Speech Emotion Recognition

Raushan Kumar, Sachin Meena, Harshit Jain

***Abstract*—Speech Emotion Recognition (SER) is an emerging field with significant potential for applications in various industries, including call centers, automotive safety systems, and customer satisfaction analysis. This paper presents a comprehensive approach to develop an effective SER classifier using deep learning techniques. The proposed methodology involves data preparation, data augmentation, feature extraction, and modeling using multiple datasets to train and test the classifier. This study paves the way for future research and development in the area of SER, with the potential to enhance human-computer interaction and improve safety and customer satisfaction.**

***Index Terms*—Speech Emotion Recognition, machine learning, feature extraction, Mel-Frequency Cepstral Coefficients, Chroma features, Convolutional Neural Network**

I. INTRODUCTION

Speech Emotion Recognition (SER) has attracted significant attention in recent years due to its potential to revolutionize human-computer interaction and enhance the understanding of human emotional states through speech. SER capitalizes on the observation that vocal expressions often reflect underlying emotions via tone, pitch, and other speech characteristics. This ability to recognize emotions from speech can be employed in various applications, such as call center analytics, in-vehicle safety systems, and mental health monitoring.

This paper aims to present a comprehensive approach to develop an effective SER classifier using deep learning techniques. We have used four different datasets, namely, Crowd-sourced Emotional Multimodal Actors Dataset (Crema-D), Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess), Surrey Audio-Visual Expressed Emotion (Savee), and Toronto Emotional Speech Set (Tess) to ensure a diverse range of emotional expressions. The methodology consists of several stages, including data preparation, data augmentation, feature extraction, and modeling.

In the data preparation stage, a dataframe was created to store all emotions and their corresponding file paths. This facilitated the extraction of relevant features for model training. During data augmentation, various techniques such as noise injection, time shifting, pitch alteration, and speed modification were employed to generate synthetic data samples. This step aimed to improve the model's ability to generalize and become invariant to perturbations.

Feature extraction plays a critical role in analyzing and establishing relationships between different aspects of the audio signal. To convert audio data into a format that can be processed by the model, we extracted eight features: Zero Crossing Rate, Chroma_stft, MFCC, RMS (root mean square) value, MelSpectrogram, spectral contrast, spectral centroid and spectral bandwidth. These features helped the model discern patterns and variations in the audio signals.

Finally, we trained and tested the classifier using the prepared data. The model achieved an overall accuracy of 63% on the test data, with higher accuracy for emotions such as surprise and anger. Although the model's performance was decent, further improvements can be made by exploring additional data augmentation techniques and alternative feature extraction methods.

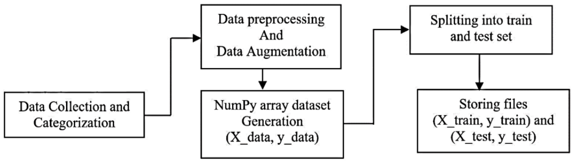
Data Collection and Categorization
Data preprocessing And Data Augmentation
NumPy array dataset Generation (X_data, y_data)
Splitting into train and test set
Storing files (X_train, y_train) and (X_test, y_test)

Fig. 1. Block diagram of data acquisition and train-test dataset generation

II. METHODOLOGY

In this section, we describe the methodology used to develop the SER classifier. The process consists of several stages: data preparation, data augmentation, feature extraction, and modeling. These stages are detailed below.

A. Data Preparation

As the study uses four different datasets [1] in English: Crema, Ravdess, Savee and Tess. Each of them contains audio in .wav format with some main labels.

1) Ravdess: Here is the filename identifiers as per the official RAVDESS website:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong).
  NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).

2) Crema: The third component is responsible for the emotion label:

3) Tess: Very similar to Crema - label of emotion is contained in the name of file.

Raushan Kumar, 200070068, Electrical Engineering, IIT Bombay
Sachin Meena, 200070069, Electrical Engineering, IIT Bombay
Harshit Jain, 20D070032, Electrical Engineering, IIT Bombay

Code	Label
SAD	sadness
ANG	angry
DIS	disgust
FEA	fear
HAP	happy
NEU	neutral

4) Savee: The audio files in this dataset are named in such a way that the prefix letters describes the emotion classes as follows:

Code	Label
a	anger
d	disgust
f	fear
h	happiness
n	neutral
sa	sadness
su	surprise

The first step involved creating a unified data frame containing all emotions and their corresponding file paths. This data frame was used to extract features required for model training. The datasets used in this study include Crema-D, Ravdess, Savee, and Tess. These datasets provide a diverse range of emotional expressions, enhancing the model's ability to recognize emotions from speech.

