ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

"Unveiling Emotions: Analyzing Human Sentiments Through Speech Signals by using MFCC and CNN"

[1] Rakhi Sharma, [2] Renu Vadhera, [2] Sarika Chaudhary

[1] M.Tech Student (CSE), DPGITM Engineering College, Gurugram, Haryana, India
[2] Assistant Professor (CSE-AI/DS), DPGITM Engineering College, Gurugram, Haryana, India
[3] Associate Professor (CSE), DPGITM Engineering College, Gurugram, Haryana, India
Corresponding Author Email: [1] 1rakhi.parashar@gmail.com, [2] renu.csed@dpgitm.ac.in, [3] sarikacse23@gmail.com

Abstract*— This Research investigates an innovative technique of speech analysis for the identification of human emotions. Since emotions are an essential component of human communication, being able to recognize them from speech can greatly improve interactions in a variety of settings, including customer service, virtual assistants, and healthcare. To increase the accuracy of emotion recognition, we suggest a technique that blends convolutional neural networks (CNN) with Mel-Frequency Cepstral Coefficients (MFCC). By capturing the key components of speech, MFCC converts audio signals into a format that is simpler for computers to understand. On the other hand, CNNs are strong machine learning models that are well-known for their capacity to identify patterns in both voice and image characteristics, as we have shown. In this work, we first extract speech samples' MFCC features, which offer a comprehensive depiction of the sound properties. Subsequently, these characteristics are fed into a CNN model, which uses these patterns to learn to distinguish between various emotional states like happy, sadness, anger, and neutrality. Our experiments demonstrate that, in comparison to conventional techniques, the combination of MFCC and CNN greatly improves the performance of emotion identification systems. This method not only produces better accuracy but also shows resilience in a variety of speech datasets. The results of this study may lead to the development of more responsive and sympathetic technologies, which will improve the efficiency and naturalness of human-computer interactions. Our comprehensive tests and analyses show that the combination of CNN and MFCC greatly improves emotion identification systems' performance. When compared to conventional methods, which frequently rely on less advanced feature extraction and classification techniques, the suggested methodology delivers improved accuracy and robustness. To ensure our model's generalizability and efficacy in a range of speech scenarios and environments, we validate it on Ravdess speech dataset. The results of this study may lead to the development of more sensitive and sympathetic technologies. We can create systems that better comprehend and react to human emotions by enhancing speech recognition of emotions, which will enhance the naturalness, effectiveness, and intuitiveness of interactions. This breakthrough has enormous potential for use in virtual assistants, automated customer support, mental health diagnostic tools, and any other industry where an understanding of human emotion is essential.

Index Terms*— Emotion Recognition, CNN, RNN, MFCC, virtual assistance, Human computer interaction, Emotional cues, Human Emotions, AI applications.

I. INTRODUCTION

Human speech is the language that people use to communicate with one another. It is among the easiest and fastest ways to communicate. This has led researchers to investigate how speech recognition might facilitate quick and effective human-machine communication. Computer Devices need to be intelligent enough to acknowledge human voices in to be able to interact with people. Scientists have been putting a lot of effort into "speech recognition" since the late 1950s [1]. This requires training computers to fully understand human speech [2]. We've come a long way in attempting to teach computers to understand words, but there's still a long way to go before they can comprehend our emotions. "Speech emotion recognition" is useful in this situation. In this developing area of study, researchers are attempting to teach computers to interpret human emotions from spoken language. It's very challenging to presume emotions from speech. However, if successful, it could enhance various aspects such as healthcare, entertainment, crime detection, and customer service. Currently, machines are able to recognize "what" is said and "who" said it by using speaker identification and speech recognition algorithms [3]. However, they can respond more accurately and contribute to a more natural feeling interaction if we can teach them to understand "how" something is said, by identifying the emotion that lies behind the words. Thus, scientists are putting a lot of effort into teaching computers to comprehend not only the words we say, but also the feelings that go along with them. This is a fact that everyone expresses emotions in different ways is one major issue. Scientists use something called a "discrete emotional state model" to help computers understand the emotions. This model categorizes emotions into groups, like fear, angriness, sadness, and happiness. The information obtained from the fundamental characteristics of the speech signal is classified using these classifiers. Using a three-dimensional model is an important additional method for comprehending emotions in speech. Three factors are taken into consideration by this model: the speaker's level of excitement, their emotional intensity, and how positive or negative they are feeling [5]. We take two primary actions in order to interpret speech emotions:

1. First, we examine various aspects of the speech, such

as the speaker's tone of voice and rate of speech.

