IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306267|

Deep Insights into Speech Emotion Recognition: Integrating Automatic Speech Recognition with Convolutional Neural Network

Pavitra Nadar, Swati S. Chopade

P.G. Student, Department of Master of Computer Applications, Veermata Jijabai Technological Institute (VJTI), Mumbai, Maharashtra, India

Head of Department, Department of Master of Computer Applications, Veermata Jijabai Technological Institute (VJTI), Mumbai, Maharashtra, India

ABSTRACT: Emotion recognition from speech signals plays a crucial role in enhancing human-computer interaction by enabling machines to understand and respond to human emotions effectively. This project explores the integration of deep learning techniques with advanced automatic speech recognition (ASR) technology to achieve accurate and real-time emotion recognition from spoken language. The research focuses on developing a deep learning model optimized for emotion classification using Mel-frequency cepstral coefficients (MFCCs) extracted from speech signals. TensorFlow and Keras are utilized for model training and evaluation, with custom metrics such as F1-score ensuring robust performance across diverse emotional states including anger, joy, sadness, and more. In parallel, OpenAI's Whisper model is employed for robust ASR, facilitating precise transcription of spoken audio into textual form.

KEYWORDS: Emotion Recognition, Neural Networks, Speech Emotion Recognition, Automatic Speech Recognition, Whisper Model

I. INTRODUCTION

With the advent of machine learning techniques, particularly those within the realm of artificial intelligence, there has been a notable surge in research aimed at developing automated systems capable of accurately detecting and interpreting emotions conveyed through speech. In the rapidly advancing fields of artificial intelligence (AI) and machine learning, the capability to discern and interpret human emotions conveyed through speech represents a compelling and increasingly pertinent area of research. Emotion detection from speech signals not only enriches human-computer interaction but also holds significant practical implications across diverse sectors such as healthcare, customer service, and entertainment. This project embodies a convergence of sophisticated technologies aimed at achieving precise emotion recognition and seamless transcription from audio recordings. Understanding emotions expressed through speech is pivotal in enhancing the effectiveness of human-computer interfaces. By enabling machines to comprehend emotional nuances, interactions can be personalized and adapted to better meet user needs. In healthcare, for instance, the ability to detect subtle emotional cues in patient speech could aid in remote monitoring of mental health conditions, facilitating timely interventions and personalized care plans.

This project leverages advanced AI technologies to achieve its objectives. Central to its framework is a deep learning model trained specifically for emotion recognition from speech data. Built using TensorFlow and Keras, the model incorporates custom metrics such as F1-score to ensure robust performance in classifying emotions like anger, joy, sadness, and more. The model's accuracy is bolstered by preprocessing techniques that extract Mel-frequency cepstral coefficients (MFCCs) from audio signals, normalizing them to enhance consistency and reliability.

Complementing the emotion recognition model is the utilization of OpenAI's Whisper model for automatic speech recognition (ASR). Whisper is adept at transcribing spoken audio into text with high fidelity, making it a suitable companion technology to extract textual content from audio inputs. This integration not only enables the transcription of spoken words but also facilitates the analysis of emotional content embedded within the transcriptions.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306267|

To interact with and showcase these capabilities, the project employs Gradio, a platform for creating customizable and interactive user interfaces. Through Gradio, users can either record audio directly using a microphone or upload existing audio files for analysis. Upon initiating the process, the system provides real-time outputs, including both the transcribed text and the detected emotional states, displayed in an intuitive interface. This functionality enhances user engagement and understanding of the system's insights into spoken content.

II. LITERATURE SURVEY

Speech Emotion Recognition (SER) is a dynamic research field within digital signal processing that has seen significant advancements over the past decade. Researchers continually introduce innovative techniques to enhance SER performance and simplify system complexity. Typically, SER systems consist of two main components: robust feature selection from speech signals and accurate classification methods for emotion recognition.

Existing state-of-the-art SER methods apply deep Convolutional Neural Networks (CNNs) trained on spectral magnitude arrays of speech spectrograms [2-5]. To achieve high accuracy, complex CNN structures must be trained on a very large number (in the order of millions) of labelled spectrograms. This method (known as “fresh training”) is computationally intense, time consuming and requires large graphic processing units (GPUs).

