Emotion in Triplicate: A CNN-Based Investigation into Varied Emotional Speech Datasets

Chowdhury Sabir Morshed
Roll no-232IT009
National Institute of Technology,Surathkal
Karnataka, India

Jai Narayan Singh
Roll no- 232IT016
National Institute of Technology,Surathkal
Karnataka, India

***Abstract*—Emotions play a crucial role in human life, serving as a vital means of expressing opinions and conveying one's physical and emotional well-being. Speech Emotion Recognition (SER) systems analyze audio signals to extract and predict the emotional tone of a speaker. Emotions are commonly categorized as Anger, Happiness, Sadness, or Neutral states, forming the basis for developing a SER system with the right resources. Through effective training, the system can learn to recognize a speaker's emotional state.**

Spectral and prosodic features are employed in speech emotion detection, as they provide essential data for determining emotional states. Mel-frequency Cepstral Coefficients (MFCC) are spectral features used in this process. Prosodic features, such as frequency, loudness, and pitch, contribute to training machine-learning models that can distinguish between various emotions in speech signals. The unique pitch of an audio signal can also be utilized to classify the speaker's gender.

Support Vector Machines (SVM), Radial-Basis Function (RBF), and Back Propagation networks are supervised learning models commonly employed in speech emotion recognition. SVMs, in particular, excel at regression and classification tasks, including gender identification in SER systems. Studies also leverage RBF and Back Propagation networks to identify human emotions in signals, focusing on specific features.

This study introduces a Speech Emotion Recognition system that surpasses existing models in terms of data, feature selection, and methodology. The goal is to enhance the accuracy of identifying speech patterns based on emotions, achieving an average accuracy rate of 55 percentage with fewer false positives. The paper emphasizes the significance of emotions in human communication, highlighting the role of SER in extracting and predicting emotional tones through audio signals. The integration of spectral and prosodic features, including MFCC, contributes to the system's effectiveness. Advanced machine learning models like SVM, RBF, and Back Propagation networks demonstrate superior performance, showcasing their potential in achieving high accuracy rates while minimizing false positives.

I. Introduction

Emotions are a big part of our lives. We express them through speech, facial expressions, and body language. But what about the emotions in our voices? That's where Speech Emotion Recognition (SER) comes in. It's like giving technology the ability to understand the feelings in our voices.

In the past, emotions were thought to be too tricky to study. But now, technology is helping us explore this emotional dimension. Imagine hearing aids that understand emotions, or call centers redirecting calls based on the caller's emotions. It's about making technology not just smart but emotionally aware.

Effective communication relies on understanding emotions. Without emotional expression, a speaker might seem dull or uninterested. Traditionally, measuring emotional content in speech has been challenging. But this study aims to make it easier. It's like creating a tool to understand the emotional tones in our voices, using methods familiar in identifying speech emotions.

We focus on Human-Computer Interaction (HCI). It wants to build a system that can detect emotions in speech, making technology more empathetic. As technology advances, especially in speech-centered devices, understanding emotional tones becomes crucial. But here's the challenge – there isn't enough emotional data in speech. It's like trying to solve a puzzle with missing pieces.

To tackle this challenge,we use techniques to represent emotional data in audio. It's like filling in the missing pieces of the puzzle. The study uses a type of technology called Echo State Network (ESN) to simplify the emotional data. Another tool called Sparse Random Projection (SRP) reduces extra information and focuses on the emotional part.

It uses speech emotion datasets as a foundation. These datasets are like books of notes for creating a system that understands emotions in different situations. Building Speech Emotion Recognition systems is not easy. Unlike other AI components, these systems need to understand human behavior. It's about creating not just smart machines but emotionally intelligent ones.

Recognizing a speaker's emotional state has many applications. From robots to online learning, the possibilities are vast. In online learning, knowing how students feel helps teachers adjust their strategies. Success in Speech Emotion Recognition depends on choosing the right emotional speech database, efficient feature extraction, and reliable machine learning classifiers.

