# Convolutional Recurrent Neural Network for Speech Emotion Recognition

MSc Research Project
Data Analytics

Aggarwal Aditi
Student ID: x18137156

School of Computing
National College of Ireland

Supervisor: Prof. Christian Horn

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Aggarwal Aditi |
| **Student ID:** | x18137156 |
| **Programme:** | Data Analytics |
| **Year:** | 2019 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Prof. Christian Horn |
| **Submission Due Date:** | 12/12/2019 |
| **Project Title:** | Convolutional Recurrent Neural Network for Speech Emotion Recognition |
| **Word Count:** | 5994 |
| **Page Count:** | 21 |

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:** | |
| **Date:** | 11th December 2019 |

## 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): | |

# Convolutional Recurrent Neural Network for Speech Emotion Recognition

Aggarwal Aditi
x18137156

**Abstract**

Speech emotion recognition (SER) has been an influential subject of research in human-machine interaction over the past decade. This subject has gained immense attention due to the fact that the human voice is a primary form of expression which reveals the mental state of the speaker and has a broad range of emotions associated with it. Deep learning techniques have been widely used with great success to design the SER system but the techniques are severely restricted due to the degradation problem and information loss in the high layer of deep neural networks. This research has proposed a deep learning model, convolutional layer embedded with a recurrent neural network (CRNN), which addresses the above-stated concerns and enables to capture of both frequency and temporal dependence. Segment level features have been extracted from the waveform to preserve the time relations through the sequence of frames and contextual information is captured by CNN. A simple recurrent unit, Long Short-Term Memory (LSTM), aggregates these frame-level features and the emotional classes are identified using the SoftMax classifier. The model is tested on the RAVDESS dataset for seven classes of emotions and the experimental results demonstrate that the proposed model can effectively determine the emotions contained in the speech.

## 1 Introduction

Speech is the most natural and fastest mode of communication between humans. This inspired researchers to believe that speech is an efficient method for human-machine interaction. The subject of speech recognition which enables us to convert speech into a sequence of words has made tremendous advancements over the years. Despite the immense growth made in speech recognition, machines are still distant from having a natural interaction with humans. A speech signal is enriched with valuable information, not just the semantic message. The sound generated with the semantic message in the speech communicates the emotional state of an individual. The process of classifying speech into different classes of emotions is well-known as speech emotion recognition. There are many applications like in-car systems to ensure the driver's safety based on mental status, a diagnostic tool for therapists, chatbots as well as call centers where call escalation is dependent on the emotional state of a speaker. This field has been researched for quite a time now and has revealed the valuable role played by emotions in shaping social interactions. But people suffering from hearing loss who fail to have such natural conversation. Therefore, these trained SER systems can be used by such an audience which would enable them to feel and understand the tonal emotions of other humans in silence.

The World Health Organization (WHO) [1] has estimated that 34 million children and 432 million adults have disabling hearing loss all over the world. It has also been predicted that this number will be doubled to approximately 900 million by the year, 2050. This disability isolates the deaf or hard of hearing (DHH) audience from Voice over Internet Protocol (VoIP) and fails the access to services of voice calls and voice messages effectively. This exclusion from communication has a powerful impact on their everyday life. WHO has also estimated the annual global cost of 750 billion US dollars posed by unaddressed hearing loss. This cost includes educational support, productivity loss, and health sector costs. Children with hearing loss rarely receive the full benefit of education, adults have higher unemployment rates and employed adults with hearing loss have lower performance grades. Automatic speech recognition (ASR) has been of great help for the DHH audience. This motivates to deliver the emotional content of human voice which would enable access to complete information conveyed through speech.

The approach for designing the SER model involves two steps: feature extraction and classification. Researchers have derived several features such as human factor cepstral coefficients (HFCC), Mel-frequency cepstral coefficients (MFCC) and Gammatone frequency cepstral (GFCC) (Sugan et al. (2018)), and also the fusion of multiple features with an intention to design an optimal model. The most valuable speech representative used in most of the research works is MFCC. Previous studies have agreed on the fact that local frame-level features contribute more than global utterance level features due to the loss of temporal information (Mirsamadi et al. (2017), Ayek et al. (2017)). The second step is classification which comprises two domains: linear and non-linear. Since the sound signal is non-stationary in nature, non-linear classifiers work better when compared to linear classifiers (Nassif et al. (2019)). Deep learning techniques evolved from the subject of machine learning have made extensive advancements in the field of speech processing. These techniques have outperformed the traditional systems due to their capability to detect complex structures. Some of the fundamental deep learning techniques are Deep Belief Network (DBM), Recurrent Neural network (RNN) and Convolutional Neural Network (CNN). The theoretical evaluation of recent research works has shown the potential of these techniques and has inspired this research to choose deep neural networks. This research propose a speech emotion recognition model based on a convolutional recurrent neural network which aims to classify seven different classes of emotion. This model could address the above-stated concern and enable the DHH audience to sense the emotional state of the speaker.

The rest of the research work is ordered as follows. Section 2 summarizes the previous research conducted on the subject of emotions, human voice, feature extraction, and classification. Section 3 details the approach used for this research and discusses the datasets and methods. Section 4 is a thorough description of the experimental setup and model design. Section 5 reports the experiments and observations and Section 7 sums up the research work.

[1] WHO Estimates: https://www.who.int/deafness/estimates/en/

# 2 Related Work

In speech emotion recognition systems, the feature extraction process has a central role in the contribution to better performance. In the following section, we review some key research related to speech emotion signals (subsection 2.1), feature extraction (subsection 2.2), and classification methods (subsection 2.3).

## 2.1 The emotional power of Audio Signal:

In layman's terms, speech is a sequence of sounds used to express feelings and thoughts. A speech signal carries information about the speaker identity, gender, message to be conveyed, language identity, and emotion. The fact that a speech signal is the most natural and preferred mode of communication has persuaded researchers to believe speech as a powerful mode of interaction between machines and humans (Eduard Frant and Stoica (2017))(Schuller (2018)). Tremendous research has been carried out on the subject of speech in the last two decades and great advancements have been made in speech recognition (Nassif et al. (2019)) but the task of speech emotion recognition still has challenges that need to be researched more in order to resolve them. The level of complexity that comes with a sound, or the power of human voice can be best defined by the fact that changing tonal attributes of voice can change the meaning of the same set of words. Therefore, voice emotion detection prevails to be a valuable subject of research.

Over the past years, psychologists have formulated a lot of theories on emotions, the two theories introduced on the structure of vocal emotions are the dimensional model and the discrete or categorical model. The discrete model is designed from a definite set of emotions such as sad, happy and fearful, which assumes to have sharp boundaries between the categories (Ekman (1992)). But (Ronan et al.; 2018) supported the argument made in several research works that defining a set of basic emotions seems deficient to express the richness of emotions. This draws attention to the dimensional structured model which deals with the affective states along broad dimensions. It is widely thought that emotion is classified into two dimensions: valence and arousal. Russell (1980) describe valence on a positive and negative continuum, whereas arousal on a high and low scale in the state of emotion. Since emotions could not be well classified in two-dimensional space, Wilhelm Max Wundt (Sarprasatham (2015)) proposed a three-dimensional model with the third dimension as tension measured on a tense and relaxed scale. Simultaneously, Fontaine et al. (2007) proposed a four-dimensional model described by two new dimensions: intensity and potency, with valence and arousal. But dimensional models were also argued for not making the clear distinction between certain emotions such as fear and anger. There is no universal consensus to measure the emotions in the field of psychology as emotion is the most biased aspect based on personal and cultural differences. In a very recent study, Cowen et al. (2018) introduced a new hypothesis, which claims that listeners can perceive more than 20 emotions in wordless noises. The authors tested their theory with three different approaches. An audio library comprised of more than 2000 sounds was evaluated with the help of 1000 people. To find the commonalities, authors merged the free-choice and forced-choice descriptions and found 24 reliable emotional categories which were plotted on an interactive map[2]. The authors were also keen to find out the existence of borders between emotions but ended up discovering that borders were fuzzy

[2] Interactive Map: `https://s3-us-west-1.amazonaws.com/vocs/map.html#modal`

and sound conveys mixed emotions too. This necessitates the need to explore and understand the complex subject of emotions in order to design an optimal SER system. The proposed model in this study intends to study the idea behind the incorrect classification of different emotional classes.