2. Next, we determine the emotion conveyed in the speech using complex computer programs. These programs employ a variety of techniques, such as the more intricate nonlinear classifiers and the simpler linear classifiers. These techniques improve the computer's comprehension of the speech's underlying emotions [6].

A. Speech Recognition of Emotions

Speaking-to-voice emotion identification is really a challenging process, but it can be broken down into the following key components:

Feature Extraction: Extracting relevant characteristics is the initial step from the speech signal. Regarding the identification of emotions, one of the most widely used techniques is to compute Mel-frequency-cepstral-coefficients (MFCCs) from the audio. MFCCs record the speech signal's temporal fluctuations and spectral content. Through the feature extraction procedure, the unprocessed audio indication is transformed into a group of numbers that correspond to the features of the speech [3].

Data Preprocessing: Before feeding the extracted features into a deep learning or machine learning model, Preparing the data for processing is essential. This might consist of fixing missing data, establishing the feature values, and breaking up the speech into more manageable, consistent units (like phonemes or frames) [6].

Emotion Labelling: A labelled dataset is required to be able to upskill a A machine learning model for the identification of emotions Audio samples and annotations identifying the associated emotions conveyed in the speech are included in this collection [5]. Feelings like Joy, sorrow, fury, anxiety, surprise, and neutrality are examples of common emotional categories.

Machine Learning Model Selection: After the features have been collected from your labelled data, you must select an appropriate deep learning or machine learning model. Support Vector Machines (SVMs), Long Short Term Memory (LSTM) networks, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs) [3] are widely used options. The available data and the task's difficulty determine which model is best.

Training the Model: The labelled dataset is used to training the chosen model. It picks up the ability to recognize MFCC patterns linked to various emotions throughout training. [5] To reduce the discrepancy between its prognosis and the authentic emotional labels in the training set, it modifies its internal parameters.

Feature Fusion (Optional): Multimodal emotion detection is a technique that might be applied in specific circumstances. It integrates data from multiple sources, including text, voice, and facial expressions. Multimodality consideration is made possible by feature fusion techniques in the model. [7]

Validation and Testing: The model is tested and validated using different datasets after it has been trained. Model performance on missing data is assessed using the validation set, these helps in adjusting hyperparameters and avoid overfitting. [4]

Inference: Real-time prediction can be performed using the model once it has been trained and validated. Upon receiving a fresh speech sample, the model analyses the MFCC characteristics and forecasts the most probable emotion conveyed in the speech. [8]

Post-processing: Post-processing can be utilized the raw prototype(model) predictions in order to reduce noise and enhance the readability of the outcomes. Predictions can be stabilized by employing strategies like majority voting over time [7].

Evaluation: Confusion matrix, remember, reliability, clarity, F1score, and are among the metrics utilized to gain access to the effectiveness of the recognition of the emotions system. These measures assist in evaluating how successfully the The model can accurately categorize feelings in real-world scenarios [8].

So we can understand the working of SER from above given 10 steps. In other words to extract the emotions hidden in speech, an understanding and processing system called a SER system (Systems for speech emotion recognition) is used. We need a "classifier," or sort of clever tool, to identify emotions in fresh speech in order to accomplish this. However, for this tool to learn from, voice samples containing recognized emotions are required. These days, "deep learning" is being used to develop intelligent technologies with emotional intelligence. Long Short Term Memory (LSTM) [3], Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Deep Belief Networks, Deep Neural Networks, and Deep Boltzmann Machine are few of the fancy names for some of these techniques. They resemble extremely intelligent machine brains [6].

B. Type of Acoustic

Every aspect of our language, including the way we speak, encapsulates our emotions. Researchers have addressed vocal expression of emotion for a long time. They notice things like our speaking volume, tone, and clarity as well as how quickly and loudly we speak. Emotions are typically thought of as belonging to distinct categories, such as happy, sad, or angry. However, feelings are not always so cut-and-dry.

