

## Speech Emotion Recognition Based on CNN

RAKSHA

Anand International College of Engineering, Jaipur  
Electronics and Communication Engineering

**Abstract**—The ability to naturally connect with computers has contributed to the rise in popularity of automatic speech emotion recognition. Voice analysis is a method for identifying feelings. However, there is quiet in speech that has nothing to do with expression. One technique to perform better is to get rid of the silence, while another is to ignore the stillness and pay more attention to the speech. In this research, we propose incorporating silence reduction with a caring model to improve speech emotion performance. The findings prove that combining the noise-cancelling and attention-focusing models is superior to using either one alone. In the subject of human-computer interaction, speech emotion recognition is a crucial and difficult topic. Several models and feature sets have been proposed for use in training the system. In this study, we tested extensively using convolutional neural networks that can learn from several perspectives. We evaluate the effectiveness of the system with input signals of varying durations, acoustic feature kinds, and emotive speech (improvement/written) styles. Our experimental results on the Ryerson Emotional Voice and Song Audiovisual Database (RAVDESS) demonstrate that the recognition performance is not related to the input function, but rather is dependent on the type of voice data. We have gathered the most recent findings from RAVDESS's unplanned voice data.

**Keywords**—*CNN, Emotion Recognition, Speech Emotion*

### 1. INTRODUCTION

In the realm of human communication, emotions play a pivotal role, shaping the quality and depth of our interactions. The ability to recognize and understand emotions conveyed through speech is not only a fundamental aspect of human interaction but also a significant challenge in the field of artificial intelligence and machine learning. Speech Emotion Recognition (SER), the task of automatically detecting emotions from spoken language, has drawn considerable attention due to its wide-ranging applications in fields such as human-computer interaction, sentiment analysis, healthcare, and virtual assistants.

While traditional approaches to SER have relied on handcrafted features and machine learning techniques, recent advancements in deep learning have revolutionized the way we approach this problem. Convolutional Neural Networks (CNNs), originally designed for image processing, have emerged as a potent tool for extracting intricate patterns from sequential data, making them an ideal candidate for speech emotion recognition. The unique characteristics of CNNs, such as their ability to learn hierarchical features and capture spatial dependencies, provide a promising foundation for enhancing the accuracy and robustness of emotion recognition models.

This research paper delves into the fascinating realm of speech emotion recognition, specifically focusing on the application of Convolutional Neural Networks. We aim to

explore the potential of CNN-based models to tackle this intricate challenge by investigating their capacity to automatically learn and extract meaningful features from the audio signals. By combining the power of deep learning with the richness of emotional cues present in speech, our work endeavors to unlock new horizons in SER, enabling more sophisticated and context-aware human-computer interactions.

In this paper, we provide a methodologies for speech emotion recognition, We present a detailed examination of the CNN architecture, its adaptability to the sequential nature of audio data, and its capacity to discern the nuanced variations in vocal expression. Through experiments and analyses, we illustrate the potential of CNNs to yield superior results in emotion recognition tasks when compared to conventional methods. Additionally, we address some of the key concerns, including the availability of annotated datasets and the interpretability of deep learning models, that are essential for the practical deployment of CNN-based SER systems. This research paper strives to shed light on the transformative role that CNNs can play in the fascinating journey of Speech Emotion Recognition, ultimately leading us towards a future where machines can better understand and respond to human emotions.

## 2. LITERATURE SURVEY

**Michael Neumann et. All (2017)** The authors of this study used a careful CNN to compare several variables for identifying the emotional tone of a speaker's voice. Prosodic characteristics yield significantly worse outcomes than logMel, MFCC, and eGeMAPS. One possible explanation for this is that there are less characteristics in the later. Similar outcomes point to the importance of model architecture and training data quantity and quality for a CNN rather than a specific set of characteristics. Improvised speech was shown to have significant variations from scripted speech, with the former yielding superior results for the authors. The performance drops off significantly as the signal duration decreases, but it's still quite good for transmissions as short as two seconds. The provided ACNN will be tested on an additional database in future developments. [1]

**Qinying Yuan et. All (2022)** The study of teaching and learning has traditionally relied heavily on classroom observation. In the past, classroom observations relied on humans and their subjective opinions to determine the relative merits of various teaching methods. More persuasive conclusions may be drawn from the more objective data acquired with a convolutional neural network model. The convolutional neural network model is labor- and time-efficient since it allows us to collect data without having to sit through a whole lecture. In this research, we present an algorithm for emotion identification in speech with varying lengths that is based on an attention mechanism that prioritizes the most important details by using an embedded attention module. In order to prioritize the information included inside the address spectrogram, this work introduces a spatiotemporal attention module. Part of the CNN's channel characteristics are focused on by the convolutional channel attention module, which then returns the relevant information. [2]

**Chenchen Huang et. All (2014)** In this study, the authors suggest a method for automatically extracting the emotional characteristic parameter from an emotional voice signal using deep belief networks (DBNs), one of the deep neural networks. We developed a classifier model that is based on DBNs and SVMs, which is a combination of the two previously mentioned methods. This approach can properly extract emotion characteristic characteristics, boosting the recognition rate of emotional speech recognition, and it has a minimal complexity and a final recognition rate that is 7% higher than typical artificial extract. On the downside, DBN's feature extraction model training took 136 hours, which is much more time than other feature extraction approaches. [3]