## 2.2 Feature Extraction:

An essential step in the design of a SER system that efficiently classifies different classes of emotions is the extraction of valuable features. There are two major challenges in the process of feature extraction. The first challenge is to determine the appropriate region of analysis, global or local features. Some studies are based on extraction of global statistics from the entire utterance and some follow traditional framework approach which divides the signal into smalls chunks and extracts local features from each chunk. Earlier researchers had a disagreement on which of global and local features are more essential to design a SER model. But the fact that audio is not stationery in nature has convinced the authors to agree that local features are superior to global ones. Mirsamadi et al. (2017) and Ayek et al. (2017) tested their SER models on both local and global features to compare the performance. Local features were extracted from multiple frames, each of 25ms segments, fetched from an utterance and global ones were fetched from the whole utterance. Both the studies have achieved higher accuracy when features were extracted from segments instead of whole utterance. The idea of global features failed because the essential temporal information enclosed in an audio signal gets completely lost (Ayadi et al. (2011)).

The second challenge is to choose the relevant features from an audio (Swain et al. (2018)). Features used in the SER model are broadly classified into two types, one is spectral and another one is temporal. Spectral features are short time demonstrations for an audio signal, it is observed that the emotional content of an audio clip has an impact on spectral energy across the frequency range. Whereas, temporal features are based on the time domain. Sugan et al. (2018) compared the performance of three cepstral features: human factor cepstral coefficients (HFCC), Mel-frequency cepstral coefficients (MFCC) and Gammatone frequency cepstral (GFCC). The proposed model was trained on these features individually and tested on two datasets. There was not much variation in the accuracy while testing a model on all the features individually. The achieved results were comparable and have a variation of 1% or 2%. This study suggests that these three features provide proportionate performance. Han et al. (2014) proposed a deep neural network model where three kinds of features were extracted: delta feature, pitch-based, and MFCC, across time. The proposed model attained 54.3% of accuracy. Zhao et al. (2017) argued the research by highlighting the potential of the MFCC feature alone. The authors proposed a model with one recurrent convolutional layer and two MLP layers. The model was trained on MFCC features and attained a comparable accuracy of 53.6%. Demircan and Kahramanli (2014) have drawn the comparison between three features: pitch-based, formant frequency, and MFCC. Based on the applied experiments, the authors stated that MFCC provides the best representation of audio signals when compared to the other two features. Pitch is thought to be the best indicator in classifying happy and neural emotions due to the huge difference in their pitch frequencies. They also observed that formant frequency satisfactorily classifies the happy emotion but fails to classify angry and neutral emotion. The study shows that different features perform differently with each emotion. This confusion of picking the appropriate features can be slightly managed by applying dimensionality reduction techniques. These techniques provide an insight into the distinguishing features and reduce the computation requirements.

## 2.3 Deep learning techniques:

The field of Artificial Intelligence is wide and has been around for a long time. Machine learning is a subfield of AI and Deep Learning is evolved from the family of Machine Learning that has artificial neural networks concerned with the function and structure of the brain. Deep learning techniques in the subject of speech emotion recognition (SER) have attracted a lot of researchers, some have used the deep learning techniques to implement their SER models that performed better when compared to traditional methods (Nassif et al. (2019)).

Zhao et al. (2017) designed a model for phoneme recognition and SER. Three SER models were developed for comparison and loaded with 25ms frame segments of audio-visual data with their respective labels. The model was tested on IEMOCAP dataset, The Interactive Emotional Dyadic Motion Capture (Busso et al. (2008), the data was filtered to 5300 utterances with 5 different emotion labels. The third model, Recurrent Convolutional Neural Network (RCNN), with two fully connected hidden layers and a recurrent convolutional layer performed the best amongst the three models with a weighted accuracy of 53.6%. Similar research with the same audio-visual data and emotional classes was carried out by Microsoft (Han et al. (2014)) where pitch based features with spectral features were used to attain the accuracy of 54.3% using deep neural network model and the research claimed that only spectral features cannot provide satisfactory results. The claim was argued by Zhao et al. (2017) by providing comparable results using only spectral features with the RCNN model. In the same year as Zhao et al. (2017) research work, Microsoft published another research (Mirsamadi et al. (2017)) and proposed an automatic SER system using Recurrent Neural Network (RNN) with local attention. The model was tested on the same IEMOCAP dataset but for four classes of emotions. The experiment, RNN with weighted pool attention, achieved 61.8% with raw spectral features and 63.5% with Low-Level Descriptors (LLDs) extracted from 20 to 50ms of short frames. Ayek et al. (2017) evaluated multiple deep learning methods for the SER system on the same dataset with five emotional classes and achieved 64.78% (frame-based). The best output was achieved using the model designed with 2 convolutional layers and 2 fully connected layers. Both the research works (Mirsamadi et al. (2017)) and Ayek et al. (2017)) demonstrated that frame-based featured models perform better than utterance-based featured models. In parallel with testing models on IEMOCAP data, Badshah et al. (2017) proposed a deep convolutional neural network model (DCNN) to detect emotions from spectrograms obtained from speech signals. Fast Fourier Transform (FFT) is applied to small chunks of spectrograms that were loaded in a DCNN system. The model was built with a softmax layer, three fully connected layers, and three convolutional layers. The model was tested on the Berlin Emotional Database (Emo-DB) (**?**) with seven emotion classes and attained 56.1% accuracy where fear, happy, and neutral emotions were poorly classified.

Chen et al. (2018) also proposed a three-dimensional attention convolution RNN (CRNN) SER model. This model passes short frames of 25 ms with deltas and deltas-deltas Mel-spectrogram features to convolutional recurrent network, attention layer, fully connected layer, and output is obtained by softmax classifier. The model was tested on two audio datasets IEMOCAP for four emotion classes and Emo-DB for seven emotion classes and achieved 64.74% and 82.82% accuracy respectively. Zhao et al. (2018) suggested another approach to test on the same datasets as used by Chen et al. (2018) except both datasets were tested on seven emotional classes. The new method, merged CNN, consists of two branches. One branch is 1D CNN to learn deep features from audio clips and another is 2D CNN to learn high-level features from Mel spectrograms. Both branches are merged and classified using softmax. This method performed better than Chen et al. (2018) model and attained 86.36% and 91.78% accuracy on IEMOCAP and Emo-DB dataset respectively.