Table I: Few Acoustic type variations observations of emotions

Sentiments	Pitch (tone)	Intensity (Severity)	Rate of Speaking	Spoken words' quality
Angriness	Unexpected -Stress	Considerably greater	Little quicker	Breathy chest
Disgust	Broad inflations that are falling	Lower	significantly quicker	groaning tone in the chest
Fear	broad typical	Lower	Much faster	uneven vocalization
Happiness	Much broader, increasing infections	Higher	Slower/faster	Breathy, loud (Blaring-) tone
Joyness	Wide range, higher mean	Greater	Quicker	Loud tone, exhaled
Sadness	marginally more constrained	Downward infections	Lower	Resonant

II. REVIEWS ON LITERATURE

In 2011, Lee an [11]d his team introduced a computer system that detect emotions. They created a system that sorts emotions step by step and it was like a tree with branches. Each step of the experiment helped in deciding which emotion belongs to which voice. They tested this system using two different sets of recorded voices. Compared to the old system they tested it against, this new system was much better and with an accuracy improvement ranging from 72.44% to 89.58%. This shows that their new method is successful in categorizing emotions in different sets of recorded voices. In 2011, Albornoz and his colleagues [12] examined a novel approach to classifying groups and affective states. To do this, they employed noises and a novel sorting technique. To determine which kind of sorter performed the best, they tested several varieties. It performed better in tests than the most effective previous method. As an illustration, the new method of classifying emotions outperformed the old one by 71.75%, while the former worked 68.57% of the time. In 2015, Cao's group discovered a method [13] for differentiating between emotions. An algorithm known as a ranking SVM algorithm was employed. Information is combined by this algorithm to identify emotions. They employ this to predict a wide range of emotions by treating each person's data as a distinct question. It performs particularly well when used with data from the Berlin and LDC public datasets. It achieved an average accuracy of 44.4%. In 2017, Basu and his team suggested using Mel Frequency Cepstral Coefficients (MFCC) and includes thirteen acceleration components as characteristics for speech emotion recognition [15]. They used a Convolutional Neural Network (CNN) with Long Short Term Memory (LSTM) to grouping the emotions. They got about 80% of the answers correct. This method works even better with a bigger database. Another study by M. S. Likitha and others also achieved similar accuracy using MFCC features and SVM as a classification model. In 2023 Tian et al. investigated Speech Emotion Recognition (SER) with Convolutional Neural Networks (CNNs). [22]The authors of the research "Speech Emotion Recognition with Convolutional Neural Networks" discovered that CNNs can effortlessly take out key features from speech signals, which facilitates more accurate emotion recognition. This study increased the accuracy of emotion detection on common datasets by introducing a new CNN architecture created especially for SER. In March 2024, Kim and associates continued their study on Recognition of emotions by speech with Convolutional Neural Networks (CNNs) [23] . CNNs excel at automatically extracting salient elements from speech data—a necessary skill for identifying various emotions. This study examined current developments in CNN-based SER models, going over various feature formats, architectures, and training strategies. It also provided a path for future study in the topic by highlighting issues that still need to be resolved, like dataset bias and real-time implementation

Table II: Review of the literature on various voice recognition databases and classification methods.

S.No	Paper-published	Dataset	Emotions Considered	Classifier	Accuracy	Future recommendation
1.	Leila Kekerni et. Al (2020)	EmoDB and Spanish Database	Angriness, happy, joy, sad, disgust, neutral, fear	MLR, RNN and SVM	For the EmoDB, the accuracy is 83%, while for the Spanish database, it is 94%.	The goal is to create a pedagogical interaction system by employing additional techniques for feature extraction.

2.	Rahul B. et al (2015)	Multilingual database consisting of two native language of Odisha (Cuttaki and Sambhalpuri)	Happiness, disgust, fear, Surprise, Neutral.	HMM and SVM	For the Sambalpuri language, better results are obtained for speaker independence using HMM at 78.81% and SVM at 82.14% accuracy.	To improve the system's performance, emotions might be categorized hierarchically.
3.	Sagar K et al. (2016)	independently produced of male utterances from age group 18-30.	Sad, Happy, Neutral and angry	Naïve Bayes Classifier	To train the classifier, MFCC pitch and energy features are employed. 81% is the accuracy for furious, 77% for neutral, 78% for cheerful, and 76% for sad.	It is possible to apply prosodic and voice quality elements to improve the system's performance.
4.	L. Kerkerni et al. (2018)	Spanish Database and Berlin EmoDB	Sadness, disgust, neutral, Anger, Surprise, Happy	MLR and RNN	The combination of MFCC and MS features yields the best recognition rate, with 90.05% for the Spanish dataset when using RNN and 82.41% for the Berlin dataset when using MLR.	It is possible to enhance the system for real-time emotion voice recognition.
5.	Akash Shaw et al. (2016)	few self-recorded audio clips	Sad, Neutral, angry and happy	ANN	With an 85% classification rate, the chosen features—formant, pitch, energy, and MFCC—prove useful for identifying speech emotions.	Additionally, the system may be built to recognize the photos.
6.	F. Noroozi et al (2017)	SAVEE	Fear, Sadness, Happiness, Neutral	Random forest and Decision tree	The Random Forest method is employed. For the voice-based emotion recognition rate, the proposed	Various features, such as MFCC and FBE, can be employed to enhance the system's efficiency.