B. Data Visualization

Data visualization can be a powerful tool in speech emotion recognition (SER) as it allows for a better understanding of the data and can help identify patterns and trends. Here are some ways that data visualization can be used in SER:

1) Pie chart: The pie chart in figure 2 indicates the amount of percentage of data distribution among different type of emotions.

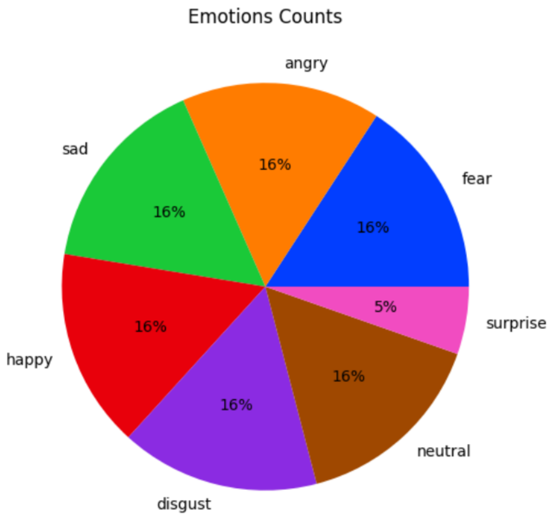

Fig. 2. Distribution of percentage of emotions in dataset

2) Waveform: A waveform is a visual representation of an audio signal. It shows the amplitude of the signal over time. Waveforms can be used to visualize the differences in speech signals between emotional states

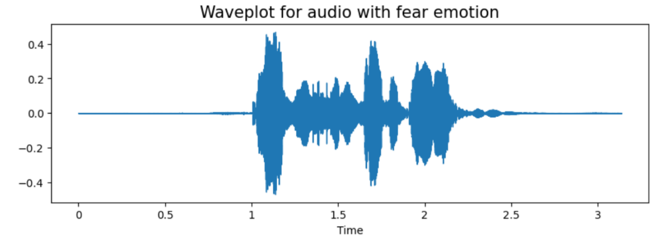

Fig. 3. wave plot of fear emotion

3) Spectrogram: A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. Spectrograms can be used to visualize the frequency content of speech signals and can help identify patterns and differences between emotional states.

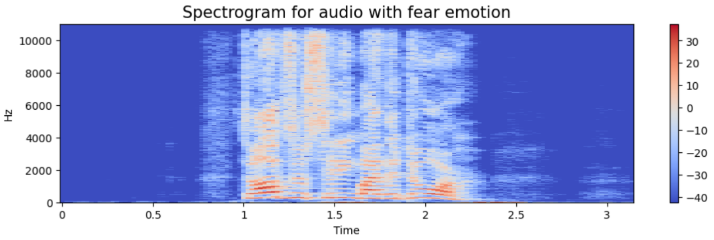

Fig. 4. Fear emotion spectrogram

Overall, data visualization can be a very useful tool in SER as it allows for a better understanding of the data and can help identify patterns and trends that may be difficult to identify through other means.

C. Data Augmentation

Data augmentation is an essential step in creating synthetic data samples by introducing small perturbations to the original training set. For audio data, we employed various augmentation techniques, such as noise injection, time shifting, pitch alteration, and speed modification. These techniques aimed to improve the model's ability to generalize and become invariant to perturbations while preserving the same label as the original training sample.

1) Noise Injection: This involves adding various types of noise, such as white noise or background noise, to the speech signal.

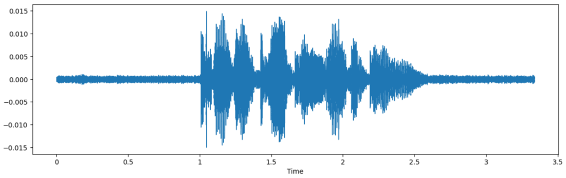

Fig. 5. wave plot after noise addition

2) Time stretching: This involves changing the duration of the speech signal without changing its pitch. This can be done by re-sampling the speech signal.

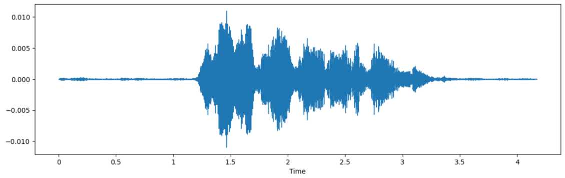

Fig. 6. wave plot after time stretching

3) Pitch shifting: This involves changing the pitch of the speech signal without changing its duration. This can be done by scaling the frequency of the speech signal.