**Li Zheng et. All (2018)** The research here use a CNN model to extract features and combines them with an RF classifier. Based on this, the NAO robot's algorithm for identifying emotions in Chinese speech was developed. The application procedure refines the preexisting Record Sound box and creates a new Recorder box. The updated Recorder device can collect voice signals that are suitable for investigating NAO-BASED speech signal analysis, Chinese words segmentation, and speech recognition, in addition to meeting the format criteria of speech emotion identification. The suggested CNN-RF model provides NAO with essential capabilities in voice emotion identification after extensive testing. [4]

**Dr. N. Herald Anantha Rufus et. all (2022)** Current Speech Emotion Detection Methods and Datasets Need to Be Analyzed and Contrasted Due to the Rapid Development of Neural Systems and the Growing Demand for Accurate and Close Actual Speech Emotion Recognition in Human-Computer Interfaces. Several studies, polls, and even Deep Learning strategies have been conducted on the topic of emotion recognition. In the long run, it will be essential to have access to a system like this that is both trustworthy and flexible. The goal of this research was to one day be able to do something called "real-time emotion detection." There is a lot of network infrastructure built on CNNs that can be used to process speech. [5]

**Abdul Ajj Ansari et. All (2020)** Deep learning algorithms used for emotion recognition are the subject of several studies and surveys. A far more dependable system like this, with infinite applications across disciplines, is essential for the future. Several datasets were looked into in an effort to solve the emotion recognition challenge using inception net in this study. I used TensorFlow to train my model. About 39% accuracy is attained. The same underlying framework can be used to provide real-time emotion recognition in the near future. [6]

**Anushka Sandesara et. All (2020)** There has been a long history of "Chatting" between humans and robots in today's society. Discussions on Speech Emotion Recognition technology have been ongoing for quite some time, with the earliest known academic publication on the issue having been authored by Daellert et al. Despite the usefulness of these applications, voice emotion identification is difficult since emotions are personal. Every person is unique in the way they process and express their emotions. No universal standards or procedures exist for classifying feelings. Humans, not even computers, are perfect at interpreting the feelings of others. This document lists and briefly describes many speech emotion recognition algorithms. This study aims to give a discussion of, and a comparison of, the many methods currently available for implementing speech emotion

identification for audio file types. SVM, MLP, RNN-LSTM, and CNN were all discussed, and the authors claimed that these four classifiers were the most accurate. [7]

**Kaibei Peng et. All (2021)** This paper proposes a CNN emotion classification model, which first extracts MFCC features of speech, then sends the extracted features to convolutional neural network, and finally outputs the category of each speech, to be used in emergency treatment of railway stations in order to reduce harm caused by danger in crowded places like railway stations. The simulation results demonstrate that this model outperforms RNN and MLP models in terms of accuracy. This paper's approach to hazard warning at train stations offers a reliable point of reference. [8]

**Apoorv Singh et. All (2020)** Several models were built until the authors found the best CNN model for the emotion differentiation challenge. The authors improved upon the accuracy of the baseline model by 71 percentage points. If the authors had additional data, their model would have been more accurate. The authors' model also excelled at identifying the difference between male and female speakers. The work of the authors can be expanded to incorporate a robot's ability to recognize a person's emotional state and use that knowledge to have a more natural conversation with that person; it can also be used to improve the quality of music recommendations made by apps like Spotify and Amazon's Recommendations service to customers. Additionally, a sequence-to-sequence paradigm for generating emotionally nuanced voices is within our reach in the near future. For instance, a sorrowful voice, an eager voice, etc. [9]

**B. Sandeep et. All (2021)** One of the most important aspects of emotional computing in HCI is verbal communication. Accurately understanding the meanings of the words or the linguistic category and identifying the emotion involved in the speech are crucial for improving performance in this sort of interaction. Emotional states including boredom, anxiety, joy, and melancholy are modeled by analyzing speech waves, which carry signals for these emotions. The goal of this research is to create a system for predicting people's emotional responses based on what they say, using Convolutional Neural Network (CNN) classifiers to identify the various expressions of emotion. Mel-Frequency Cepstral (MFCC) is the extracted spectral features. The suggested method uses the LIBROSA module in the Python programming language to discern between emotions including happiness, surprise, rage, apathy, sorrow, fear, and more from the Ryerson AudioVisual Database of Emotional Speech and Song (RAVDESS). The process of feature selection (FS) was utilized to identify the best collection of features to use. The results suggest that CNN yields the highest performance gains. Model accuracy is determined by comparing actual values to those predicted by the model, which are obtained after training the model with the training dataset and testing it with the test data set. [10]

### 3. METHODOLOGY

#### 3.1 Audio feature extraction

Our SMILE online audio analysis tools is a freely available resource for calculating features and functions. It uses the same benchmark sound function set as the 2011 audio/visual emotion challenge. It includes 25 energy and spectrum low-level descriptors (42 functions), 6 voice-related low-level descriptors (32 functions), 25 incremental energy/spectrum LLD coefficients (23 functions), 6 voice-related LLD incremental coefficients (19 functions), and 10 voiced/unvoiced duration characteristics. In total, there are 1941 functions. Table 2 and Table 3 provide in-depth data on LLD and functions, respectively. Common functions for analyzing auditory signals and emotions are represented in the LLD set. The set is based on a comparable set, such as the one used at the Interspeech 2011 Speaker State Challenge, but it has been manually trimmed to make sure the LLD-function combination does not provide a consistent result with little information and/or a considerable degree of noise. There are some cases when the LLD/function combination doesn't tell you anything, as when the "minimum pitch" is 0.