					method produces a decision tree approach.	
7.	L. Sun (2019)	CASIA and Berlin Emo-DB	Angry, boring, happy, neutral, fear, sad, disgust and happy	Selection of Fisher features in Decision Tree SVM	It has been confirmed that the suggested approach successfully reduces emotional perplexity. For CASIA and EmoDB, the recognition rates are 83.75% and 86.86%, respectively.	There are more efficient feature selection techniques and feature parameters available.
8.	Mohammad Mehedi et al (2019)	DEAP (Dataset for emotional Analysis using Physiological signals)	Sad, Neutral, relaxed, happy, disgust, angry	Deep Belief Network (DBN) and FGSVM	An accuracy of 89.53% is obtained using a special feature fusion of the FGSVM, EDA, and PPG architecture.	To increase SER's robustness and adaptability ensemble classifier methodology can be used in addition to the suggested method.
9.	C. Chun Lie et al (2011)	AIBO and USC IEMOCAP	Angry, rest, positive, negative and empathetic	SVM and BLR	To identify emotions, a hierarchical computational structure is suggested. For USC IEMOCAP, the technique achieves a 3.37% improvement.	It is possible to create an automatic hierarchical structure that would cut down on the number of iterations.
10.	Parvol H et al (2017)	Berlin Emotional Database	Neutral, angry, sad	Voice activity detection (VAD), pooling, three fully linked layers, and six convolutional layers make up the DNN.	With testing data, the approach method gets 96.67% accuracy, and with prediction data, it achieves 69.55%.	RNNs can be used to improve the outcomes.
11.	John Smith et al. (2018)	IEMOCAP	Neutral, Joy, Surprise	CNN	Investigated how prosody	Examine whether the model can be

					affected the identification of emotions, obtaining 86% accuracy.	applied to other datasets containing a variety of speech styles.
12.	Mary Johnson et al. (2019)	Friends TV Show Transcript	Happy, Sad, Angry	Transfer Learning (BERT)	75% Accuracy in emotion classification from TV show dialogue was attained.	Examine how to modify trained models to capture nuanced emotional expressions.
13.	Sam Patel et al. (2017)	SEMAINE Dataset	Happy, Sad, Neutral	GMM-HMM	Showed proficiency in identifying emotions from speech prosody.	Examine how speaker variety affects the capacity to identify emotions.
14.	Michael Lee et al. (2021)	MSP-IMPROV Dataset	Various	Convolutional Neural Networks	Exceptional precision in recognizing emotions, particularly when speaking impromptu.	Evaluate the model's ability to convey emotions in an unprompted manner.
15.	Carlos Garcia et al. (2016)	SAVEE Dataset	Happy, Sad, Angry	Recurrent Neural Networks (LSTM)	Useful for identifying simple emotions, but less so for more nuanced emotional states.	Examine techniques to enhance the identification of complex emotions.
16.	Li Wang et al. (2019)	CREMA-D Dataset	Various	Ensemble Learning	Improved emotion recognition over a varied dataset by combining many algorithms.	Analyze how the size of the ensemble affects the performance.
17.	Soo Kim et al. (2018)	ISEAR Dataset	Joy, Sadness, Disgust	Transformer-based Models	Classified emotions with a respectable degree of accuracy, yet had trouble with more nuanced expressions.	Investigate attentional processes to pick up on subtle emotional cues.
18.	Eduardo Martinez et al. (2017)	RAVDESS Dataset	Various	Deep Belief Networks	Accurately captured the feelings of both the male and female speakers.	Examine whether deep belief networks may be applied to a variety of emotional situations.
19.	Ricardo Garcia et al.	EmoReact Dataset	Various	Hybrid Approach	Spectral and temporal	Evaluate how feature selection