At the same time, neural network techniques were also applied to other available datasets. Jannat et al. (2018) implemented the CNN model on RAVDESS audio-only dataset, The Ryerson Audio-Visual Database of Emotional Speech and Song. The CNN model was only trained on 2 classes of emotions: Happy and Sad, and achieved 66.41% of accuracy. Gumelar et al. (2019) implemented a deep neural network model RAVDESS dataset with five different emotional classes. The author passed spectral and prosodic features as input to a two-layered network comprised of a convolutional and a max-pooling layer. The model achieved 78.83% of accuracy. Deep learning techniques applied in above research works and their performance has greatly inspired the proposed model of this research.

# 3 Research Methodology

An audio signal is an essential source for emotional expression and the human voice is enriched with emotional information. Speech emotion recognition will become an essential component of HCI but there are still many concerns that need to be addressed to obtain an optimal SER system. Deep learning techniques are currently considered as a flourishing research subject for processing speech signals. This research study aims to explore the potential of the deep neural technique to design an SER model and also to understand the puzzle of emotions. The proposed research has followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Wirth and Hipp (2000)) approach to attain the goal. The proposed SER system comprises of three vital tasks. First is the data collection and preparation. The second step is the extraction of speech representatives. These features are imported to the machine learning classifier for emotion recognition. The block diagram of the proposed model is described in Figure 1. and the approach followed for this research has been outlined below.

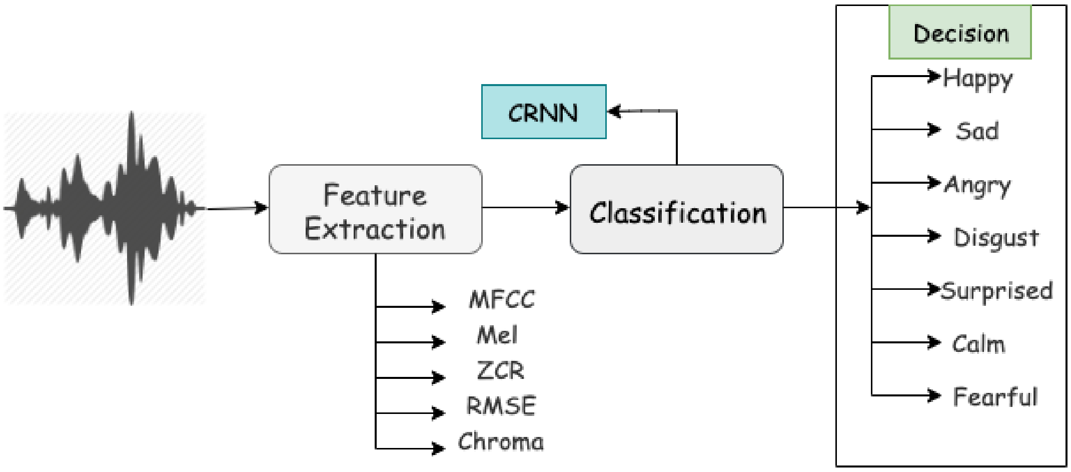


Figure 1: Block Diagram

## 3.1 Data Preparation

### 3.1.1 Data set

Speech data can be easily fetched from any source such as YouTube, movies, TV shows, etc but the SER model demands reliable and high-quality data. The choice of a dataset has an impact on the performance of the model. This research has extracted the validated audio dataset from The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[3]. Livingstone and Russo (2018) prepared this dataset with the help of 24 professional actors with balanced gender, 12 female and 12 male actors. The vocals are recorded in a neutral North American accent. The speech dataset comprises eight different emotions: happy, surprised, angry, disgust, calm, sad, fearful and neutral. This RAVDESS dataset was rated 10 times by 247 untrained individuals on intensity, emotional validity and genuineness to ensure the quality. The dataset is available in three different formats: audio only, video only and audio-visual. This work has extracted 1440 speech clips where each expression was recorded at two different levels of emotional intensity: normal and strong. Neutral emotion does not have recordings with strong intensity. Two different statements are used to record each of the emotions. Each audio file has been named with a unique numerical identifier that defines the audio characteristics. This research experiments the model on seven classes of emotion. The intention to choose dataset with high number of emotional classes is motivated with the hypothesis introduced by Cowen et al. (2018) which states that people can perceive approximately 20 emotions. This inspired to explore large number of classes as there is a long way to meet an optimal solution classifying all emotions.

### 3.1.2 Data Pre-processing

Audio is one-directional in nature and transmitted as sound waves. The foremost step is to feed the sound signal into the model. The sampling method is used to convert the audio signal into a digital signal. The raw speech data is sampled at the rate of 22.05 kHz which is sufficient to cover the human speech frequency range. There are 1440 audio

[3] Dataset: `https://zenodo.org/record/1188976#.Xe04R5P7RN0`

clips with variable duration time. Those clips are transformed into equal lengths which enables us to divide the signal into equal-sized frames to fetch local features. The thought to fetch local features for this model is inspired by two research studies (Mirsamadi et al. (2017) and Ayek et al. (2017)) where authors demonstrated the power of local features when compared to global ones. Then the speech signal is split into frames and each frame has an overlap of 50% samples from the previous frame in order to ensure that no information is lost as shown in Figure 2. This step facilitates to fetch the frame-level features from each utterance. After preparing the audio clips with the above-mentioned steps, below mentioned features are extracted for this research.

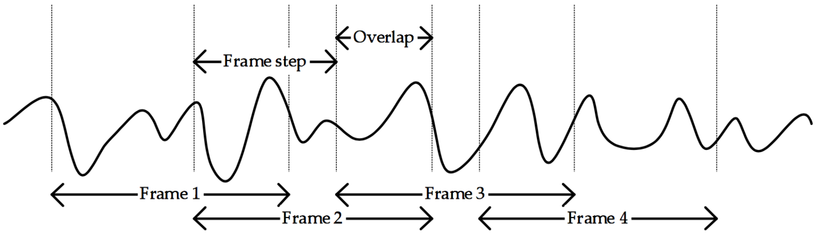


Figure 2: Framing of audio signal

### 3.1.3 Speech Representations

The next crucial step for data pre-processing is feature extraction. This step is essential for audio signals as these cannot be inserted directly into the model. This study aims to extract five features that intend to gather all the characteristics of emotion from a voice signal. The collection of these five features acts as a representative of raw audio information to the model. The below-mentioned features are fetched from each frame of an audio signal:

**Mel Frequency Cepstral Coefficient (MFCC):** This feature is the most used representation of spectral characteristics of sound signals. The potential of this feature is best described by Demircan and Kahramanli (2014) and Zhao et al. (2017) where MFCC alone is considered to be superior to other features. This calculates the energy spectrum and Fourier transforms for each frame. Mel-frequency scale is used to map these values. Then MFCC is estimated by applying discrete cosine transform (DCT) on Mel log energies. These coefficients best represent the way sound is heard by the human ear. In this study, 40 MFFCs are computed over 104 frames of each audio clip.

$$\text{Mel}(f) = 2595log(1 + \frac{f}{700}) \tag{1}$$

**Mel Spectrogram:** A non-linear transformation of frequency scale results in the Mel scale, which is on the y-axis of a spectrogram. The reason to convert the frequency scale to the Mel scale is that the frequency scale understands the difference between 500 Hz and 1000 Hz but fails to notice the difference between 7500 Hz to 8000 Hz.

**Chroma:** This feature represents the tonal or harmonic information of voice signals. Chroma uses the magnitude (energy) spectrum instead of the power spectrum and it associates with 12 different classes of the pitch. This feature has shown the promising performance when used along with MFCC justified by Han et al. (2014).