#### 3.2 Emotion recognition model

To predict both categorical (such as "happiness") and continuous (such as "arousal level") labels, the emotion identification model is a basic feedforward neural network trained with multi-task learning. Each job has its own independent hidden layer that is present before the output layer. The eGeMAPS, or extended Geneva Minimalist Acoustic Parameter Set, is a feature vector developed for emotion prediction and serves as the input. The eGeMAPS function was motivated by the fact that it successfully modeled perception-related modifications to speech. The low-level descriptor (LLD) has 88 total functions, each of which is a discourse-level function. Energy, spectrum, cepstrum, prosody, and speech descriptors are all part of LLD.

#### 3.3 Data Pre-processing and Exploration

Before we go into pre-processing and data scanning, we'll cover the fundamentals of feature selection. Mel's scale deals with the way humans interpret frequencies. The pitch, or volume, on this scale is universally agreed upon by

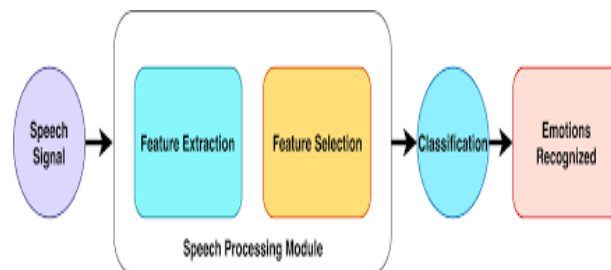
listeners. The rate is proportional to it. The faster the speed at which sound vibrations travel, the higher the frequency. The octave-wide logarithmic spectrum of wave amplitude is utilized to calculate the period in seconds.

### 3.4 Data

The North American Emotion Database is used for all investigations. It includes voice, video, and face motion capture data from the two sequences recorded: script replay and spontaneous speech. The pivotal annotation includes both discrete dimensions and labelled categories (such as joy, sadness, and anger). Activation, valence, and utility all have discrete values from 1 to 5. In this analysis, we classified emotions into the same four broad categories of anger, happiness, truth, and apathy. Joy is the umbrella term for a wide range of positive emotions. As a result, we fixed the maximum length (mean duration + standard deviation) that every given sample may have in order to satisfy CNN's requirements.

### 3.5 LSTM Classifier

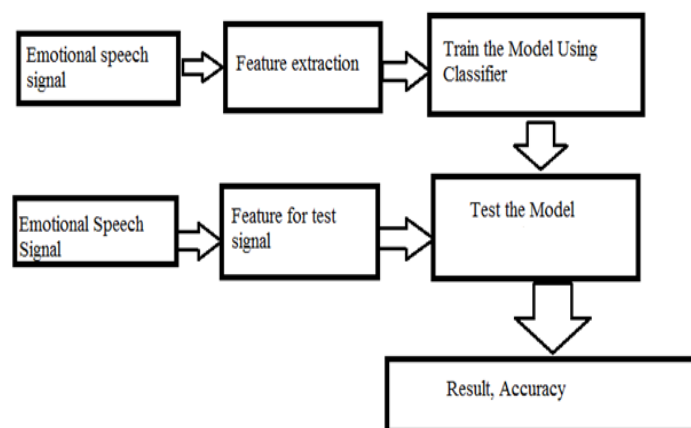
Composed of the standard building blocks of any model recognition process (emotional speech recognition included). The suggested method improves upon previous approaches by employing a silence-cancellation function before extraction and a two-way, long-term storage network (LSTM) classification model. The suggested system is illustrated here:



**Figure 3.1: “Proposed speech emotion recognition system with silence removal and bidirectional LSTM classifier”**

### 3.6 Speech Emotion Recognition System

This schematic depicts the brain's emotional circuitry for understanding spoken language. The voice emotion recognition system consists of three primary modules: the emotion signal, the function extraction, and the classification and display output. The figure below depicts a very basic architecture for identifying emotions from a person's speech. First, the voice signal from the speaker must be analyzed to determine relevant properties. Functions like this will be the primary descriptors. Words like "happy," "neutral," "sad," "calm," and "anguish" are all examples of emotions. Irritation and dread. Both CNN and DT are used to ensure the user is producing the desired results. There is a consistent user signal. The quality of the database that feeds into the system for recognizing speech emotion is taken into account in the evaluation.