Feature extraction is like isolating the emotional parts of speech. Many studies propose different emotional features like pitch, energy, and frequency. But choosing the right ones is crucial. Too many features make learning complicated, and too few lose the richness of emotions. It's about finding the right balance.

In the final stage, the system categorizes raw audio data.

It's like the climax of a play where machine learning models bring the system to life. There are different models like Neural Networks and Gaussian Mixture Models. Researchers propose various approaches, like combining different methods for better results.

The applications of recognizing emotions in speech are diverse. From Human-Computer Interaction to the Internet of Things, it's about making technology understand human emotions. In the IoT industry, applications like Alexa and Google Home depend on speech. This study predicts that 10 percent of IoT applications will rely on vocal commands soon.

Understanding speech signals is crucial, whether in self-driving cars or interactive applications. The study envisions a future where cars respond not just to commands but also to drivers' emotions. Call centers benefit from Speech Emotion Recognition too. Calls can be transferred based on the caller's emotions, providing a more personalized experience. The applications extend to lie-detection systems and humanoid robots.

In simple terms, Speech Emotion Recognition is a technology that understands emotions in voices. It's not just about algorithms; it's about decoding the emotional messages in our voices. In today's world, SER systems are a growing field, paving the way for a future where technology understands not just what we say but also how we feel. It's about creating a world where the conversation between humans and machines goes beyond the basics and embraces the full range of human emotions. This study is like a song celebrating the union of technology and emotion, showing how they dance together in our daily lives

II. Methodology

The main goal is to figure out how people feel by analyzing their voice. The system we're using, called Speech Emotion Recognition (SER), can do this by focusing on a feature called MFCC. Our main aim is to make this system work better, giving us accurate results and reducing mistakes.

People have already done a lot of research on this, mostly looking at the words used to express emotions. They often categorize emotions into three types: anger, joy, and neutral. One way they do this is by comparing the training data (recordings in a dataset) with new data (new recordings). Another way is by recognizing specific parts of speech that sound angry, happy, or neutral, and using a method called feature extraction along with a classifier called SVM.

In our system, we're using a feature called MFCC to group the data into different emotion categories. We're also using a tool called CNN, which is good at recognizing patterns like MFCC. CNN is straightforward and uses fewer settings to train the model, making it great for Speech Emotion Recognition. This helps us find a good balance between how much computing power we need and how accurate our system is in real-time.

Think of the SER system as a kind of smart learning model we've created. We've fine-tuned and adjusted it to work better, following steps similar to other smart learning projects. You can see a picture of how this all works in Figure 1.