In recent years [1], deep learning approaches, particularly 2D-CNN models, have gained popularity in SER for extracting hidden cues from spectrograms. Some researchers employ transfer learning techniques, using pre-trained models such as AlexNet and VGG on spectrograms to identify speakers' emotional states. Additionally, 2D-CNNs are utilized to extract spatial information from spectrograms, while Long Short-Term Memory (LSTM) or Recurrent Neural Networks (RNNs) capture sequential and temporal information from speech signals.

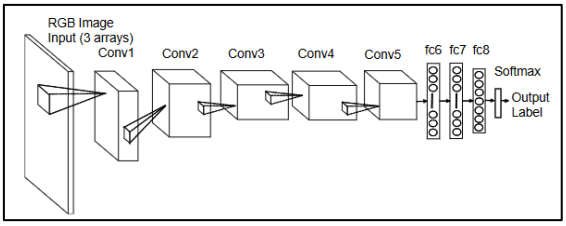

Figure 1: Structure of AlexNet. The input consists of 3 arrays (3 input channels). Feature extraction is performed by 5 convolutional layers (Conv1-5), and classification by 3 fully-connected layers (fc5-8). The output is given in the form of soft labels indicating probability of each class. (Source [1])

Researchers have investigated various input types for SER, including spectrograms, log Mel spectrograms, speech signals, and Mel Frequency Cepstral Coefficients (MFCCs), aiming to recognize speakers' emotional states. Some studies combine traditional approaches with advancements, leveraging pre-trained CNNs to extract salient information from audio data and using traditional classifiers for emotion classification [6]. Techniques have been developed to introduce new deep learning models for SER, utilizing datasets like RAVDESS [7] and IEMOCAP [8] with fewer parameters to achieve high accuracy and reduced computational complexity.

III. PROPOSED METHODOLOGY

The primary objective is to develop and deploy a deep learning model proficient in classifying a wide spectrum of emotions conveyed through speech. Emotions such as anger, disgust, fear, happiness, neutrality, sadness, and surprise are targeted for classification. Leveraging TensorFlow and Keras frameworks, the model is trained on extensive datasets annotated with emotional labels. This training equips the model to accurately analyze speech features and infer the underlying emotional states. Complementary to emotion recognition, the project integrates advanced automatic speech recognition (ASR) technology provided by OpenAI's Whisper model. Whisper excels in transcribing spoken audio into textual form with high fidelity, even in noisy environments or with diverse accents. The accurate transcription of spoken content into text enables further analysis and interaction, facilitating seamless integration with downstream applications such as sentiment analysis or semantic understanding. The utilization of deep learning,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306267|

particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), enables the model to effectively learn hierarchical representations of speech features. Mel-frequency cepstral coefficients (MFCCs) serve as primary features extracted from audio signals, capturing the spectral characteristics crucial for emotion classification. The model is trained using labeled datasets, fine-tuning its parameters to distinguish subtle variations in emotional expressions. OpenAI's Whisper model represents a significant advancement in ASR technology. Trained on vast corpora of diverse speech data, Whisper excels in transcribing spoken content accurately while adapting to varying speech patterns and acoustic environments. This robust ASR capability ensures that the transcribed text faithfully represents the spoken content, forming the foundation for subsequent emotion analysis and understanding.

A. DATASET SOURCES

1. Toronto Emotional Speech Set (TESS): This dataset consists of 2800 audio recordings featuring two female speakers (aged 26 and 64 years), each repeating seven different emotions while saying the phrase "Say the word" [9].
2. Surrey Audio-Visual Expressed Emotion (SAVEE): Comprising 480 audio files, SAVEE features recordings from four male speakers (aged 27 to 31), expressing 15 TIMIT sentences per emotion [9].
3. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): With 2452 audio files recorded by 12 male and 12 female speakers, RAVDESS features two repeated statements ("Kids are talking by the door" and "Dogs are sitting by the door") expressed in eight different emotions, each with two intensity levels [9].
4. Crowd-sourced Emotional Multimodal Actors Dataset (CREMA): This dataset comprises 7442 audio files recorded by a group of 48 male and 43 female speakers of various races and ethnicities. The recordings include 12 statements expressing six different emotions across four emotional levels [9].