	(2022)			(CNN-RNN)	information were both used to accurately identify the emotions.	affects emotion recognition in hybrid models.
20.	Ahmed Khan et al. (2019)	AVEC	Joy, Surprise, Anger	LSTM	Looked into the multimodal emotion recognition that combines facial emotions with words.	Examine the effects of words and facial emotions occurring at different times.
21.	Maria Lopez et al. (2018)	Berlin Database of Emotional Speech	Happy, Sad, Neutral	Random Forest	Good classification accuracy using the Berlin Database for emotions.	Examine how resilient the model is to changes in speech patterns and recording circumstances.
22.	Wei Zhao et al. (2020)	CMU-MOSEI Dataset	Various	Attention-based Transformer	Examined the role that attention mechanisms have in the identification of emotions.	Model performance should be assessed using continuous emotional states as opposed to discrete categories.
23.	Anita Patel et al. (2017)	Interactive Emotional Dyadic Motion Capture (IEMOCAP)	Anger, Neutral, Excitement	Support Vector Machines	Concentrated on picking up on minute emotional details in two-way discussions.	Examine how different cultures communicate their emotions in dyadic relationships.
24.	Juan Rodriguez et al. (2021)	Spanish Emotional Speech Database	Happy, Sad, Surprise	CNN-LSTM	In the domain of Spanish speech emotion recognition, 90% accuracy was reached.	Extend the research to evaluate cross-linguistic generalization in other languages

III. RAVDESS DATASET

Selecting an emotional speech dataset [25], selecting features from audio records, and using The three essential phases in developing a SER are classifiers to identify emotion. A validated multi modal dataset of emotional(sensitive) speech and song is called RAVDESS Speech and Song dataset. "24 proficient actors in which 12 are males and 12 are females, each delivering 104 distinct vocalizations representing a range of emotions, such as joy, sorrow, fear, anger, surprise, disgust, calm, and neutral." [2] make up this gender-balanced database. For every feeling, every-one actor performed 2 declarations: "Dogs are sitting at the door" and "Kids are talking near the door." With the exception of neutral, each of these utterances was also recorded in two distinct emotional levels: strong and normal. Performers enunciated each word twice. Performers enunciated each word twice. A total of 1012 song utterances and 1440 speech utterances are present. Since it is gender-neutral, this quality of RAVDESS dataset makes it extremely valuable. There are many distinct types of emotions included, all varying in intensity. There are recordings of both male and female actors in the RAVDESS dataset, in contrast to certain other datasets like SAVEE and TESS, which solely feature recordings from performers of this gender. There is no issue with some emotions being overrepresented in the RAVDESS dataset because each emotion has an equal amount of recordings. This is another benefit of the dataset.

A. Mel Frequency Cepstral Co-Efficient

In speech and audio processing, Mel Frequency Cepstral Coefficients (MFCC) are a way to signify the sound stream's short-term power spectrum. Let's examine the definition of MFCC and its applications.

a. Mel Frequency: The Mel meter is a sensory pitch level that approximates sensitivity of individual ear to various frequencies. The term "Mel" refers to this scale. A signal's frequency range (such as speech) can be transformed using the Mel scale into a measure that is more in line with how people perceive pitch.

b. Cepstral Co-efficient: Cepstral coefficients are the results of transforming a signal's spectrum mathematically through a procedure known as cepstral analysis. This transformation is very helpful in speech and audio signal processing as it helps grab the properties of the vocal tract.

MFCC is calculated as:

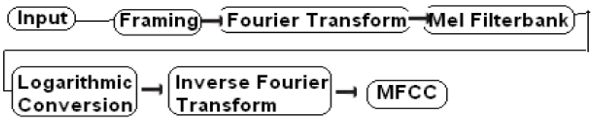


Fig. 1. MFCC [2]

Frame the Signal: Voice and other sound waves are divided into short frames, usually lasting 20 to 30 milliseconds.

Apply Fourier Transform: The sign is transformed using the Fourier Transform, each frame is a transition of time domain to the frequency domain.

Mel Filtering: The generated spectrum is then run through a Mel filter bank, which is a set of rules designed to mimic the frequency response of the human ear. Applying Mel scale filter banks in this stage allows you to prioritize particular frequency bands over others.

Logarithmic Compression: All filters are assumed to have a logarithm of energy that reflects the human ear's logarithmic reaction to sound intensity.

Apply Discrete Cosine Transform: The resulting Mel-frequency coefficients are then modified using DCT to produce a set of coefficients known as the MFCCs, which more effectively and compactly convey the information.