**Root Mean Square Energy (RMSE):** This feature is average power or energy for each frame of audio. RMS is found between the quietest sections and loudest sections (the peak, 0dB) of an audio signal.

**Zero-Crossing Rate (ZCR):** This feature is a rate of change in sign across the signal from positive to negative or vice-versa. The is the number of times signal passes the horizontal axis.

The last two features are experimented in this research to evaluate their potential to represent a speech signal.

## 3.2 Modeling

Deep learning methods perform better than traditional methods as conveyed by Nassif et al. (2019). This has motivated to implement deep neural networks for this study. The research work carried out by Chen et al. (2018) proposed a CRNN model and the model was tested on local features. The performance attained by the model when tested on the Emo-DB dataset was exceptional which has inspired the model design of this research. This study has proposed a Convolutional Recurrent Neural Network (CRNN) model which extracts the high-level features for speech emotion recognition. The convolutional layers extract the local information and the recurrent layer combines this over a longer temporal context. The type of recurrent layer used here is Long Short-Term Memory (LSTM). This layer consists of a gating mechanism and memory cells and is highly efficient in learning long term dependencies. The recurrent structure of the model makes it capable of learning and utilizing the features from the input which enables it to overcome short-term deficiencies. The intention behind choosing CNN LSTM architectures is that they enable an analysis of inputs over longer periods when compared to RNN architectures. Extensive research carried out on previous studies has motivated to implement this CRNN architecture for designing the SER model.

## 3.3 Evaluation

The performance of the proposed model is evaluated based on four metrics derived from the confusion matrix. This matrix provides a summary of prediction and can accommodate more than two classes. It also specifies the errors made instead of just providing the insights of errors. The outcome of the model is compared with previous research works. The model performance was also tested based on the above metrics for different frame sizes: 25ms, 50ms and 100ms, and a different sets of emotions.
**Accuracy:** Number of correct predictions over total predictions made by the model. This metric is reliable when target classes are nearly balanced.
**Recall:** This metric is the percentage of total relevant outcomes. Higher values convey a larger number of true positives.
**Precision:** This metric is the percentage of relevant outcomes. Higher values convey a smaller number of false positives.
**F-measure:** This score presents both Recall and Precision. This metric uses harmonic mean to put emphasis on positive counts.

# 4 Design Specification and Implementation

## 4.1 Data Exploration

Dataset contains 1440 audios for eight different emotions. The distribution of different emotional classes is demonstrated in Figure 3 which reveals the class imbalance problem in the dataset. The total number of audios for neutral emotion is far less than the number of audios of other classes. The audio clips of neural emotion have been removed in order to balance the classes and further steps were processed on 1344 audio clips for seven different emotions. Each emotion is recorded by 24 actors for two intensities, two utterances and each is recorded twice.

$$\text{Total Clips} = \underbrace{8}_{\text{emotions}} \times \underbrace{2}_{\text{utterances}} \times \underbrace{2}_{\text{intensities}} \times \underbrace{2}_{\text{repetitions}} \times \underbrace{24}_{\text{actors}} \tag{2}$$

$$\text{Total Clips} = 24 \times (64 \text{ - } 4) = 1440 \tag{3}$$

Note: Neutral emotion is recorded only with normal intensity.

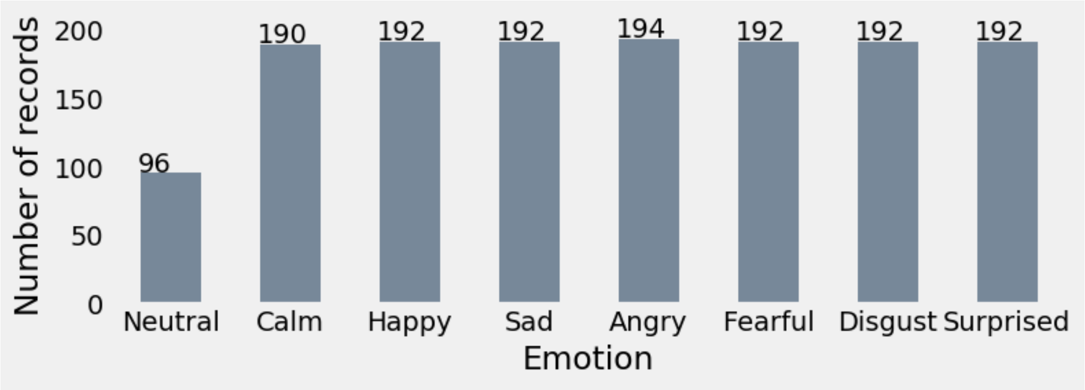


Figure 3: Class Distribution

Table 1 shows the plots of eight classes of emotions with the same semantic message recorded by the same speaker. The raw audio signal can be demonstrated as a wave plot shown in Table 1(a). Each audio clip has been recorded at the sampling rate of 22.05 kHz which signifies that a second of the audio clip has 22050 samples. The first column (a) shows the change in amplitude or air pressure at discrete moments in time. The second and third plots in Table 1(b)(c) are the speech representatives: MFCC and Mel spectrogram, extracted from the raw signal. The color in both plots in columns (b) and (c) defines the volume, darker shades indicate the low volume, pauses or noise and brighter shades indicate speech or loud sound. The fourth plot in Table 1(d) is an energy plot which indicates the level of loudness of sound. High frequencies in the fourth column (d) indicate loud sounds and low frequencies indicate soft sounds.