B. PREPARATION

Dataset preparation plays a crucial role in the success and effectiveness of speech emotion recognition (SER) systems. A well-prepared dataset should accurately represent the diversity of emotions, speakers, languages, and cultural backgrounds encountered in real-world scenarios. Ensuring a diverse and representative dataset enables SER models to generalize better across different contexts and demographics. Incorporating variability in speaker characteristics such as age, gender, accent, and linguistic background is essential for developing robust SER models. Dataset preparation should include recordings from a diverse pool of speakers to capture the range of vocal expressions and acoustic features associated with different demographic groups. : A well-prepared dataset facilitates the development of SER models that generalize well to unseen data and exhibit robust performance across diverse settings.

EDA is performed to better understand the new dataset that has been created. According to the graph in Figure 2, all the emotions have almost equivalent count of audio files except for the emotion ‘surprise’.

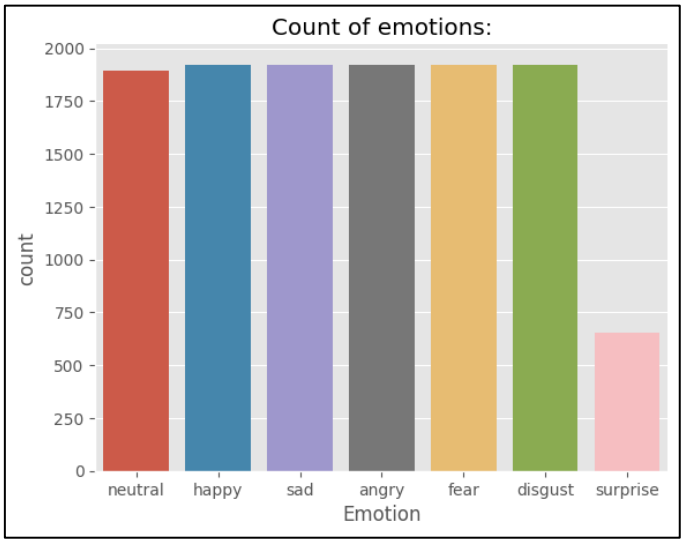

Figure 2: Distribution of emotions in the dataset.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306267|

C. DATA AUGMENTATION

Data augmentation techniques are used to increase the size and diversity of a dataset by applying transformations to existing data samples while preserving their semantic content. The four data augmentation techniques are as follows.

1. Noise Injection: This technique involves adding background noise to audio recordings. By introducing various types and levels of noise, such as white noise, pink noise, or environmental sounds, the model becomes more robust to noisy environments. Noise injection helps the model learn to focus on the relevant signal amidst background noise, improving its ability to generalize to real-world scenarios.

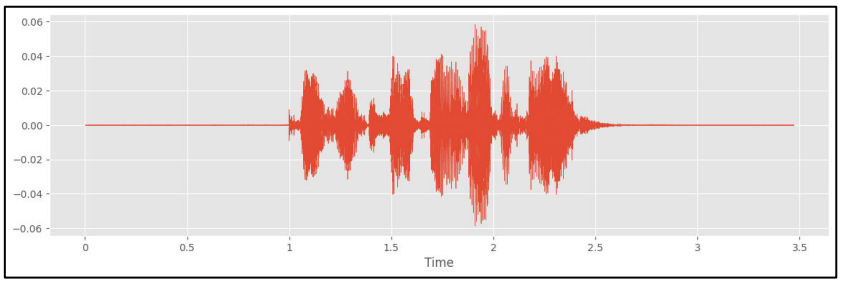

Figure 3: Noised audio of a file that is classified as exhibiting happy emotion.

2. Stretching: Stretching, involves altering the duration of audio recordings while preserving their pitch and content. This technique is commonly used to simulate variations in speaking rates or to generate longer or shorter versions of audio samples. By compressing the time axis, the model learns to recognize emotions across different speech speeds or durations.