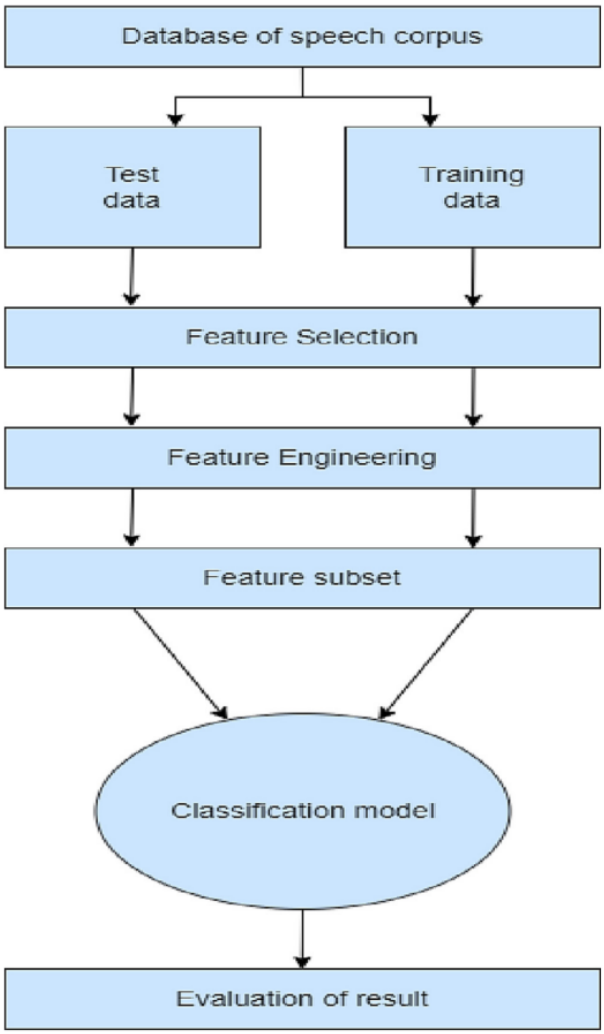

Fig. 1. Overview of the procedures in the Speech Emotion Recognition System

Now, the traditional way of doing Speech Emotion Recognition involves three steps. First, we clean up the speech signal to make it easier for the system to understand. Then, we extract features that show the emotional aspects of the speech. Finally, we use a model to recognize these emotional features. The last two steps are especially important and affect how well our system works, as you can see in Figure 1

III. Implementation

The Speech Emotion Recognition (SER) system is an ingenious application of artificial intelligence, specifically a convolutional neural network (CNN), designed to comprehend and categorize human emotions based on speech patterns. It's a fascinating exploration into the intersection of technology and human expression, where machines are taught to discern and interpret the intricate nuances of our emotions.

To embark on this journey, the first crucial step is gathering the right datasets. In the realm of SER, researchers often rely on various speech emotion datasets available on the internet. For our model, we've chosen three prominent datasets: RAVDESS, TESS, and CREMA-D. RAVDESS, standing for Ryerson Audio-Visual Database of Emotional Speech and Song, is particularly favored for its extensive collection and diversity. It features statements spoken by twenty-four professionally trained voice actors in a neutral American accent, covering emotions like sadness, anger, calmness, happiness, fear, surprise, and disgust.

The TESS dataset, an acronym for Toronto Emotional Speech Set, adds another layer of complexity to our training. Recruiting two English-fluent actresses from the Toronto locality, TESS offers recordings of pre-planned target words, portraying seven different emotions: anger, happiness, pleasant surprise, disgust, fear, sadness, and neutrality. The University of Toronto oversees the creation and curation of this dataset, ensuring its relevance and authenticity.

CREMA-D, short for the Crowd-sourced Emotional Multimodal Actors Dataset, contributes valuable vocal emotional expressions in sentences spoken in a range of basic emotional states. This dataset includes emotions like happiness, sadness, anger, fear, disgust, and neutrality. Each dataset brings its unique flavor to the training process, enriching our model's ability to comprehend a broad spectrum of emotional cues in speech.

Once armed with these datasets, the next step is data preprocessing. In the realm of machine learning, this is equivalent to getting the data dressed for success. Raw data is often noisy and disorganized, akin to a cluttered room. The preprocessing stage involves techniques to clean and organize the data, making it suitable for training the desired machine learning models. This could include correcting spelling errors, reducing the number of replicated characters, and disambiguating abstruse abbreviations.

For our SER model, we've taken it a step further. We've converted all uppercase letters into lowercase, ensuring uniformity in our data. Punctuation marks have been stripped away, simplifying the data for the model. It's like tidying up the room, removing unnecessary distractions, and creating an environment conducive to effective learning.

With the datasets cleaned and prepped, the next. In this case all the three datasets are ready to be fed into the model for training. The metaplot library comes to our aid at this stage, helping generate various graphical representations of the data. Visualizing the data is crucial, offering insights into its characteristics and distribution.

Data visualization, represented through charts, graphs, and maps, is an integral part of the process. Fig. 3, for instance, provides a count plot graphically displaying the total number of emotions present in the dataset. It's like having a visual menu of emotions, showcasing the diversity and distribution of emotional expressions in our dataset. Characteristics extraction follows the data preparation, a phase crucial for classification and problem depiction. The audio signal, es-