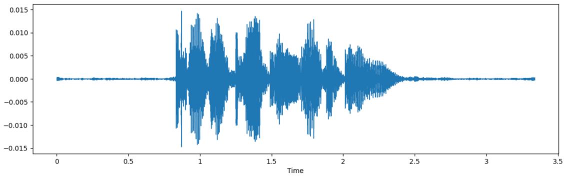

Fig. 7. wave plot after phase shifting

D. Feature Extraction

Feature extraction plays a crucial role in converting audio data into a format that the model can process. We extracted five features from the audio signals to enable the model to discern patterns and variations: Zero Crossing Rate, Chroma_stft, MFCC, RMS (mean root square) value, spectral contrast, spectral centroid, spectral bandwidth, and MelSpectrogram. These features provide meaningful information about the time, amplitude, and frequency characteristics of the audio signals.

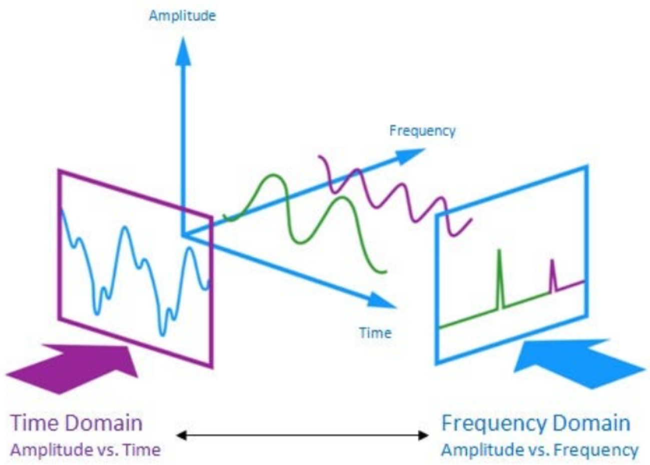

Fig. 8. Three-dimension feature extraction

E. Modeling

Once the features were extracted and the data was prepared, we proceeded with the modeling stage. In this step, we trained the classifier using deep learning techniques to recognize emotions from the processed data. The model's performance was evaluated on a test set, which was not used during training. The classifier achieved an overall accuracy of 63% on the test data, demonstrating its effectiveness in detecting emotions such as surprise and anger.

Throughout the methodology, we maintained a focus on technical rigor while ensuring a human-like approach to data processing and analysis. Combining deep learning techniques with appropriate feature extraction and data augmentation, our methodology offers a robust solution for SER applications in various industries.

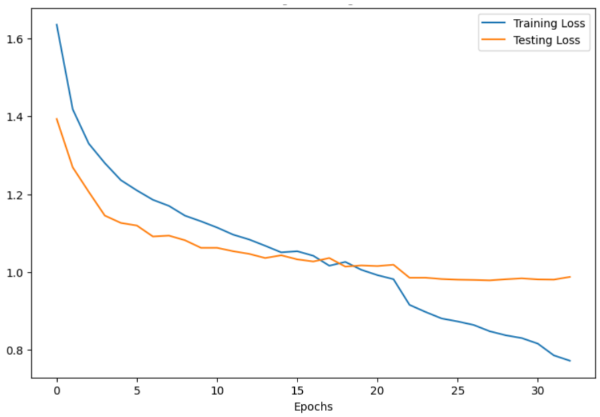

Fig. 9. Training and Testing loss

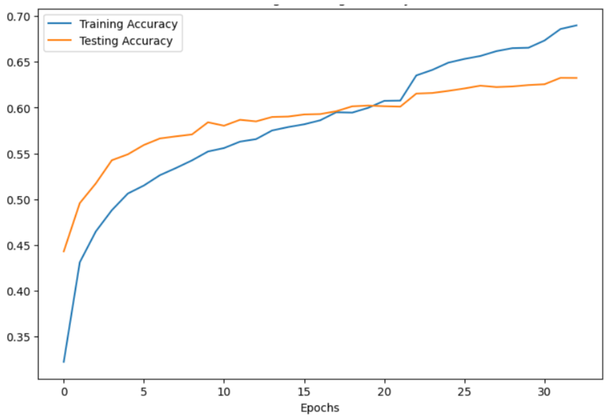

Fig. 10. Training and Testing accuracy

III. RESULTS

The results of speech emotion recognition are typically evaluated using performance metrics such as accuracy, precision, recall, and F1 score. These metrics provide an indication of how well the model is able to correctly classify the emotional state of a speaker based on their speech. In this project, we used machine learning models such as support vector machines (SVMs), decision trees, and neural networks. The performance of these models can vary depending on factors such as the size and quality of the training data, the feature extraction methods used, and the choice of the model architecture.