**Figure 3.2: “Speech Emotion Recognition System”**

### 3.7 Audio Features

In audio processing, determining good feature representations and classifier designs is typically treated as a separate issue. The function that is created may not be optimal for the classification goal, which is a drawback of this method. The extraction of features while optimizing for a purpose (such as classification) is what we mean when we talk about a deep neural network (DNN). When it comes to recognizing languages, for instance,

Mohamed et al. Research indicates that the activation of a DNN's lower layers may be viewed as a speaker adaption characteristic, while the activation of its upper layers can be viewed as a class-based distinction. Mel Frequency Cepstral Coefficients (MFCCs) have been the go-to acoustic function for audio analysis for decades. These are Discrete Cosine Transform (DCT) whitened and compressed approximations of amplitude spectra that have been projected onto a series of reduced frequency bands. Since the latter eliminates data and breaks down spatial relationships, it was shown to be superfluous in deep learning models. If you do that, you'll end up with a log mel spectrum, which is rather common in the music industry. Research into the physiology of the ear and how we interpret speech served the inspiration for the Mel filter bank. The capture transposition as a translation representation is preferable for several jobs. A common component, the compensation of the logarithmic frequency dimension, is used to scale the fundamental frequencies and overtones, therefore transferring the pitch. The constant Q spectral quality of a suitable filter bank is used to accomplish this frequency scaling. Log mel (or constant Q) spectrograms are time series representations of frequency spectra. The natural sound boxes next to one another have a temporal and spectral relationship, just like in nature photographs. However, due to the physical qualities of sound creation (harmonic), there exist various correlations of frequencies with the same basic frequency. The amplitude of the harmonic sequence can be generated directly by local models (such as CNN) with the addition of a third dimension. Furthermore, the values throughout the frequency bands are distributed very differently from the illustration. The spectrogram can be normalized on a band-by-band basis to address this problem. The window size in the spectrum balances the temporal resolution (small window) with the frequency resolution (large window). While both the log mel and constant Q spectra may be obtained with small window sizes, the spectrograms get flattened unevenly at higher frequencies, rendering them unsuitable for use in local space models. Computer spectra are another alternative; they can be projected into the same frequency band and then processed independently as channels by adjusting the window size. The author also looked at what happens when you mix and match different spectral features. Several alternatives to the predetermined filter bank were developed to further streamline the feature extraction process and postpone it during data-driven statistical model learning. The Mel spatial triangle filter was abandoned in favor of a data-driven filter. Put the data-driven filter to good use in tandem with the other network, and make use of the high-resolution amplitude spectrum and the original audio waveform representation as input. This permits direct optimization of the learnt filter in terms of the goal. Logarithm-mel spectroscopy computations are simulated at a lower model layer, while all filter parameters are learnt at a higher level. In this scenario, we may train the causal regression model of time-domain waveforms from scratch, ditching the filter bank notion in the process.

In conventional sound signal processing, the MFCC comes second to a unique waveform or spectrogram and then to deep learning. The most often used form for conventional audio signal processing is the MFCC. The authentic waveform precludes any manual design features. These operations should optimize the learning representation for the job and make greater use of the deep learning modeling capabilities. It's not always easy to reap the practical benefits of this, as it leads to increased calculation costs and data requirements. Methods that employ the logmel spectrogram typically require less data and less training to get results equivalent to today's technical standard since it gives a more compact representation for analytics applications like ASR, MIR, and ambient sound detection. Efficiency while employing the first-generation audio setup. Reconstructing the phase using (log-mel) amplitude spectrograms is a challenging operation when the goal is to synthesize sounds of high audio quality, such as in source separation, audio enhancement, TTS, or sound distortion. The original waveform or complex spectrogram is typically favored above the input representation.

### **3.8 The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)**

RAVDESS is a multi-modal database that has been shown effective through emotion and music. Twenty-four professional players with neutral North American accents and matching vocabularies were used to balance out the database. Communication encompasses a wide range of emotions, from serenity to anger to fear to surprise to contempt. Songs may evoke a wide range of feelings, from serenity to pleasure to sadness to rage and fear. There are three variations of each sentence available: an upbeat, a downbeat, and a neutral option. In every circumstance, face-to-face and voice communication is the best option. All 7,356 recordings meet the 10-point standard for sincerity, depth of feeling, and overall quality. These characteristics were found in 247 people who participated in the research as untrained North Americans. The results from the second round of testing came from 72 more participants. Reports indicate that the test rehearsal has a high degree of emotional validity and reliability.

### 3.9 Distinguishing features of the RAVDESS

RAVDESS has five unique characteristics that improve upon standard devices. - Variability. To begin, RAVDESS features 7356 clips whereas many episodes have fewer than 200. The RAVDESS factorial layout is seen in Sets 1 and 2. To the best of our knowledge, only three previous records have more than a thousand instances of multi-modal dynamic interaction. The RAVDESS is comprised of 24 professional athletes. Each performer has access to 104 distinct vocalizations. Feelings might range from happiness to anger to shock to contempt to serenity to indifference. Each actor's (AO) performance may be found in three different recordings: audio-only (AO), audio-only (AV), and audio-only (video). Since imaging studies have revealed that key brain regions are continually engaged for the same stimuli, this variety may be useful in the design of repeating tests. Researchers in the field of machine learning also have access to a vast library of recordings. Supervised learning and other machine learning techniques benefit greatly from validated databases since they include a wealth of data that can be used to train and evaluate various algorithms, such as sentiment classifiers. Second, the strength of one's feelings; there are only two degrees of expression (weak and strong) for each given emotion. We all know that the other two factions hold sway over the level of intensity. One of the most crucial features of emotions is their intensity, which has been the focus of several theoretical frameworks. The phrases "strength" and "activation" are used synonymously in these texts. In these frameworks, intensity is typically represented as one of multiple orthogonal axes in a multidimensional emotional space. It is more easily recognized perceptually than less anxious facial and vocal emotions. An observer is better able to mimic an intense facial expression than one with less nerve activity. Therefore, powerful displays might be helpful for researchers who need concrete examples of how people actually feel. However, typical intensity expression may be required for researchers wanting a depiction similar to those in ordinary life while researching small variances in emotional perception.