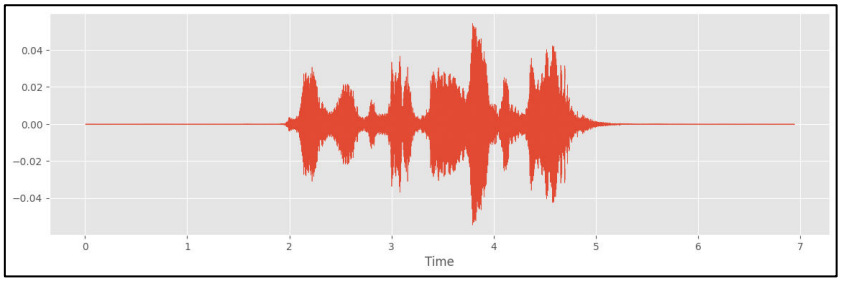

Figure 4: Stretched audio of a file that is classified as exhibiting happy emotion.

3. Shifting: Shifting, or time shifting, involves shifting the entire waveform along the time axis. This technique is used to simulate temporal variations in speech, such as slight changes in timing or onset/offset points. By randomly shifting the waveform within a specified time range, the model learns to recognize emotions despite temporal variations in speech patterns.

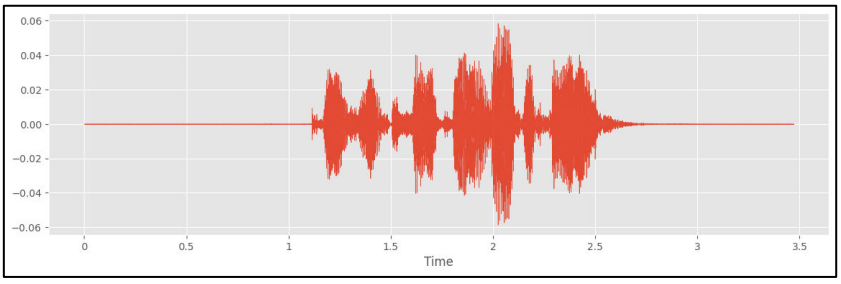

Figure 5: Shifted audio of a file that is classified as exhibiting happy emotion.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306267|

4. Pitching: Pitching, also known as pitch shifting, involves altering the pitch or frequency of audio recordings while preserving their duration and content. This technique is commonly used to simulate variations in pitch, intonation, or vocal characteristics. By shifting the pitch up or down by a certain number of semitones or percentage, the model learns to recognize emotions across different vocal registers or pitch ranges.

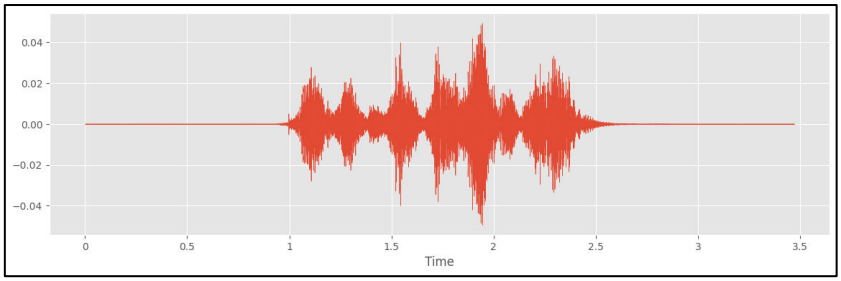

Figure 6: Pitch shifted audio of a file that is classified as exhibiting happy emotion.

For our data augmentation we will use noise and pitch and combination with both of it.

