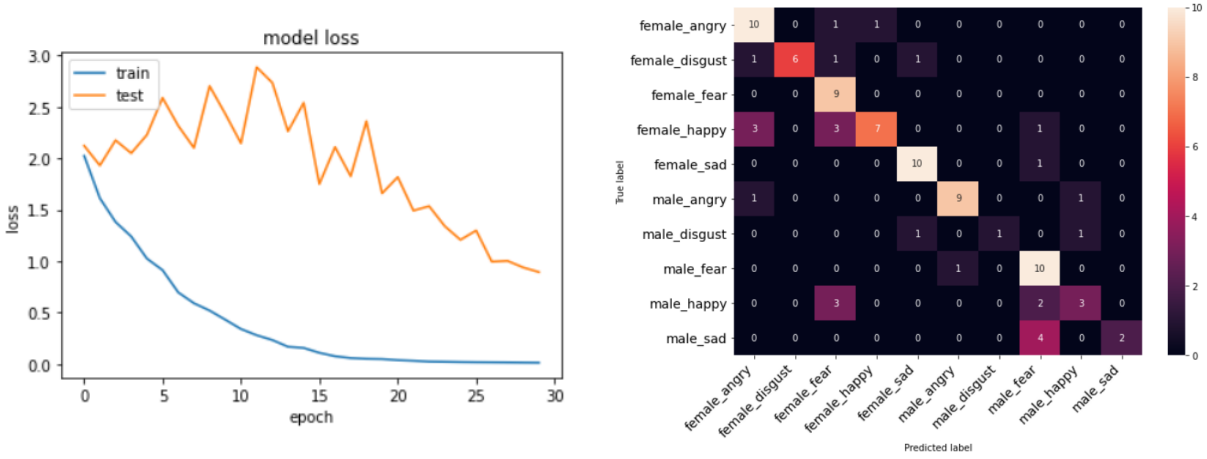


Figure 10: Evaluated results obtained for EMO-DB dataset

in Figure 11 This was a statement with angry emotions and was predicted correctly by our classifier. This experiment helps to find an answer to the research question. As the SER model using deep learning can classify emotions on practical data, such models can further be improved and used in future in real-time applications to detect emotions.

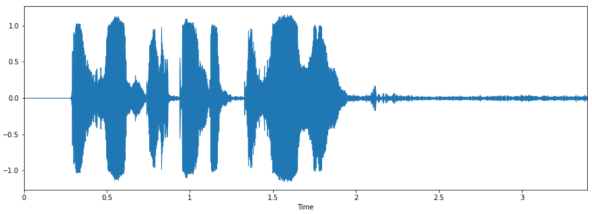


Figure 11: Wave Plot of Audio Recorded for testing

6.5 Discussion

All of the above studies yielded important data that may help in establishing the answer to the research question. On various datasets, the SER model produced a range of outcomes based on a number of factors. One might say that the SER model performs significantly differently for female and male speakers. Female speakers are more accurate at expressing emotions than male speakers. To prevent misguided training of the SER model for speakers of various genders, it is essential to train the model using data from both genders. Deep learning methods have shown excellent performance in detecting emotions in recorded speech. The suggested model, which was trained on four merged datasets including over 12000 utterances, successfully identified emotions using a hybrid model of CNN and LSTM with a 67.58% on test data. Additionally, this model obtained 71.28% accuracy on the previously unknown EMO-DB dataset. This dataset was in German, demonstrating that the SER model performs well regardless of the language of the speech. Additionally, the model was capable of correctly detecting emotions in speech recorded by my own voice. All of these findings contribute to the SER model

being considered trustworthy and capable of being utilized in real-time applications.

7 Conclusion and Future Work

This paper presents a novel approach to emotion recognition using 2D CNN LSTM networks. The method of feature extraction from raw audio clips is being investigated in this research. The role of MFCC features in extracting emotional elements from raw speeches is discussed and used as an input to the deep learning model. Using the proposed SER model, multiple experiments are carried out that help in determining the reliability of SER in real-time scenarios. The results obtained of 71% validation accuracy by running the SER model on unseen data in the German language help to state that this model can work on any real-time data irrespective of the language. Also, by involving multiple datasets with multiple speakers, the SER model has been trained in such a way that if given an input irrespective of the gender of the speaker, the model can accurately classify the emotion of the recorded speech.