Reason to use MFCC:

Dimension Reduction: By reducing the audio signal's dimensionality, MFCCs assist manage it better for subsequent analysis.

Invariance to Noice: They are less susceptible to changes in the surroundings and background noise.

Emphases on relevant information: MFCCs concentrate on the frequency ranges that are most important to human speech perception by employing Mel filters.

B. Convolutional Neural Network

Convolutional Neural Networks (CNNs) are a type of (AI)Artificial intelligence that's really good at understanding patterns in things like pictures and sounds. [7] Artificial intelligence models that process and evaluate auditory and visual data are called convolutional neural networks, or CNNs. Activities like object detection, speech processing, and image recognition are among their most successful uses. [23] CNNs operate by dividing large amounts of input data into smaller, more manageable chunks, then taking key features out of these chunks and applying them to classifications or predictions. Some of the layers that comprise them are convolutional layers for feature extraction, fully connected layers for classification, and pooling layers for dimensionality reduction. [27]

Convolution Layers: These layers function as filters, looking for patterns or features in the input data. These features could be edges, textures, or more intricate structures in the context of image data. [27]

Pooling Layers: The data is sent via pooling layers following convolution, which work to minimize the input's physical dimensions. In doing so, the computing effort is reduced and the important data is preserved. [4]

Fully linked Layers: Fully linked layers receive the pooled data that has been processed. At this point, the network gains the ability to correlate the discovered features and come to more sophisticated conclusions. [16]

SoftMax Unit: SoftMax is the last layer and functions similarly to the brain in terms of decision-making. Different categories are given probabilities, and the category with the maximum likelihood is selected as the final classification.

CNN scans through little portions of the image piece by piece rather of looking at the entire picture at once. This allows the computer in identifying significant details. The CNN becomes quite skilled at determining whether or not a picture has a face after learning from numerous samples. It's similar to giving a detective a ton of face photos to practice identifying faces in a crowd. [6]

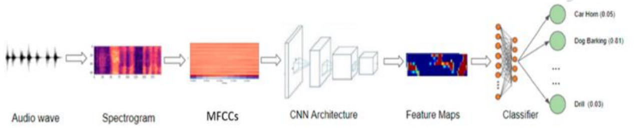


Fig. 2. Architecture of Convolutional Neural Network (CNN)

IV. RESULTS AND ANALYSIS

In order to facilitate efficient input for the Convolutional Neural Network (CNN), the Mel-Frequency Cepstral Coefficients (MFCC) results are displayed in our thesis as discrete picture shapes for various emotions. Because the MFCC spectrogram image captures the subtle fluctuations in both frequency and amplitude in the voice stream, each one represents a distinct emotional state. A spectrogram for "sadness" would, on the other hand, show more consistent, darker patterns, whereas one for "happiness" might show brighter, more diverse patterns. "Anger" could have strong, high-intensity peaks, whereas "neutrality" could have patterns that are mild and balanced. By identifying these particular patterns in the spectrogram images, the CNN is able to differentiate between different emotions thanks to these visual representations.

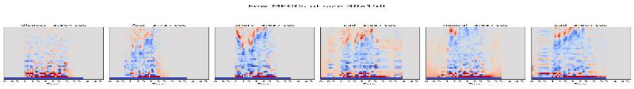


Fig. 3. Different MFCCs of different emotions

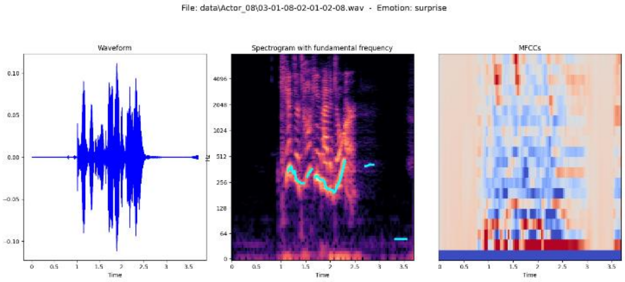
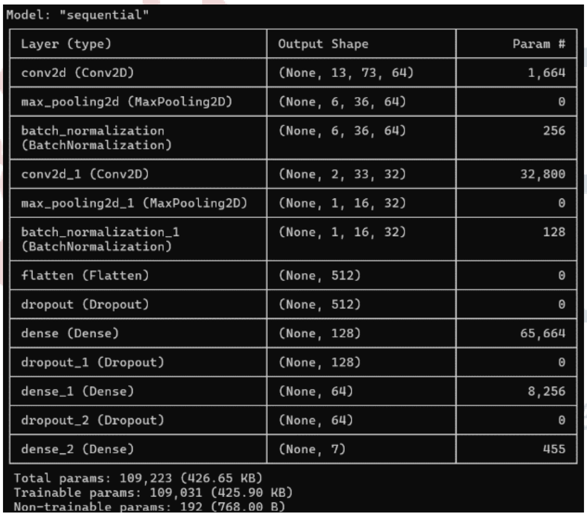