D. FEATURE EXTRACTION

1. Zero Crossing Rate: The frequency at which the signal changes sign within a given frame.
2. Energy: The normalized sum of squares of the signal values within a frame.
3. Entropy of Energy: A measure of the randomness in the distribution of sub-frame normalized energies, indicating abrupt changes.
4. Spectral Centroid: The "center of mass" of the spectrum, indicating where the center of gravity of the spectral distribution is located.
5. Spectral Spread: The variance or second central moment of the spectrum, reflecting the spread of the spectral components.
6. Spectral Entropy: The entropy of the normalized spectral energies across sub-frames, indicating the complexity of the spectrum.
7. Spectral Flux: The squared difference in the normalized magnitudes of spectra between successive frames, indicating spectral changes over time.
8. Spectral Roll-off: The frequency below which 90% of the spectral magnitude distribution is concentrated.
9. MFCCs (Mel Frequency Cepstral Coefficients): A set of coefficients representing the short-term power spectrum of a sound, where the frequency bands are spaced according to the mel scale, approximating the human ear's response more closely than the linear frequency scale.

```
ZCR:  (107,)
Energy:  (107,)
Entropy of Energy : (107,)
RMS : (107,)
Spectral Centroid : (107,)
Spectral Flux:  ()
Spectral Rollof:  (107,)
Chroma STFT:  (1284,)
MelSpectrogram:  (13696,)
MFCC:  (2140,)
```

Figure 7: The output after feature extraction.

Experimentally, it was decided to use just 3 main features for this task: ZCR, RMS and MFCC. Also in experimental way was decided to use just 2.5s duration with 0.6 offset - in the dataset first 0.6s contains no information about

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306267|

emotion, and most of them are less then 3s. Once data is processed, the features are saved into a dataframe for further processing.

E. MODEL

The Convolutional Neural Network (CNN) model, implemented using Keras with TensorFlow as the backend, plays a pivotal role in various machine learning tasks, particularly in image and signal processing domains. In this specific instance, a 1D CNN architecture is employed, tailored for processing sequential data such as audio signals. The model architecture begins with a 1D convolutional layer featuring 512 filters, each with a kernel size of 5 and a stride of 1, utilizing "same" padding to ensure consistency in output dimensions. This layer applies the Rectified Linear Unit (ReLU) activation function to introduce non-linearity into the network. The input shape of the data is defined by the shape of the training data (`X_train`), allowing for flexibility in handling inputs of varying dimensions.

To enhance the stability and speed of training, a batch normalization layer is incorporated after each convolutional layer. Batch normalization serves to normalize the activations of the previous layer, thereby mitigating issues related to internal covariate shift and accelerating convergence during training.

Following the convolutional layers, 1D max pooling layers are introduced with a pool size of 5, a stride of 2, and "same" padding. Max pooling is employed to downsample the spatial dimensions of the feature maps, preserving the most salient information while reducing computational complexity.

Subsequently, a flatten layer is added to convert the 2D feature maps into a 1D vector, facilitating the transition to fully connected layers. A fully connected layer with 512 units and ReLU activation function is then incorporated, allowing the network to learn high-level representations of the input data.

Finally, the output layer consists of 7 units, corresponding to the 7 possible emotions in the classification task. The softmax activation function is employed to compute the probability distribution over the different emotion classes, making the model suitable for multi-class classification problems.