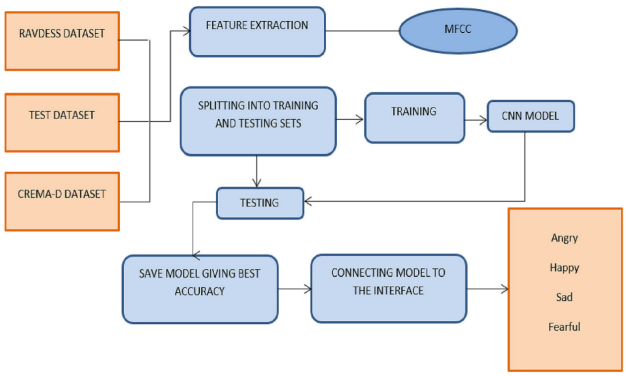

Fig. 2. Overview of the procedures in the Speech Emotion Recognition System

Dataset Name	No. of Speakers	No. of Emotions	Language	Duration	Sampling Rate	Annotation Method	Year
RAVDESS	24	8	English	1.5 hr	48 kHz	Actor-reported	2016
TESS	2	7	English	~1hr	16 kHz	Actor-reported	2005
CHREMA-D	91	6	English	~10hr	44.1 kHz	Crowd-sourced	2018

Fig. 3. Graphical representations of the data in the dataset

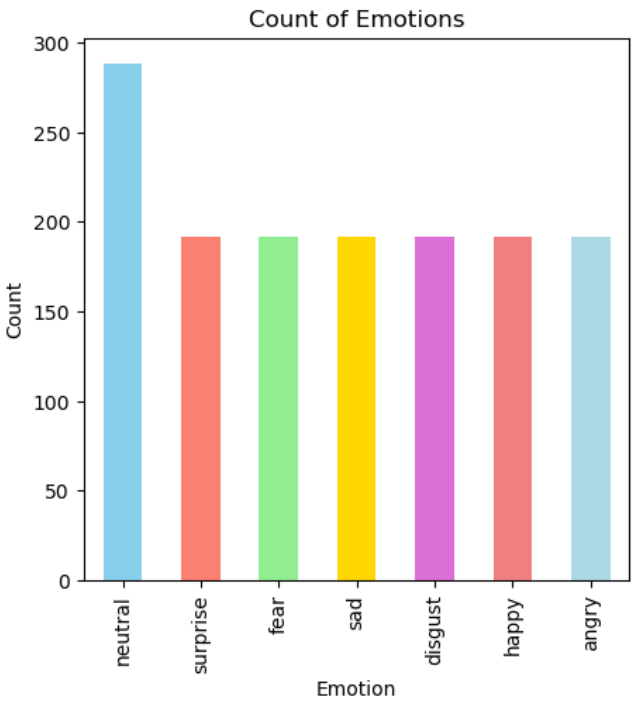

Fig. 4. Visualization of RAVDESS datset

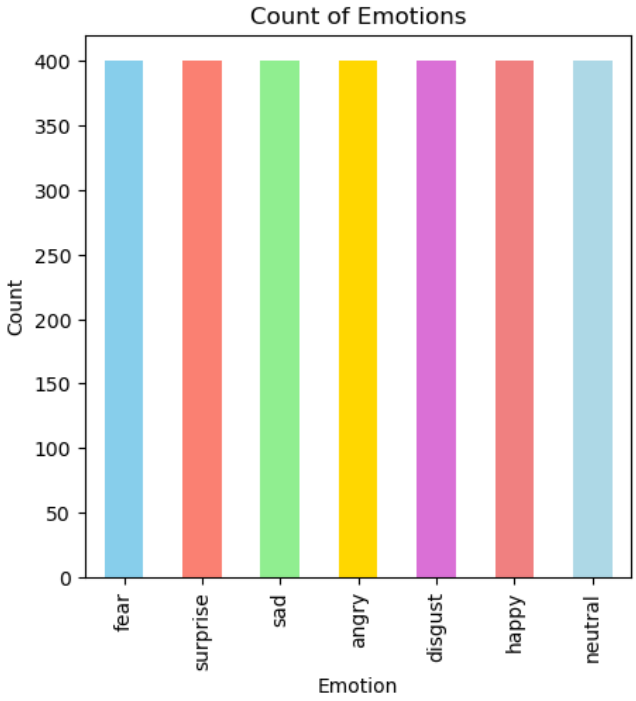

Fig. 5. Visualization of TESS dataset

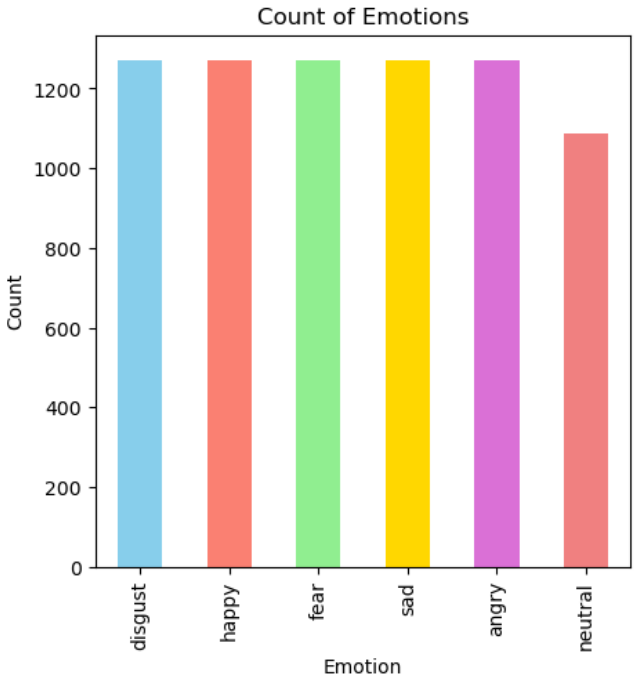

Fig. 6. Visualization of CREMA-D dataset

sentially a three-dimensional signal where the axes indicate time, amplitude, and frequency, becomes the focus. Librosa, a powerful Python library, becomes our tool of choice for this stage.

Librosa enables us to analyze and extract the required characteristics of the audio signal. Its display functionality is of vital importance, representing the audio files in various forms such as wave plots, spectrograms, and color maps. Wave plots, for example, utilize the loudness of the audio at a particular time, offering a visual representation of the audio signal's intensity over time. Spectrograms, on the other hand, display the various frequencies for a particular time with their amplitude, providing a frequency distribution map.

Feature extraction is the next vital phase in the process. Here, CNN deep learning algorithms are applied over the all three datasets. These techniques address challenges with data representation and data quality, refining the dataset for optimal learning. The sample audio serves as input, and for the audio files, the Spectrogram and Waveform are plotted.

Enter MFCC, or Mel-Frequency Cepstral Coefficients, a group of coefficients representing the short-term power spectrum of a sound signal. The Mel scale, a perceptual scale of pitches, is applied to convert the frequency values into a more human-friendly representation. The process involves using an algorithmic model constructed using the generated sampling rate value obtained using the Librosa package and MFCC function.