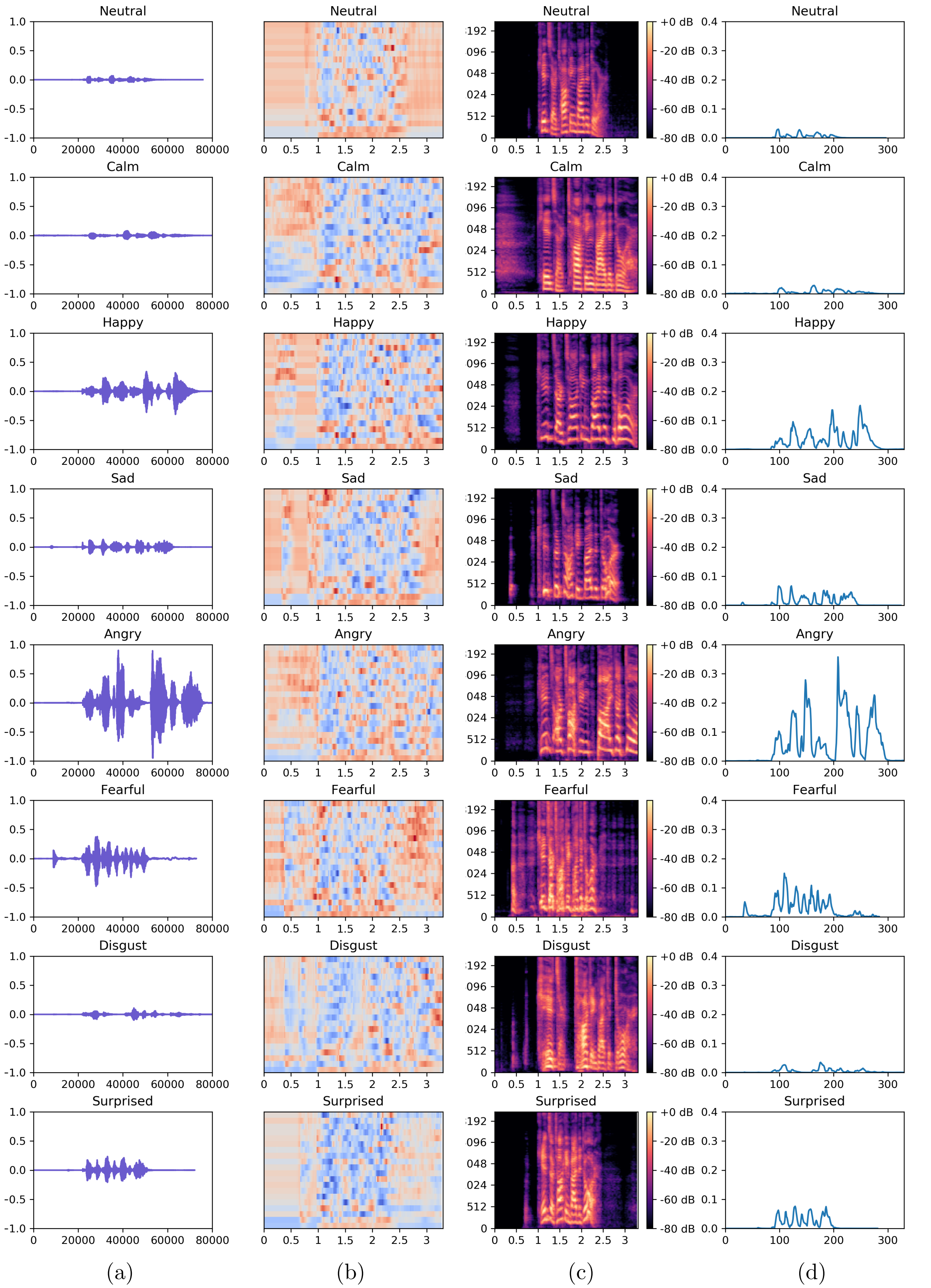


Table 1: Visual representations of eight emotional states in the RAVDESS dataset. Visualizations: (a) Raw signal waveform (amplitude vs time) (b) MFCC (MFFC coefficients vs time) (c) Mel spectrogram (log of frequency vs time and color axis is amplitude transformed to decibels) and (d) Energy (energy vs time).

## 4.2 Padding and Framing

The clips are recorded at a sampling rate of 22.05kHz. Those clips are transformed into equal lengths of 5.3s, and zero paddings are applied for the files with a duration of less than 5.3s. Then the speech signal is split into frames of three different sizes as described in Table 2 and each frame overlaps half of the samples from a previous frame.

Table 2: Framing

| **Frame Size (in ms)** | **Samples per frame** | **Number of frames per clip** | **Total frames** |
|---|---|---|---|
| 25 | 552 | 422 | 567168 |
| 50 | 1104 | 210 | 282240 |
| 100 | 2208 | 104 | 139776 |

## 4.3 Feature Extraction

The five vital speech representatives discussed in subsection 3.1 are extracted from each frame of the signal.

- 40 elements of MFCCs
- 12 element chroma feature vector
- 128 elements of Mel
- One feature element from ZCR
- One feature from RMSE

A feature vector of 182 elements is used to train the model.

## 4.4 Labels Encoding

The labels (classes of emotion) are categorical and do not have any relations based on number series. To ensure that the model does not misunderstand the numerical identifier of emotional class, One hot encoding is applied on labels data which represents the obtained categorical variables as binary vectors.

## 4.5 Train/Test Split

The dataset is split into two sets: train set with 75% of data and test set with 25% using the hold-out method which ensures that both datasets are properly separated. The proposed model is trained on 1008 audio files and tested on 336 files which implies one test per emotion.

## 4.6 Model Design

This research has proposed a convolutional recurrent neural network that consists of 1D convolutional layers, batch normalization layer, 1D max-pooling layers, dropout layer, LSTM layer, and a dense layer. The series of the layers have been discussed below:

*Input:* Frame-level features are inserted into the model after pre-processing the audio signal.

*Convolution:* This model uses 56-time filters with a kernel size of 5 in order to extract features from each frame of the audio signal. The output of this layer is determined by a rectified linear unit (ReLu).

*Batch normalization:* This layer standardizes the input layer and momentum value is set to 0.8 for this model. High momentum leads to slow training. This method also speeds up the neural network training.

*Max-pooling:* This layer down-samples the input representation and is operated on individual features maps which reduces the amount of computation and parameters in the network. The pool size is set to 2.

*Dropout:* This layer prevents overfitting of the model by randomly dropping some number of layer outputs to train the model with a different sets of values. This value is set to 0.1 for this model.

The series of Convolution, Batch normalization and then Max-pooling is implemented three times.

*LSTM layer:* This layer captures the contextual information and dimensionality of the output space is set to 96. The activation function is set to default (tanh). This layer is followed by a dropout layer with the value set to 0.4.

*Dense layer:* This layer outputs the arrays of shape 64 and the activation function used is SoftMax which assigns probabilities for each label independently. This layer followed by a dropout layer with a value set to 0.4.

L2 regularization (Ridge Regression) is applied to the network with l2 penalty set to 0.001 with an intention to reduce the overfitting. The model is compiled using Adam optimizer which handles sparse gradients and categorical cross-entropy loss function.

## 4.7 Hyperparameter Optimization

The optimization approach used for this model is Bayesian Optimization with an intention to tune six parameters: epochs, batch size, dropout, activation function, learning rate and decay for Adam optimizer. The Gaussian process is used as a surrogate function which determines the best set of parameters based on their performance. These hyperparameters are applied to fitness objective function and update the surrogate model with new outcomes. This process is repeated until maximum calls are reached which is set to 20 for this model. Table 8 displays the values of tuned parameters for three different frame sizes:

Table 3: Optimized values of each parameter

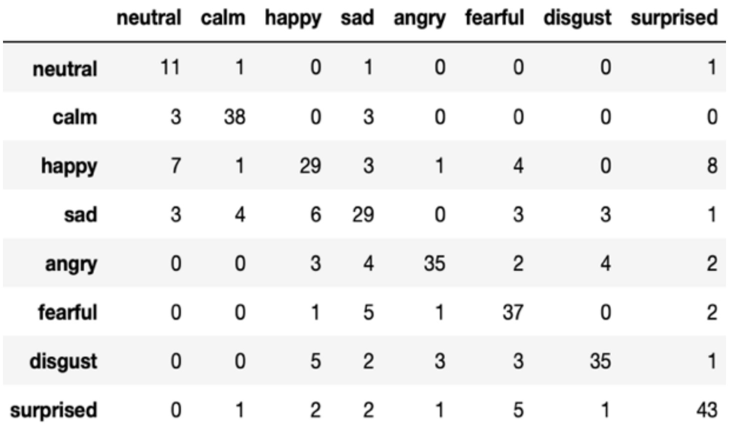
| Frame Size | Epochs | Batch Size | Dropout | Activation Function | Learning Rate | Decay |
|---|---|---|---|---|---|---|
| 25ms | 200 | 4 | 0.0 | reLu | 0.00042 | 0.0000001 |
| 50ms | 183 | 15 | 0.038 | reLu | 0.0003 | 5.682641 |
| 100ms | 189 | 34 | 0.06929 | reLu | 0.00067 | 6.248 |