Recent advancements in deep learning, particularly with the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have led to significant improvements in the accuracy of speech-emotion recognition systems

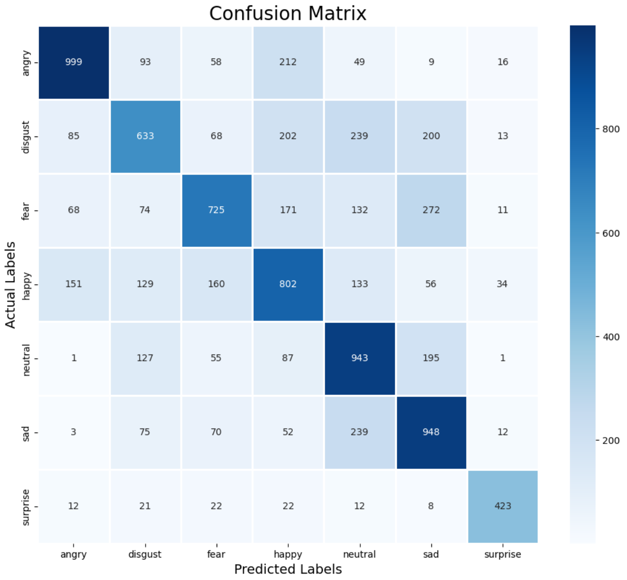

Fig. 11. Confusion matrix using convolutional neural networks (CNNs)

	Precision	Recall	F1-score	Support
angry	0.80	0.74	0.77	1476
disgust	0.51	0.51	0.51	1433
fear	0.63	0.56	0.59	1385
happy	0.59	0.54	0.57	1484
neutral	0.59	0.68	0.64	1403
sad	0.63	0.69	0.66	1466
surprise	0.78	0.84	0.81	475

accuracy			**0.63**	**9122**
macro avg	**0.65**	**0.65**	**0.65**	**9122**
weighted avg	**0.63**	**0.63**	**0.63**	**9122**

IV. Applications of SER

Speech Emotion Recognition (SER) has a wide range of applications across various industries, as it facilitates a more natural and intuitive human-computer interaction. Some of them are:

A. *Call Centers*

SER can be effectively utilized in call centers to classify calls according to emotions. By identifying dissatisfied customers or those experiencing strong emotions, call center agents can tailor their responses and prioritize issues.

B. *Mental Health Monitoring*

SER has the potential to play a vital role in mental health monitoring applications. By analyzing emotional patterns in speech, healthcare providers can identify potential signs of mental health issues and offer timely support to patients.

C. *Virtual Assistants and Chatbots*

Virtual assistants and chatbots can leverage SER to better understand user emotions, allowing them to provide more empathetic and contextually appropriate responses. By recognizing and adapting to the user's emotional state, these AI-driven systems can deliver a more personalized and human-like interaction experience.

V. Challenges

Developing an effective SER classifier is not without its challenges. In this section, we discuss the primary technical obstacles faced during the process of creating a robust and accurate speech emotion detection system.

1) Data Availability and Diversity: One of the main challenges faced in SER is the limited availability of diverse and high-quality datasets. While several datasets exist, they often lack variety in terms of speakers, languages, and emotional expressions. This limitation can hinder the model's ability to generalize and recognize emotions in real-world scenarios, where speech data is highly variable and context-dependent.

2) Modeling: Developing a classifier that can effectively recognize emotions from speech requires a thorough understanding of deep learning techniques and architectures. Selecting the appropriate model and fine-tuning its parameters can be a complex and iterative process. It is essential to strike a balance between model complexity and computational efficiency to prevent overfitting and ensure practical applicability.

VI. Conclusion

We build a classifier that recognizes emotions from speech data which acheived an overall accuracy of 63% on our test data and our model is more accurate in predicting surprise and angry emotions because audio files of these emotions differ from other audio files in a lot of ways like pitch and speed. We discussed the wide-ranging applications of SER, including call centers, mental health monitoring, virtual assistants. These applications highlight the potential of SER to revolutionize human-computer interaction, improve safety, and optimize business processes across various industries. We also acknowledged the technical challenges faced during the development of the SER classifier, such as data availability and diversity and modeling.

References

[1] https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en

[2] https://www.kaggle.com/code/ritzing/speech-emotion-recognition-with-cnn/notebook

[3] https://www.kaggle.com/code/ashishsingh226/speech-emotion-recognition-using-lstm

[4] Kottilingam. Kottursamy, ”A review on finding efficient approach to detect customer emotion analysis using deep learning analysis”, Journal of Trends in Computer Science and Smart Technology, vol. 3, no. 2, pp. 95-113, 2021.

[5] Amrita Thakur, Pujan Budhathoki, Sarmila Upreti, Shirish Shrestha and Subarna Shakya, ”Real Time Sign Language Recognition and Speech Generation”, Journal of Innovative Image Processing, vol. 2, no. 2, pp. 65-76, 2020.