Upon defining the model architecture, it is compiled using the RMSprop optimizer, categorical cross-entropy loss function, and additional metrics such as accuracy and a custom metric called f1_m, which evaluates the model's performance based on precision and recall.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 2376, 512)         3072
_________________________________________________________________
batch_normalization (BatchNo (None, 2376, 512)         2048
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 1188, 512)         0
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 1188, 512)         1311232
_________________________________________________________________
batch_normalization_1 (Batch (None, 1188, 512)         2048
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 594, 512)          0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 594, 256)          655616
_________________________________________________________________
batch_normalization_2 (Batch (None, 594, 256)          1024
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 297, 256)          0
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 297, 256)          196864
_________________________________________________________________
batch_normalization_3 (Batch (None, 297, 256)          1024
...
Total params: 7,193,223
Trainable params: 7,188,871
Non-trainable params: 4,352
```

Figure 8: Summary of the model.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306267|

Overall, this model consists of several convolutional layers with batch normalization and max pooling, followed by fully connected layers and an output layer.

F. AUTOMATIC SPEECH RECOGNITION

The choice of OpenAI's Whisper model for automatic speech recognition (ASR) in this project is driven by its advanced capabilities in handling diverse speech characteristics. Whisper has been specifically designed and trained on extensive datasets to excel in transcribing spoken audio into accurate textual representations [10]. Its robustness extends to accommodating various accents, mitigating background noise, and deciphering different speech patterns with high fidelity. These attributes are critical for ensuring reliable and precise transcription of spoken content, which forms the foundation for subsequent analysis and emotion classification. From the preprocessed text, the system extracts Mel-frequency cepstral coefficients (MFCCs) to capture acoustic features that are indicative of emotional expressions in speech. Utilizing TensorFlow and Keras, a deep learning model is employed to analyze these extracted features. The model is trained to classify emotional states such as anger, joy, sadness, and more, ensuring it interprets emotional cues effectively.

IV. EXPERIMENTAL RESULTS

The experimental results of applying a 1D Convolutional Neural Network (CNN) on a combined dataset of RAVDESS, CREMA, TESS, and SAVEE datasets indicate a significant achievement in speech emotion recognition. The reported accuracy of 96.03% suggests that the model successfully classified emotional states expressed in speech signals across various datasets.

By combining multiple datasets, the model benefits from increased diversity and variability in the data, which enhances its robustness and generalizability. This approach acknowledges the variability in speakers, emotions, and recording conditions present in different datasets.

The graphs in Figure 9 suggest that the model is performing well, with high accuracy and low loss on both training and testing data, though the difference between training and testing metrics indicates a potential for slight overfitting.

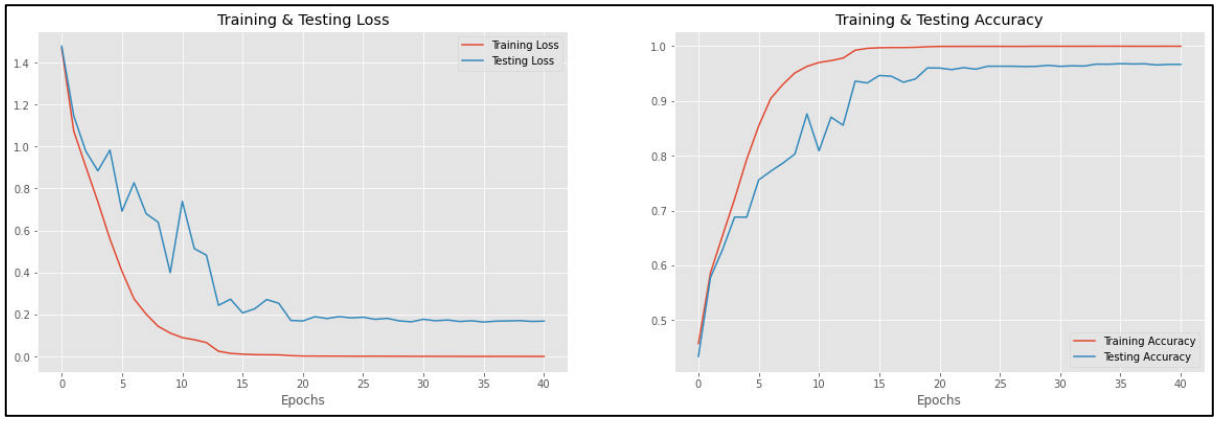

Figure 9: The graphs show the model's performance over 40 epochs, with training loss and accuracy improving significantly and stabilizing, while testing metrics indicate good generalization despite minor fluctuations.

The utilization of a 1D CNN architecture demonstrates the effectiveness of deep learning techniques in analyzing sequential data such as speech signals. 1D CNNs are adept at capturing local patterns and temporal dependencies, making them suitable for emotion recognition tasks.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306267|

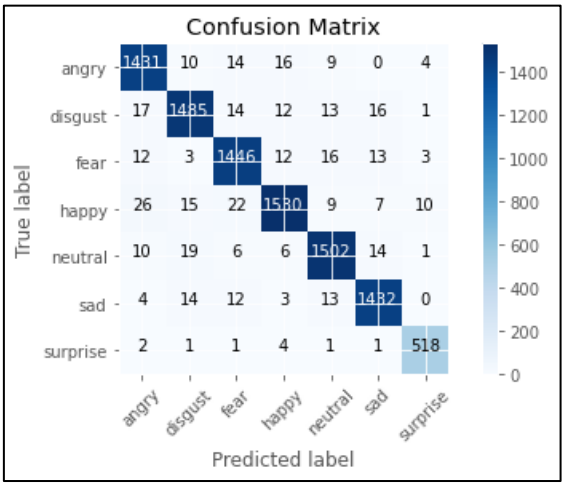