# 5 Evaluation

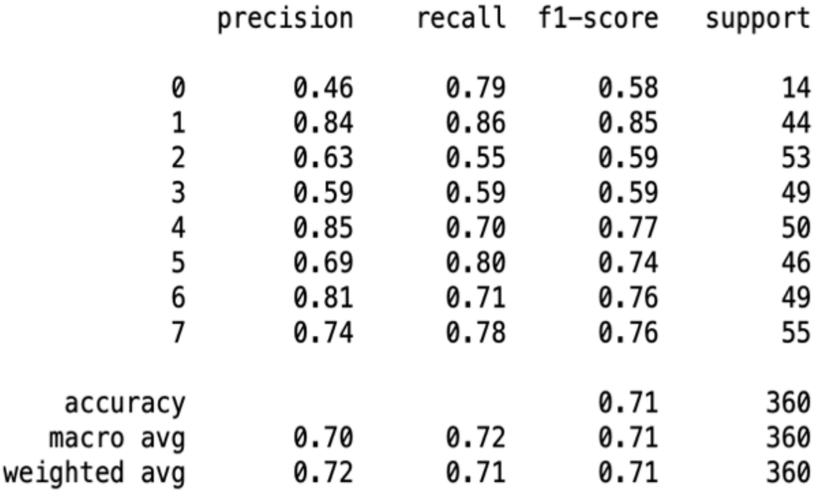
The proposed model is compared with previous research work (Gumelar et al. (2019)) where the author designed a deep neural network model and tested it on the same dataset as used in this research. Their model classified five emotions with gender dependence and achieved 78.8% accuracy whereas this proposed model classified seven classes of emotions and achieved 83.69% accuracy independent of gender. Another comparison made with the research work (Jannat et al. (2018)) where only two emotions are classified using the CNN model and obtained an accuracy of 66.41%. This model has outperformed CNN and DNN models proposed in previous studies. The model is used to conduct four experiments which are evaluated based on metrics discussed in subsection 3.3.

## 5.1 Experiment 1: Classifying eight classes of emotions with 25ms frame size

The model was tested with the class imbalanced dataset. Based on the outcome in Table 4, it is observed that calm emotion obtained the highest recognition rate and happy obtained the lowest. In addition, happy is most confused with surprised and neutral emotion. Also, surprise class got confused with all other classes at least once except neutral. The average values of precision, recall, and F1 score indicate that the model has performed well. The model classified four classes of emotions with accuracy more than 75% except the samples of happy, sad and angry emotions. The overall accuracy achieved by the model for this experiment is 71.2%.



| | neutral | calm | happy | sad | angry | fearful | disgust | surprised |
|---|---|---|---|---|---|---|---|---|
| **neutral** | 11 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| **calm** | 3 | 38 | 0 | 3 | 0 | 0 | 0 | 0 |
| **happy** | 7 | 1 | 29 | 3 | 1 | 4 | 0 | 8 |
| **sad** | 3 | 4 | 6 | 29 | 0 | 3 | 3 | 1 |
| **angry** | 0 | 0 | 3 | 4 | 35 | 2 | 4 | 2 |
| **fearful** | 0 | 0 | 1 | 5 | 1 | 37 | 0 | 2 |
| **disgust** | 0 | 0 | 5 | 2 | 3 | 3 | 35 | 1 |
| **surprised** | 0 | 1 | 2 | 2 | 1 | 5 | 1 | 43 |

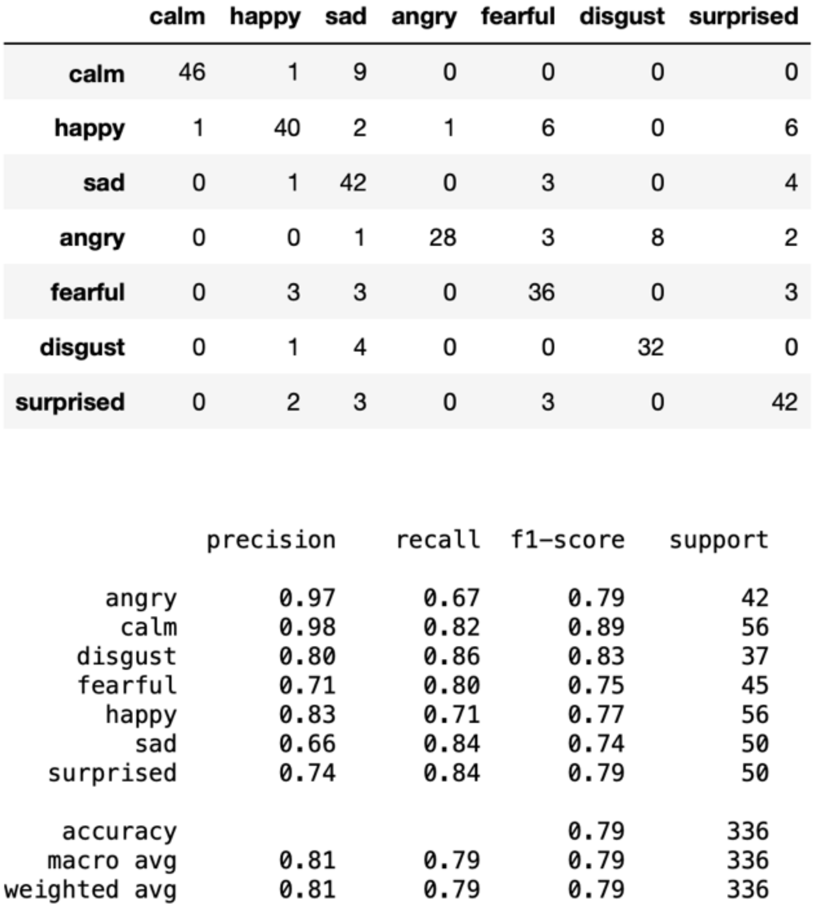


```
              precision    recall  f1-score   support

           0       0.46      0.79      0.58        14
           1       0.84      0.86      0.85        44
           2       0.63      0.55      0.59        53
           3       0.59      0.59      0.59        49
           4       0.85      0.70      0.77        50
           5       0.69      0.80      0.74        46
           6       0.81      0.71      0.76        49
           7       0.74      0.78      0.76        55

    accuracy                           0.71       360
   macro avg       0.70      0.72      0.71       360
weighted avg       0.72      0.71      0.71       360
```

Table 4: Confusion matrix and Classification report

## 5.2 Experiment 2: Classifying seven classes of emotions with 25ms frame size

The audio clips of 'neutral' class has been removed from the database for this experiment to resolve the issue to class imbalance. Based on the results described in Table 5, calm emotion has obtained the highest recognition rate and got most confused with sad emotion and angry emotion obtained the least rate which is mostly confused with disgust emotion. The average precision value of 0.81 reveals that the number of false positives predicted is quite low and the average recall value of 0.79 indicates that a high proportion of actual positives are correctly identified. The model classified all the classes of emotions with accuracy more than 75% except the samples of happy and angry emotion. The overall accuracy achieved by the model for this experiment is 79%.