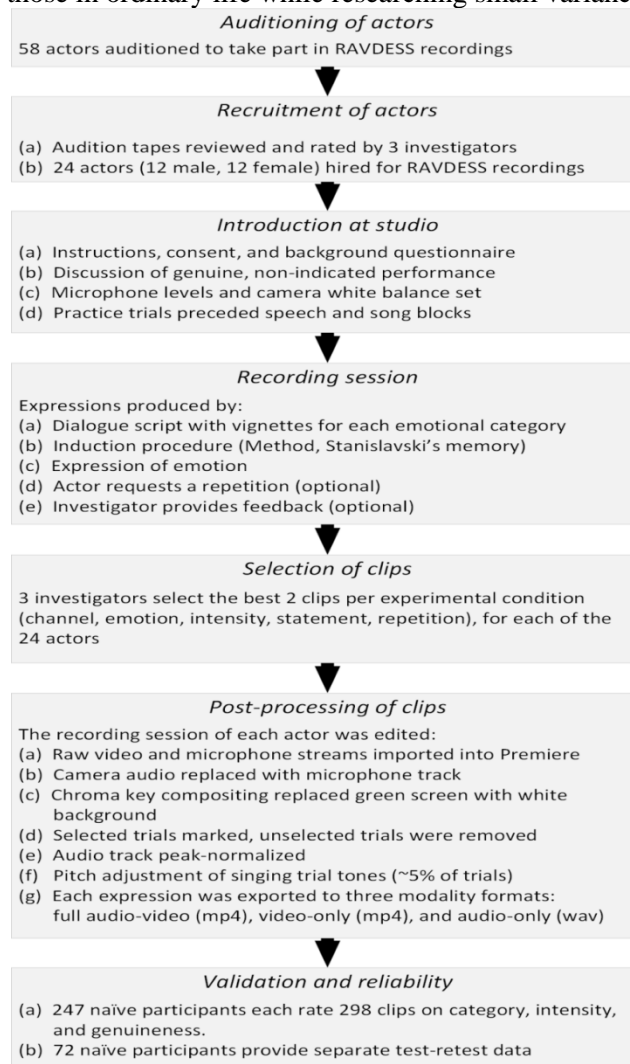
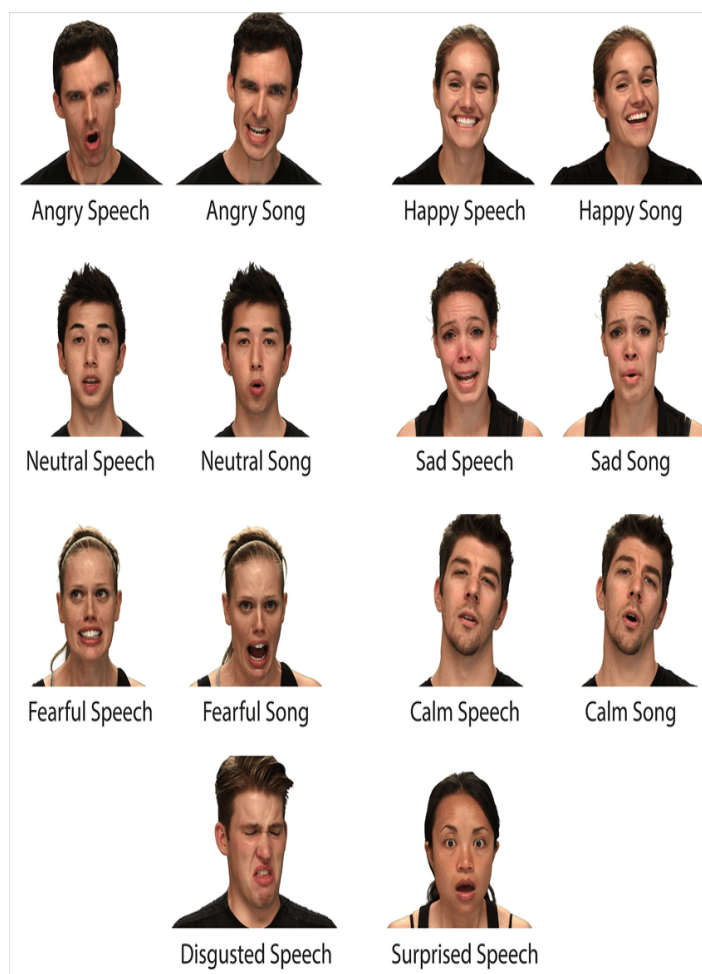


Figure 3.3: “Flow Diagram of RAVDESS Creation and Validation”



**Figure 3.4: “Examples of the Eight RAVDESS Emotions”**

Each actor's microphone track peak should be normalized to -3 dBFS in Adobe Audition CS6. If you want to keep the differences in volume that occur across different states of mind, peak normalization is the way to go. Melodyne was used to change the pitch of the vocal recording so that each of the three melodies would sound unique. It is called "coordinated" when the interval is between 35 and 100 cents, and "uncoordinated" when the interval is between 50 and 100 cents. If a note's frequency is off by more than 35 cents, it will be corrected such that it is within 35 cents of the intended frequency.