This stage, often considered the core of machine learning work, involves separating the data into training and testing sets. The CNN model, our brain in this analogy, is constructed using Keras, a high-level neural networks API. This model utilizes a machine learning algorithm to understand the data, training itself to respond to any new data it encounters.

As we initiate the training, the importance of MFCC values becomes apparent. Among all the features retrieved from the audio signal, MFCC values closely relate to the emotional tone of the speaker. It's like teaching the computer to focus on specific aspects of speech that carry emotional information. Speech, in this context, is represented as an image with three layers, and the CNN is designed to consider the first and second derivatives of the speech image with time and frequency.

CNN, or Convolutional Neural Network, is a powerful tool for pattern recognition. It can predict and analyze speech data, learning from speeches to identify words or utterances associated with different emotions. The process significantly reduces the dimensionality of the training dataset, decreasing the overall computation time of the model.

In essence, the process of Speech Emotion Recognition is akin to teaching a computer to understand the emotional content in human speech. It involves exposing the computer to examples, cleaning up the data, using visuals to understand emotional nuances, and evaluating its ability to grasp and interpret emotions. It's a journey toward creating machines that can navigate the intricate world of human emotions with finesse and understanding. As technology progresses, the

potential applications of SER extend beyond research labs, promising a future where machines can truly connect with us on an emotional level.