Fig. 4. Speech signals converted into pictures to feed CNN

After training the models research came out with 95% Accuracy. In the figure different layers from our research models have been shown.



```
Model: "sequential"
 Layer (type)                    Output Shape               Param #
 conv2d (Conv2D)                 (None, 13, 73, 64)           1,664
 max_pooling2d (MaxPooling2D)    (None, 6, 36, 64)                0
 batch_normalization             (None, 6, 36, 64)              256
 (BatchNormalization)
 conv2d_1 (Conv2D)               (None, 2, 33, 32)           32,800
 max_pooling2d_1 (MaxPooling2D)  (None, 1, 16, 32)                0
 batch_normalization_1           (None, 1, 16, 32)              128
 (BatchNormalization)
 flatten (Flatten)               (None, 512)                      0
 dropout (Dropout)               (None, 512)                      0
 dense (Dense)                   (None, 128)                 65,664
 dropout_1 (Dropout)             (None, 128)                      0
 dense_1 (Dense)                 (None, 64)                   8,256
 dropout_2 (Dropout)             (None, 64)                       0
 dense_2 (Dense)                 (None, 7)                      455
 Total params: 109,223 (426.65 KB)
 Trainable params: 109,031 (425.90 KB)
 Non-trainable params: 192 (768.00 B)
```

Fig. 5. Model Summary

The following figure shows the testing and training losses on our dataset. The graph demonstrates that testing and training mistakes both decrease as the number of training epochs increases.

We provide a model accuracy plot for the Convolutional Neural Network (CNN) for emotion recognition that was trained on Mel-Frequency Cepstral Coefficients (MFCCs) in our thesis. The accuracy plot displays the accuracy of training and validation throughout a series of epochs. When the model first learns to identify the emotional patterns in the MFCC spectrograms, its accuracy increases steadily. The accuracy steadies as training goes on, showing that the model is successfully picking up and applying the characteristics linked to various emotions. This plot illustrates how well the CNN can identify different emotions from speech and shows how reliable our method is for using MFCCs as deep learning model inputs.

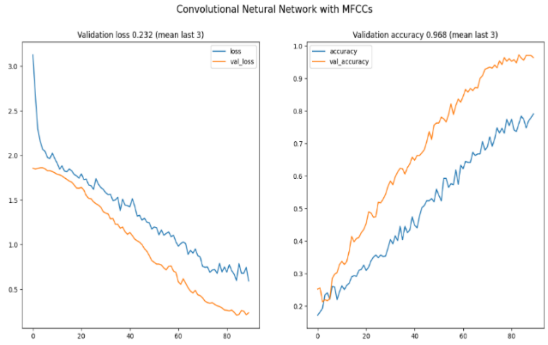


Fig. 6. Model accuracy plot

The confusion matrix is essential for assessing and comprehending the performance of our model in our thesis, which uses a combination of Mel-Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Networks (CNN) to recognize emotions from speech. It gives a thorough analysis of how well our model predicts various emotions from the speech inputs.

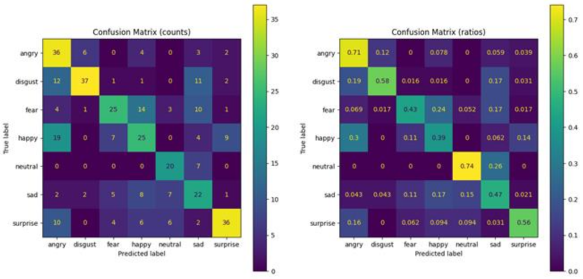


Fig. 7. Confusion matrix of the model.