### 3.10 Data Preprocessing Steps

1. “Train, test split the data”
2. “Normalize Data — To improve model stability and performance”
3. “Transform into arrays for Keras”
4. “One-hot encoding of target variable”
5. “Reshape data to include 3D tensor”

### 3.11 Read Data Algorithm flow chart

This diagram illustrates the logic behind our method. After an algorithm reads a dataset, a data frame is prepared, the dataset is cleaned, audio file features are extracted, noise is introduced, the pitch of the data is adjusted, and finally the dataset is merged with the augmented data.

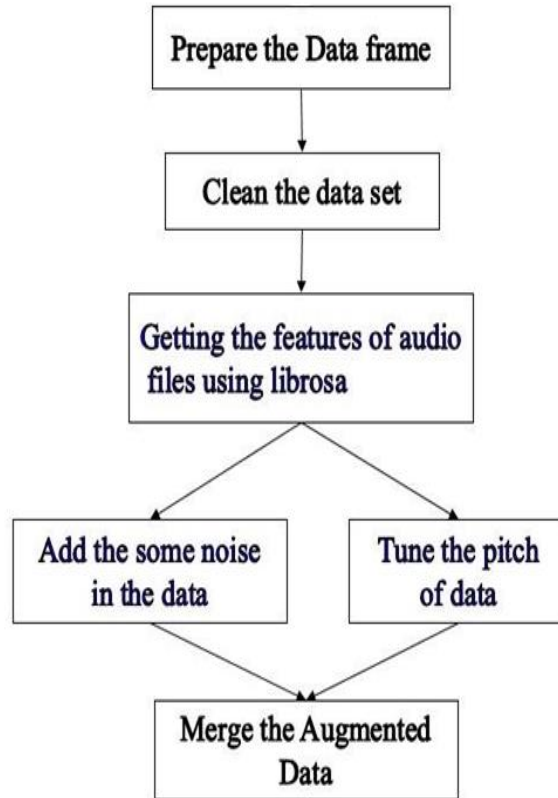
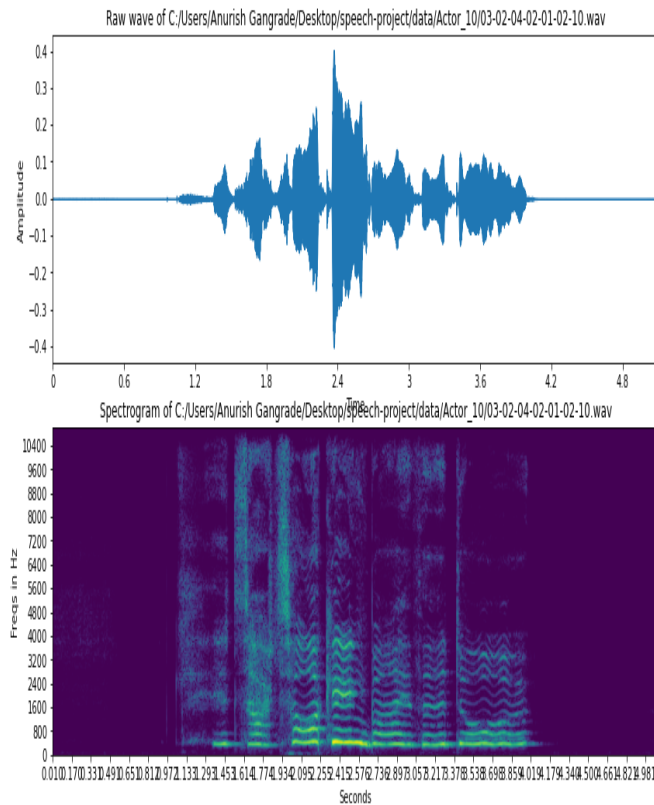


Figure 3.5: “Algorithm Flow Chart for Read Data”