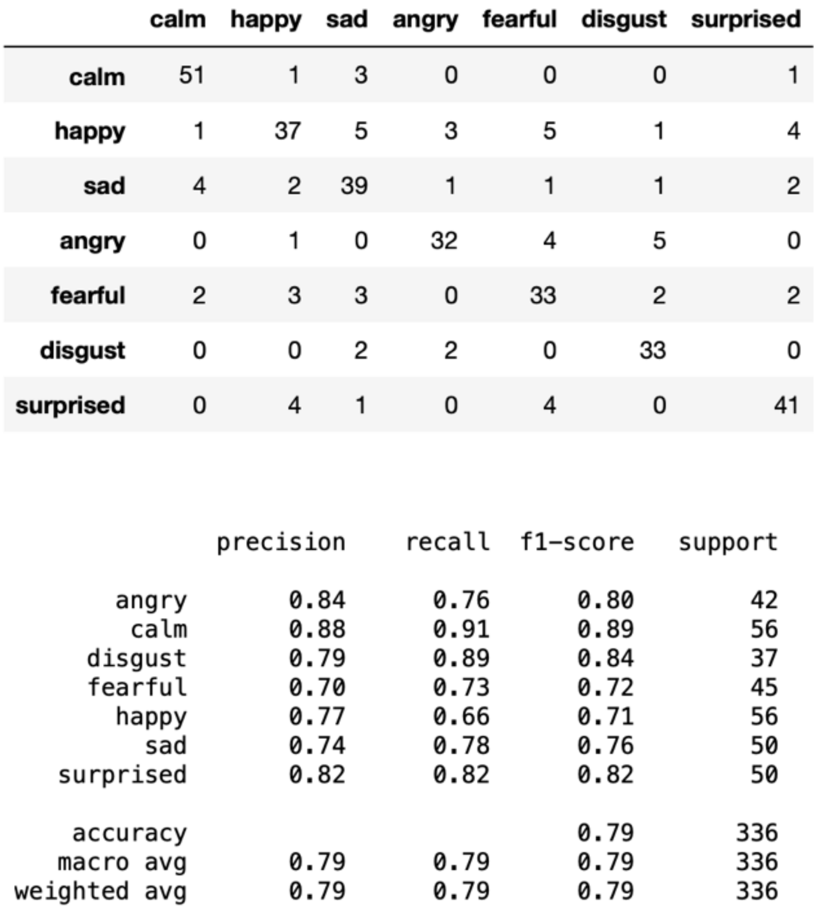
| | calm | happy | sad | angry | fearful | disgust | surprised |
|---|---|---|---|---|---|---|---|
| **calm** | 46 | 1 | 9 | 0 | 0 | 0 | 0 |
| **happy** | 1 | 40 | 2 | 1 | 6 | 0 | 6 |
| **sad** | 0 | 1 | 42 | 0 | 3 | 0 | 4 |
| **angry** | 0 | 0 | 1 | 28 | 3 | 8 | 2 |
| **fearful** | 0 | 3 | 3 | 0 | 36 | 0 | 3 |
| **disgust** | 0 | 1 | 4 | 0 | 0 | 32 | 0 |
| **surprised** | 0 | 2 | 3 | 0 | 3 | 0 | 42 |

```
              precision    recall  f1-score   support

       angry       0.97      0.67      0.79        42
        calm       0.98      0.82      0.89        56
     disgust       0.80      0.86      0.83        37
     fearful       0.71      0.80      0.75        45
       happy       0.83      0.71      0.77        56
         sad       0.66      0.84      0.74        50
   surprised       0.74      0.84      0.79        50

    accuracy                           0.79       336
   macro avg       0.81      0.79      0.79       336
weighted avg       0.81      0.79      0.79       336
```

Table 5: Confusion matrix and Classification report

## 5.3 Experiment 3: Classifying seven classes of emotions with 50ms frame size

Based on the results described in Table 6, calm emotion has again obtained the highest recognition rate and sad samples obtained the least recognition rate. Happy samples are most misclassified as fearful and sad classes and Angry samples as disgust. The model classified all the classes of emotions with accuracy more than 75% except the samples of fearful emotion. The average values of recall obtained for this experiment is same as but there is a slight difference in the value of precision when compared to experiment 2. The overall accuracy achieved by the model for this experiment is 79.16%.



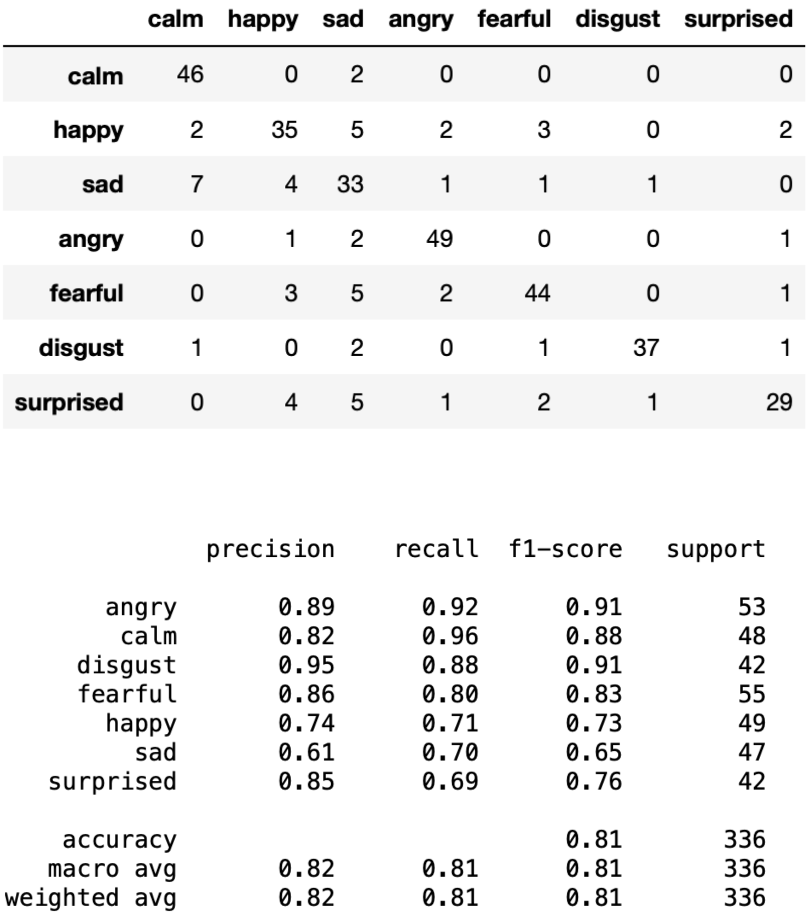
|  | calm | happy | sad | angry | fearful | disgust | surprised |
|---|---|---|---|---|---|---|---|
| **calm** | 51 | 1 | 3 | 0 | 0 | 0 | 1 |
| **happy** | 1 | 37 | 5 | 3 | 5 | 1 | 4 |
| **sad** | 4 | 2 | 39 | 1 | 1 | 1 | 2 |
| **angry** | 0 | 1 | 0 | 32 | 4 | 5 | 0 |
| **fearful** | 2 | 3 | 3 | 0 | 33 | 2 | 2 |
| **disgust** | 0 | 0 | 2 | 2 | 0 | 33 | 0 |
| **surprised** | 0 | 4 | 1 | 0 | 4 | 0 | 41 |

```
              precision    recall  f1-score   support

       angry       0.84      0.76      0.80        42
        calm       0.88      0.91      0.89        56
     disgust       0.79      0.89      0.84        37
     fearful       0.70      0.73      0.72        45
       happy       0.77      0.66      0.71        56
         sad       0.74      0.78      0.76        50
   surprised       0.82      0.82      0.82        50

    accuracy                           0.79       336
   macro avg       0.79      0.79      0.79       336
weighted avg       0.79      0.79      0.79       336
```