Steps carried out in the process of speech emotion recognition.

A. Data Collection

Data collection is the first and essential step in any machine learning project. In this step, speech data samples are collected from various sources, such as online databases, audio recordings,or user-generated content. The quality and diversity of the collected data determine the accuracy of the SER model.

B. Data Pre-processing

In this step, the collected speech data is pre-processed to improve the model's performance. The pre-processing step includes filtering out background noise, normalization of speech signals, segmentation of speech signals, and converting speech signals into a format suitable for feature extraction.

C. Data Splitting

The pre-processed and combined data is then split into two subsets: a training set and a testing set. The training set is used to train the SER model, while the testing set is used to evaluate the model's performance. Splitting the data ensures that the model is tested on data it has never seen before, and helps prevent overfitting.

D. Feature Extraction

Feature extraction is the process of extracting relevant features from the pre-processed speech data. In SER, features such as MFCC, recognize various emotions such as happiness, sadness, anger, fear,or neutrality.

E. Model Training

In this step, the extracted features are used to train the SER model. The SER model is built using machine learning algorithms CNN. The model is trained to recognize various emotions such as happiness, sadness, anger, fear,or neutrality.

IV. RESULTS

fig 7 summarize the performance of the CNN model on the RAVDESS test dataset.fig 8 summarizes the performance of the CNN model on the TESS test dataset.fig 9 summarizes the performance of the CNN model on the CREMA-D test dataset.fig 10 shows the normalized confusion matrix for the CNN model applied to the RAVDESS test dataset.fig 11 shows the normalized confusion matrix for the CNN model applied to the TESS test dataset.fig 12 shows the normalized confusion matrix for the CNN model applied to the CREMA-D test dataset In the SER system, the most commonly used normalization method is normalization by row, as it provides information about the accuracy of the model's predictions for each emotion class.The normalized confusion matrix can then be used to calculate evaluation metrics such as accuracy, precision, recall, and F1 score for each emotion class, which can help in improving the model's performance. To make the distribution of observed values more evident, it displays the various matrix values in distinct colours. The values in the row are all standard. The colours of the diagonal cells differ significantly from those of the other cells, which is an exact indication that the model performs relatively well. fig 13, 14,15 shows training and validation loss for RAVDESS, TESS and CREMA-D dataset respectively. Using RAVDESS dataset our model is detecting the emotions that belong to the set – Angry, Disgust, Fear,Happy, Neutral, Sad, Surprise and using CREMA-D dataset our model is detecting all emotion of later except surprise. Model using TESS dataset is also giving same set as that of model using RAVDESS dataset. Accuracy of the implemented model with respect to different datasets are represented in fig 16.

	precision	recall	f1-score	support
angry	0.56	0.61	0.59	44
disgust	0.47	0.33	0.39	27
fear	0.48	0.54	0.51	37
happy	0.21	0.15	0.18	33
neutral	0.53	0.69	0.60	67
sad	0.28	0.20	0.23	41
surprise	0.40	0.41	0.41	39
accuracy			0.45	288
macro avg	0.42	0.42	0.41	288
weighted avg	0.43	0.45	0.44	288

Fig. 7. Performance analysis of RAVDESS Dataset

	precision	recall	f1-score	support
angry	0.79	0.92	0.85	74
disgust	0.96	0.96	0.96	69
fear	0.88	0.99	0.93	74
happy	0.97	0.89	0.93	87
neutral	0.99	0.88	0.93	92
sad	1.00	0.91	0.95	85
surprise	0.89	0.95	0.92	79
accuracy			0.92	560
macro avg	0.93	0.93	0.92	560
weighted avg	0.93	0.92	0.92	560

Fig. 8. Performance analysis of TESS Dataset

	precision	recall	f1-score	support
angry	0.61	0.63	0.62	246
disgust	0.32	0.34	0.33	245
fear	0.36	0.17	0.23	252
happy	0.38	0.46	0.42	278
neutral	0.30	0.44	0.36	209
sad	0.47	0.38	0.42	259
accuracy			0.40	1489
macro avg	0.41	0.41	0.40	1489
weighted avg	0.41	0.40	0.40	1489