Such a system can further be used as a baseline approach for developing numerous applications in the entertainment sector.This model can be further improved by using different combinations of neural networks as well as including data augmentation techniques. The model can be trained with audio files having ample background noise such that the emotions can be classified easily from the real time data.

Acknowledgements: The Authors want to express their gratitude to the National College of Ireland for providing the essential research facilities and required infrastructure.

References

Abbaschian, B. J., Sierra-Sosa, D. and Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models, *Sensors* **21**(4): 1249.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A. and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset, *IEEE transactions on affective computing* **5**(4): 377–390.

Chen, C. and Li, Q. (2020). A multimodal music emotion classification method based on multifeature combined network classifier, *Mathematical Problems in Engineering* **2020**.

Chernykh, V. and Prikhodko, P. (2017). Emotion recognition from speech with recurrent neural networks, *arXiv preprint arXiv:1701.08071* .

Dupuis, K. and Pichora-Fuller, M. K. (2011). Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set, *Canadian Acoustics* **39**(3): 182–183.

Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L. and Schmauch, B. (2018). Cnn+ lstm architecture for speech emotion recognition with data augmentation, *arXiv preprint arXiv:1802.05630* .

Florence, S. M. and Uma, M. (2020). Emotional detection and music recommendation system based on user facial expression, *IOP Conference Series: Materials Science and Engineering*, Vol. 912, IOP Publishing, p. 062007.

Fu, C., Liu, C., Ishi, C. T. and Ishiguro, H. (2021). An end-to-end multitask learning model to improve speech emotion recognition, *2020 28th European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 1–5.

Hizlisoy, S., Yildirim, S. and Tufekci, Z. (2021). Music emotion recognition using convolutional long short term memory deep neural networks, *Engineering Science and Technology, an International Journal* **24**(3): 760–767.

Iqbal, A. and Barua, K. (2019). A real-time emotion recognition from speech using gradient boosting, *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, pp. 1–5.

Issa, D., Demirci, M. F. and Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks, *Biomedical Signal Processing and Control* **59**: 101894.

Jackson, P. and Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database, *University of Surrey: Guildford, UK* .

Kowtha, V., Mitra, V., Bartels, C., Marchi, E., Booker, S., Caruso, W., Kajarekar, S. and Naik, D. (2020). Detecting emotion primitives from speech and their use in discerning categorical emotions, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7164–7168.

Kwon, S. et al. (2020). Clstm: Deep feature-based speech emotion recognition using the hierarchical convlstm network, *Mathematics* **8**(12): 2133.

Langari, S., Marvi, H. and Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction, *Informatics in Medicine Unlocked* **20**: 100424.

Liu, C.-L. and Chen, Y.-C. (2018). Background music recommendation based on latent factors and moods, *Knowledge-Based Systems* **159**: 158–170.

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, *PloS one* **13**(5): e0196391.

Mishra, P. and Sharma, R. (2020). Gender differentiated convolutional neural networks for speech emotion recognition, *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, IEEE, pp. 142–148.

Panda, R., Malheiro, R. and Paiva, R. P. (2020). Novel audio features for music emotion recognition, *IEEE Transactions on Affective Computing* **11**(4): 614–626.

Pandey, S. K., Shekhawat, H. and Prasanna, S. (2019). Deep learning techniques for speech emotion recognition: A review, *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, IEEE, pp. 1–6.

Patel, N., Patel, S. and Mankad, S. H. (2021). Impact of autoencoder based compact representation on emotion detection from audio, *Journal of Ambient Intelligence and Humanized Computing* pp. 1–19.

Sönmez, Y. Ü. and Varol, A. (2020a). In-depth analysis of speech production, auditory system, emotion theories and emotion recognition, *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE, pp. 1–8.

Sönmez, Y. Ü. and Varol, A. (2020b). A speech emotion recognition model based on multi-level local binary and local ternary patterns, *IEEE Access* **8**: 190784–190796.

Wang, Y., Liang, W., Li, W., Li, D. and Yu, L.-F. (2020). Scene-aware background music synthesis, *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1162–1170.

Zhang, B., Quan, C. and Ren, F. (2016). Study on cnn in the recognition of emotion in audio and images, *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, IEEE, pp. 1–5.

Zhao, J., Mao, X. and Chen, L. (2019). Speech emotion recognition using deep 1d & 2d cnn lstm networks, *Biomedical Signal Processing and Control* **47**: 312–323.