Figure 10: Confusion matrix between the actual and the predicted labels is shown.

The high accuracy obtained implies the model's proficiency in learning discriminative features from the combined dataset. This accomplishment holds promise for real-world applications of speech emotion recognition, including human-computer interaction, sentiment analysis, and emotion-aware systems in various domains.

This model is integrated with OpenAI's Whisper model within the Gradio framework to create a user-friendly interface. This allows users to either upload an audio file or record audio using their microphone. The combined models work together to transcribe the speech and detect the emotion conveyed in the speech.

V. CONCLUSION

This project marks a significant advancement in artificial intelligence, emotion recognition, and human-computer interaction by integrating deep learning, automatic speech recognition (ASR), and user interface design. It addresses key challenges in understanding and responding to emotions conveyed through speech. Integrating OpenAI's Whisper model for ASR ensures reliable transcription of spoken audio, handling accents, background noise, and speech variations effectively. Whisper's capabilities enhance data processing accuracy, making the system usable in real-world settings with diverse speech inputs.

Despite notable accuracy, potential limitations warrant further exploration. Additional metrics like precision, recall, and F1-score should be considered for comprehensive assessment. Experimenting with larger, diverse datasets and exploring alternative architectures or hybrid models could enhance accuracy and robustness.

The project's impact spans numerous applications. In healthcare, remote monitoring and emotion assessment can improve telemedicine and mental health support. In customer service, real-time emotion understanding enhances service quality and customer satisfaction. Educational applications benefit from insights into emotional dynamics, enabling personalized and supportive experiences.

In summary, this project demonstrates the transformative potential of AI-driven emotion recognition and speech processing technologies. By bridging human emotions and machine understanding, it advances technical capabilities and paves the way for empathetic and responsive human-computer interactions, enhancing human emotional experiences in meaningful ways.

REFERENCES

[1] M. Lech, M. Stolar, R. Bolia, M. Skinner "Amplitude-Frequency Analysis of Emotional Speech Using Transfer Learning and Classification of Spectrogram Images", Advances in Science, Technology and Engineering Systems Journal, vol. 3, no. 4, pp. 363-371 (2018).

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306267|

[2] H. Fayek, M. Lech, and L. Cavedon, “Towards real-time speech emotion recognition using deep neural networks”, ICSPCS, 14-16 December 2015, Cairns, Australia, pp. 1-6.
[3] H.M. Fayek, M. Lech and L. Cavedon L, “Evaluating deep learning architectures for speech emotion recognition”, Neural Networks, March 2017, Special Issue 21, pp. 1-11.
[4] Z. Huang M. Dong, Q. Mao, and Y. Zhan, “Speech emotion recognition using CNN”, ACM 2014, November 3-7, 2014, Orlando Florida, USA, pp. 801-804.
[5] W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and recurrent neural networks”. In Proceedings of the Signal and Information Processing Association Annual Summit and Conference, Jeju, Korea, December 2016, pp. 1–4.
[6] Mustaqeem and S. Kwon, "1d-cnn: speech emotion recognition system using a stacked network with dilated cnn features," Computers, Materials & Continua, vol. 67, no.3, pp. 4039–4059, 2021.
[7] S. R. Livingstone and F. A. Russo. (2018). “The ryerson audio-visual database of emotional speech and song (RAVDESSA dynamic, multimodal set of facial and vocal expressions in north american english,” PLoS One, vol. 13, no. 5, pp. e0196391.
[8] S. Kang, D. Kim and Y. Kim. (2019). “A visual-physiology multimodal system for detecting outlier behavior of participants in a reality TV show,” International Journal of Distributed Sensor Networks, vol. 15, pp. 1550147719864886.
[9] Dataset source https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en
[10] Radford, Alec & Kim, Jong & Xu, Tao & Brockman, Greg & McLeavey, Christine & Sutskever, Ilya. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details