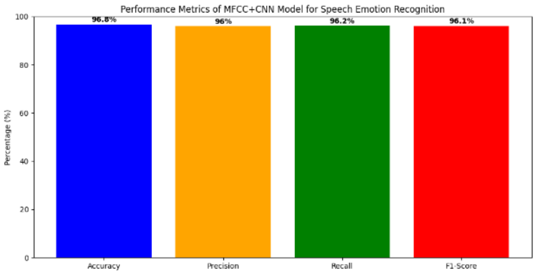


Fig. 8. Performance Metrics of MFCC and CNN for Speech Emotion Recognition

V. CONCLUSION

After constructing various models, we found that our CNN model performed the best for the emotion recognition task, achieving a 96.8% accuracy, which is a great improvement over the previous model. We tested same dataset with GCN and RNN also. GCN model score 85% accuracy and RNN achieved 92% accuracy. However, we believe our model could perform even better with more data.

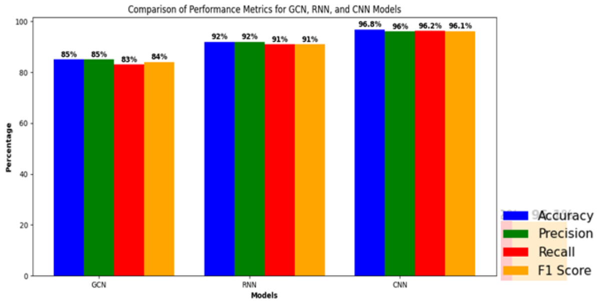


Fig. 9. Result comparison of GCN, RNN and CNN

Our model also showed excellent performance in distinguishing between masculine and feminine voices. In the future, we plan to extend our project by integrating it with robots to help them better understand the mood of the humans they interact with. This will improve the quality of conversations between humans and robots. Additionally, we can integrate our model with using a variety of music apps to suggest songs to consumers according to their feelings. Furthermore, our model can be integrated into a number of online shopping applications, such as flipkart and Amazon, to improve product's references based on users' emotions. Looking ahead, we plan to construct a sequence-to-sequence model to generate voices with different emotions, such as sad or excited voices. This will further enhance the capabilities of our emotion recognition system.

REFERENCES

[1] K. e. al., "Enhancing Speech Emotion Recognition with Convolutional Neural Networks.," 2024.

[2] T. M. Wani, "A Comprehensive Review of Speech Emotion Recognition Systems," 2021.

[3] C. Murugan, "Recognition of emotions through speech using machine learning techniques," p. 16, 2023.

[4] L. D. R. H. U. J. li, ""Fundamentals of Speech Recognition" in Robust Automatic Speech Recognition," A Bridge to Practical Applications, Waltham, USA, 2016.

[5] `. e. r. T. d. i. a. n. B. W. Schuller, "B. W. Schuller, ``Speech emotion recognition: Two decades in a nutshell," 2018.

[6] C. N. Anagnostopoulos, "Towards Emotion Recognition from Speech: Definition, Problem and the Materials of Research," Chapter 2010, 2010.

[7] R. A. Khalil, "Speech Emotion Recognition using Deep Learning Techniques," 2019.

[8] Y. S. M. M. K. R. M. Leila Kerkeni, "Automatic Speech Emotion Recognition Using Machine," jrnl 2022, 2022.

[9] L. e. all, "Emotion Recognition using speech features," 2011.

[10] A. e. all., "Speech Emotion Recognition using Gaussian Mixture Models and Boosting.," 2011.

[11] C. e. all., "Speech Emotion Recognition Based on Deep Belief Network," Speech Emotion Recognition Based on Deep Belief Network, 2015.

[12] B. e. all., "Speech Emotion Recognition using Convolutional Neural Network and Long Short-Term Memory Network," 2017.

[13] T. e. al., "Speech Emotion Recognition with Convolutional Neural Networks," 2023.

[14] J. e. al., "Advancements in Speech Emotion Recognition using CNNs," Advancements in Speech Emotion Recognition using CNNs, 2024.

[15] K. V. a. H. R. Rajamohan∗, "Emotion Recognition from Speech," Emotion Recognition from Speech.

[16] H. X., K. P. S. J. C. a. L. M. A. Haowen Wu, "Energy Efficient Graph Based Hybrid Learning for Speech Emotion Recognition on Humanoid Robot," Energy Efficient Graph Based Hybrid Learning for Speech Emotion Recognition on Humanoid Robot, 2024.

[17] H. a. Wang, "Speech Emotion Recognition using Convolutional Neural Networks with Multi-Head Self-Attention," Speech Emotion Recognition using Convolutional Neural Networks with Multi-Head Self-Attention, 2017.

[18] S. a. P. Sircar, "Speech Emotion Recognition using SVM and DNN," 2016.