Table 6: Confusion matrix and Classification report

## 5.4 Experiment 4: Classifying seven classes of emotions with 100ms frame size

Based on the results described in Table 7, the calm class has consistently obtained the highest recognition rate with 96% of correct classification and sad samples are mostly wrongly classified. The model classified all the classes of emotions with accuracy more than 75% except the samples of sad and fearful emotions. This experiment has obtained the best average values of precision and recall (0.84) when compared to other three experiments. The overall accuracy achieved by the model for this experiment is 81.25%.



| | calm | happy | sad | angry | fearful | disgust | surprised |
|---|---|---|---|---|---|---|---|
| **calm** | 46 | 0 | 2 | 0 | 0 | 0 | 0 |
| **happy** | 2 | 35 | 5 | 2 | 3 | 0 | 2 |
| **sad** | 7 | 4 | 33 | 1 | 1 | 1 | 0 |
| **angry** | 0 | 1 | 2 | 49 | 0 | 0 | 1 |
| **fearful** | 0 | 3 | 5 | 2 | 44 | 0 | 1 |
| **disgust** | 1 | 0 | 2 | 0 | 1 | 37 | 1 |
| **surprised** | 0 | 4 | 5 | 1 | 2 | 1 | 29 |

```
              precision    recall  f1-score   support

       angry       0.89      0.92      0.91        53
        calm       0.82      0.96      0.88        48
     disgust       0.95      0.88      0.91        42
     fearful       0.86      0.80      0.83        55
       happy       0.74      0.71      0.73        49
         sad       0.61      0.70      0.65        47
   surprised       0.85      0.69      0.76        42

    accuracy                           0.81       336
   macro avg       0.82      0.81      0.81       336
weighted avg       0.82      0.81      0.81       336
```

Table 7: Confusion matrix and Classification report

## 5.5 Discussion

Experiment 5.4 where seven classes of emotions are classified with 100ms frame size has attained exceptional performance when compared to other three experiments. It is noted that in all the four experiments 'calm' class was consistently confused with 'sad' class and this is because both the classes of emotion have low energy and similar speech rate when compared to other classes of emotions. The other consistent confusion was between 'happy' and 'surprised' classes because the both emotions have wide pitch range. The proposed model has outperformed other previous research works conducted on the same dataset with an accuracy of 81.25%. The results attained from experiments have been summarised in Table 8.

Table 8: Experimental results

| Frame Size (in ms) | Accuracy (in %) |
|---|---|
| 25 | 79 |
| 50 | 79.16 |
| 100 | 81.25 |

# 6 Conclusion and Future Work

This research has proposed a convolutional recurrent neural network model. The five major speech representatives: MFCC, Mel spectrogram, Chroma, ZCR, and RMSE are extracted at frame-level and applied to the model for each class of emotion. The set of features was chosen with an intention to make the best feature set. The model was designed and then optimized using Bayesian Optimization for six different parameters: epochs, batch size, dropout, activation function, learning rate and decay for Adam optimizer. The model experimented on three different frame sizes: 25ms, 50ms, and 100ms obtained from each audio signal. It is noted that the best performance was attained on the largest frame in size. Experiments on the RAVDESS dataset demonstrates the dominance of our proposed system compared with state-of-the-art based on the overall accuracy of 81.25%. The experiments performed have motivated to research more on the subject to receive the answers to all the questions appeared during this study. Some of those have outlined below:

- Another element of interest is to analyze the influence of gender in identifying the emotion in a speech signal since the fundamental frequency and pitch period vary. The fundamental frequency for men lies between the range of 85 Hz – 155 Hz whereas it lies between 165 Hz – 255 Hz for women. Likewise, the pitch period value for men is estimated to be 8ms and 4ms for women.

- This research focussed on the effect of frame size and tested the model on three different sizes. It is observed that the model performed better with the larger frame size. This influences to experiment with the other frame sizes in order to find the optimal one.

# 7 Acknowledgement

I would like to show my gratitude to my supervisor, Prof. Christian Horn for his exceptional support and guidance which has extensively contributed to shaping this research. His precision to detail, enthusiasm, and knowledge have been a motivation and kept my tasks on track. I would also like to extend my gratitude to the National College of Ireland which has supported this study and provided me appropriate resources to conduct this research.

# References

Ayadi, M., Kamel, M. S. and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition* **44**: 572–587.

Ayek, H., Lech, M. and Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition, *Neural Networks* **92**: 60–68.

Badshah, A. M., Ahmad, J., Rahim, N. and Baik, S. W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network, pp. 1–5.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower Provost, E., Kim, S., Chang, J., Lee, S. and Narayanan, S. (2008). Iemocap: Interactive emotional dyadic motion capture database, *Language Resources and Evaluation* **42**: 335–359.

Chen, M., He, X., Yang, J. and Zhang, H. (2018). 3-d convolutional recurrent neural networks with attention model for speech emotion recognition, *IEEE Signal Processing Letters* **25**(10): 1440–1444.

Cowen, A., Elfenbein, H., Laukka, P. and Keltner, D. (2018). Mapping 24 emotions conveyed by brief human vocalization, *American Psychologist* .

Demircan, S. and Kahramanli, H. (2014). Feature extraction from speech data for emotion recognition, *Journal of Advances in Computer Networks* **2**: 28–30.

Eduard Frant, Ioan Ispas, V. G. M. d. E. Z. and Stoica, I. C. (2017). Voice based emotion recognition with convolutional neural networks for companion robots, *Romanian Journal Of Information Science and Technology* **20**(3): 222–240.

Ekman, P. (1992). An argument for basic emotions, *Cognition and Emotion* **6**: 169–200.

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B. and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional., *Psychological science* **18 12**: 1050–7.

Gumelar, A. B., Kurniawan, A., Sooai, A. G., Purnomo, M. H., Yuniarno, E. M., Sugiarto, I., Widodo, A., Kristanto, A. A. and Fahrudin, T. M. (2019). Human voice emotion identification using prosodic and spectral feature extraction based on deep neural networks, *IET Signal Processing* pp. 1–8.

Han, K., Yu, D. and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* .

Jannat, S. R., Tynes, I., Lime, L., Adorno, J. and Canavan, S. (2018). Ubiquitous emotion recognition using audio and video data, pp. 956–959.

Livingstone, S. and Russo, F. (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**: e0196391.

Mirsamadi, S., Barsoum, E. and Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention, pp. 2227–2231.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M. and Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review, *Softw., Pract. Exper.* **7**: 19143–19165.

Ronan, D., Reiss, J. D. and Gunes, H. (2018). An empirical approach to the relationship between emotion and music production quality.

Russell, J. (1980). A circumplex model of aect, *Journal of Personality and Social Psychology* **39**: 1161–1178.

Sarprasatham, M. (2015). Emotion recognition: A survey, *International Journal of Advanced Research in Computer Science* **3**: 14–19.

Schuller, B. (2018). Speech emotion recognition: Two decades in a nutshell benchmarks and ongoing trends, *Communications of the ACM* **61**: 90–99.

Sugan, N., Srinivas, N., Kar, N., Kumar, L., Nath, M. and Kanhe, A. (2018). Performance comparison of different cepstral features for speech emotion recognition, pp. 266–271.

Swain, M., Routray, A. and Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review, *International Journal of Speech Technology* **21**.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* .

Zhao, J., Mao, X. and Chen, L. (2018). Learning deep features to recognise speech emotion using merged deep cnn, *IET Signal Processing* **12**(6): 713–721.

Zhao, Y., Jin, X. and Hu, X. (2017). Recurrent convolutional neural network for speech processing, *American Scientist* **89**: 344–350.