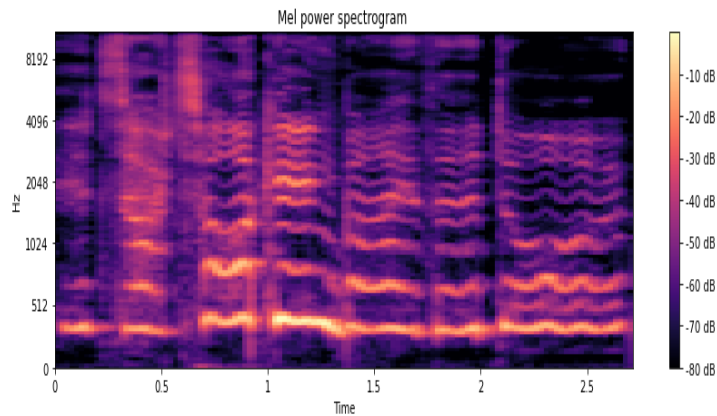
#### 4. RESULT



**Figure 4.1: Audio Wave File and Spectrogram**

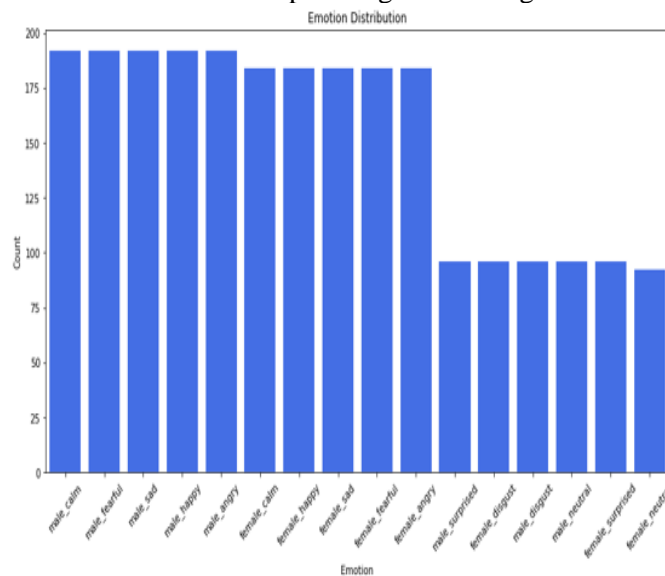
Figure 4.1 displays the waveform and spectrogram of a sample wave file taken from an audio collection. When the amplitude of the waveform is shown on the Y axis, the audio samples are presented as a time series. The amplitude is typically calculated from the pressure shift that occurred when the sound was captured by the microphone or receiver. If audio samples are metadata, then these time series signals are just training data for a machine learning algorithm.

**Mel Power Spectrogram**



**Figure 4.2: Mel Power spectrogram**

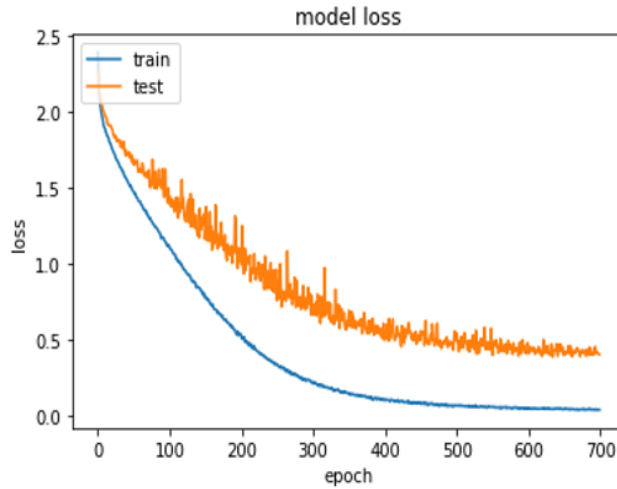
The Mel power spectrogram from the Urbansound8k sample melofrequency dataset is displayed in Figure 4.2. The time-varying visual representation of a signal's frequency spectrum is called a spectrogram. Digital spectrograms can be thought of as a stacked view of periodograms throughout the course of some time interval.



**Figure 4.3: Emotion Distribution**

Figure 4.3 is a scatter plot of feelings vs counts, and it displays the emotional distribution. Clearly, the range of possible count values is from zero to two hundred in this diagram. Male\_calm, male\_fearful, male\_sad, male\_happy, male\_angry, female\_calm, female\_fearful, female\_sad, female\_happy, female\_angry, male\_surprised, female\_disgust, male\_disgust, male\_neutral, and female\_surprised are the emotions employed in this narrative. The layout gets lower in elevation as one moves from left to right.

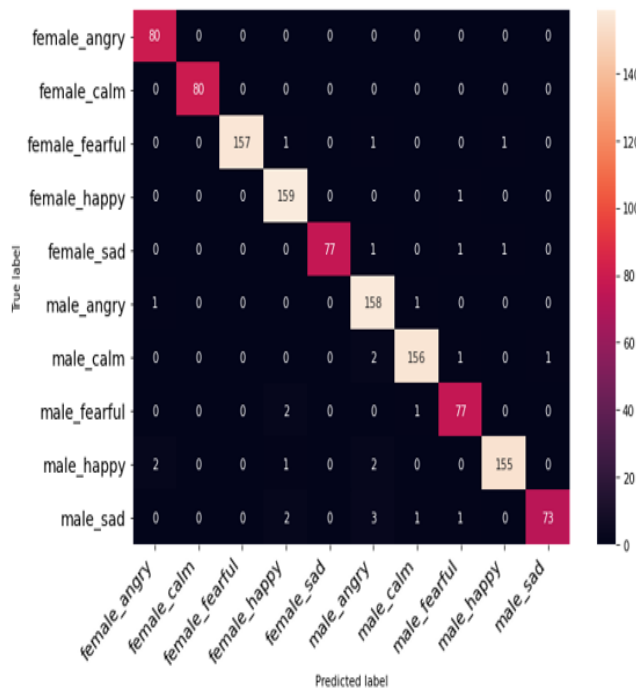
### Plotting the Train Valid Loss Graph



**Figure 4.4: Model Loss**

Model loss is plotted against time in Figure 4.4 (epoch on the x-axis, loss on the y-axis). In this case, the blue line denotes the training set, whereas the orange line represents the testing set. Here, we'll do some routine checks to verify if the model has successfully learnt a Lagrangian approximation for the system. After collecting data from 700 epochs, we find that the loss ranges from 0.0 to 2.5.

### Actual v/s Predicted



**Figure 4.5: Actual v/s predicted emotions**

Observed and expected emotions are plotted against one another in Figure 4.5. For both the female\_angry and female\_calm classes, 80 are accurate. One misclassification with female\_happy, one with male\_angry, and one with male\_happy occur in the female\_fearful class, for a total of 157 valid classifications. One was incorrectly assigned to the male\_fearful category, while 159 were assigned to the female\_happy category. Seventy-one instances of the female\_sad class were accurately predicted, whereas one instance each was mismatched with the male\_angry, male-fearful, and male-happy classes. One prediction was off for the female\_angry category, and another was off for the male\_calm category, for a total of 158 valid male\_angry predictions. There were 156 valid predictions for the male\_calm class, 2 mismatches with male\_angry, 1 mismatch with male\_fearful, and 1

mismatch with male\_sad. Seventy-seven are correctly labeled as male\_fearful, while two are incorrectly matched as female\_happy and one as male\_calm. 155 males were appropriately assigned to the joyful class, whereas 2 females, 2 males, and 1 female were misclassified as furious. There are 73 valid predictions for the male\_sad class and 3 mismatches to male\_angry, 2 to female\_happy, 1 to male\_calm, and 1 to male\_fearful.

## 5. CONCLUSION AND FUTURESCOPE

### 5.1 Conclusion

We can learn a great deal about human behavior and interactions from advances in automatic emotion recognition. Typically, high-dimensional features from a meticulously selected dataset are used in the creation of an emotion identification system. However, there is a downside to this approach: it is difficult to examine the dataset within the high-dimensional feature space due to its restrictions. Our suggested system uses a 9-layer convolutional neural network (CNN) architecture, which allows us to achieve a high accuracy of 87. Most notably, we tested our approach on the North American English dataset, where the gold standard currently sits at 76% accuracy. As a result, our proposed method shows a significant improvement of around 11%.

### 5.2 Future scope

Multimodal emotion recognition is one of the rapidly expanding research fields; it combines speech-based emotion detection with that based on other modalities, such as facial expressions and physiological signs. A deeper comprehension of human emotions and the signs they provide might result from such an amalgamation. Real-time recognition applications, such as customer service, virtual assistants, and mental health monitoring, would benefit greatly from the ongoing efforts to improve the speed and efficiency of CNN-based SER systems. Researchers are also concentrating on making these systems more resilient to noise, accents, and varying emotional expressions. This will allow them to function well in a wide variety of practical settings. Another potential direction is cross-lingual emotion recognition, which would allow SER models to identify emotions spoken in a variety of languages and dialects. More widespread international applications and cooperation could benefit from this. CNN-based SER has potential uses in healthcare settings, including the early diagnosis of emotional abnormalities, mood disorders, and patient stress. Tools that employ SER to measure students' involvement and emotional responses to lessons are gaining traction in the education sector as a means of improving students' learning experiences. As SER is included into devices and interfaces, it will make possible more natural and sympathetic interactions between humans and computers. Guidelines and laws protecting user rights, addressing privacy problems, and ensuring informed consent are all necessary because of ethical considerations.

## REFERENCES

- [1.] Michael Neumann, Ngoc Thang Vu "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech" arXiv:1706.00612v1 [cs.CL] 2 Jun 2017.
- [2.] Qinying Yuan "A Classroom Emotion Recognition Model Based on a Convolutional Neural Network Speech Emotion Algorithm" Occupational Therapy International 2022.
- [3.] Chenchen Huang, Wei Gong, Wenlong Fu, and Dongyu Feng "A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM" Mathematical Problems in Engineering 2014.
- [4.] Li Zheng, Qiao Li, Hua Ban, Shuhua Liu "Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest" IEEE 2018.
- [5.] Dr. N. Herald Anantha Rufus, M.Zaheer, S. A. V. Dolendrakumar "SPEECH EMOTION RECOGNITION USING DEEP LEARNING" International Research Journal of Modernization in Engineering Technology and Science 2022.
- [6.] Abdul Ajj Ansari, Ayush Kumar Singh "Speech Emotion Recognition using CNN" International Research Journal of Engineering and Technology (IRJET) 2020.
- [7.] Anushka Sandesara , Shilpi Parikh, Pratyay Sapovadiya, Mrugendrasinh Rahevar "A Comparative Study On Speech Emotion Recognition" International Journal of Research in Engineering, Science and Management 2020.
- [8.] Kaibei Peng, Liuyi Wu, Xiaoshu Wang "Speech Emotion Recognition Based on Convolutional Neural Network for Emergency System of Railway Station" CCIIS 2021.
- [9.] Apoorv Singh, Kshitij Kumar Srivastava, Harini Murugan "Speech Emotion Recognition Using Convolutional Neural Network (CNN)" International Journal of Psychosocial Rehabilitation 2020.

- [10.]B. Sandeep, Dr. R. Sivaranjani “Speech based Emotion Recognition using CNN Classifier” International Journal of Advance Research, Ideas and Innovations in Technology 2021.