Fig. 9. Performance analysis of CREMA-D Dataset

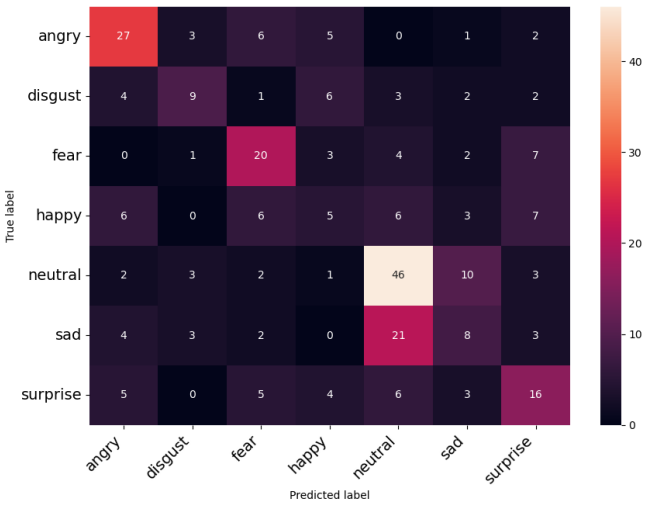

Fig. 10. Confusion matrix of RAVDESS Dataset

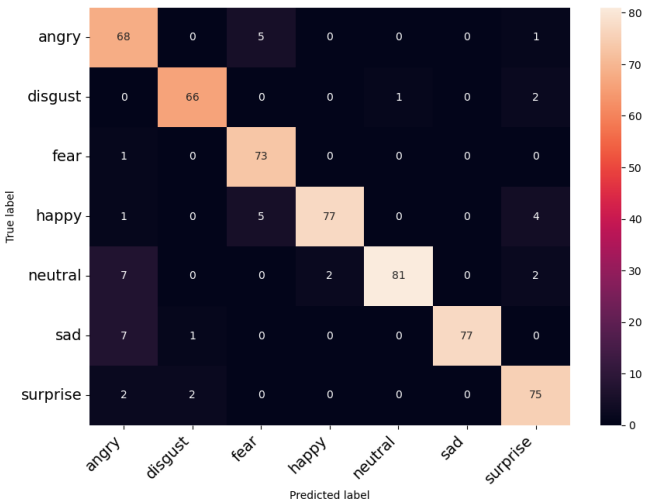

Fig. 11. Confusion matrix of TESS Dataset

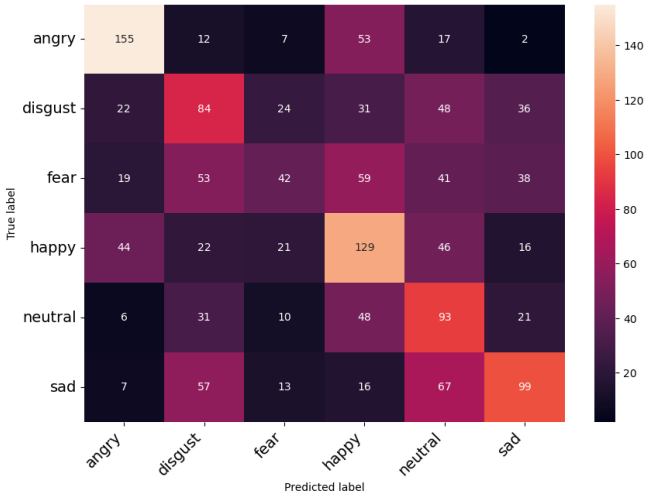

Fig. 12. Confusion matrix of CREMA-D Dataset

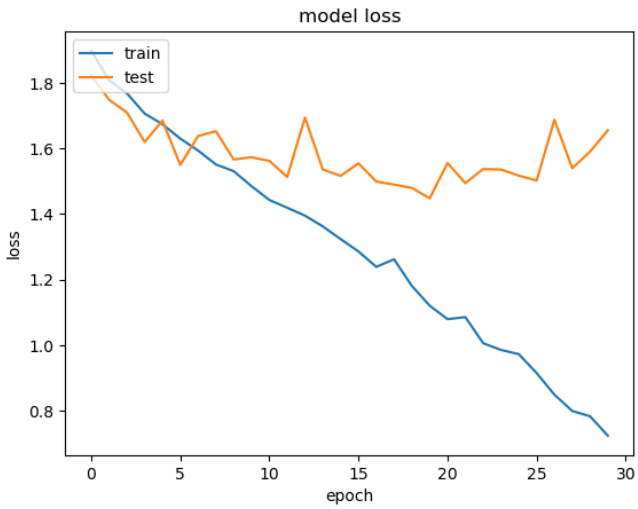

Fig. 13. Training and validation loss of RAVDESS Dataset

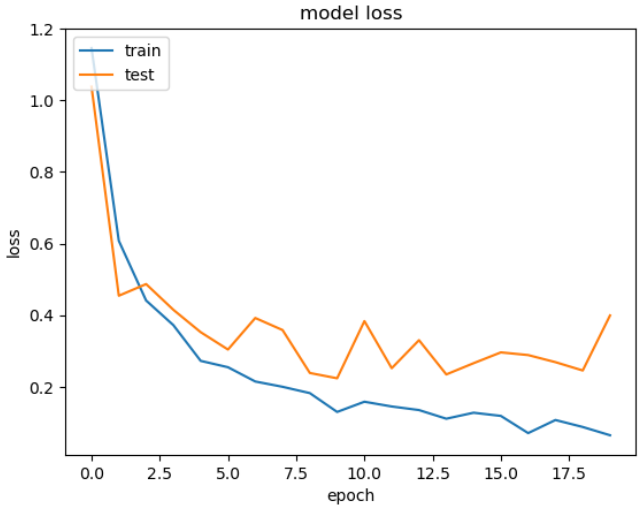

Fig. 14. Training and validation loss of TESS Dataset

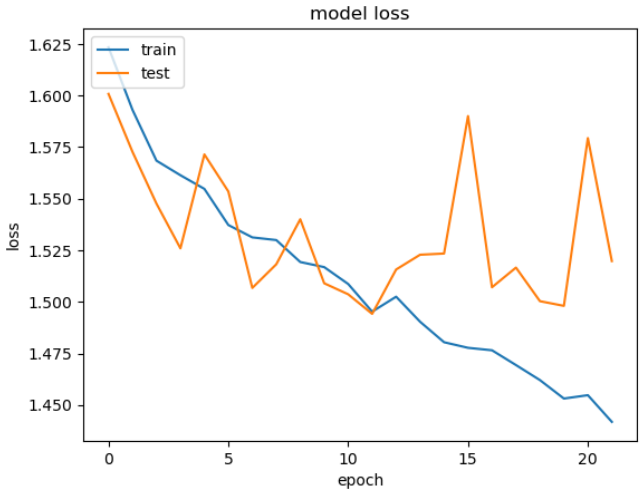

Fig. 15. Confusion matrix of RAVDESS Dataset

DATASET	Accuracy
RAVDESS	45.49%
TESS	92.32%
CREMA_D	40.43%

Fig. 16. Confusion matrix of RAVDESS Dataset

V. Conclusions and Future Enhancement

[htbp] In conclusion, our study on recognizing emotions in speech using three datasets—RAVDESS, TESS, and CREMA-D—via Convolutional Neural Networks (CNN) has given us some useful insights. We found that the TESS dataset performed the best in accurately classifying emotions.

The TESS dataset stood out, likely because of specific factors like its recording conditions, diverse speakers, and well-captured emotional expressions. These aspects helped create a strong training set for the CNN model, making it better at learning the subtle patterns linked to recognizing emotions in speech.

However, it's crucial to note that the success with the TESS dataset doesn't mean the RAVDESS and CREMA-D datasets are less important. Each dataset has its own characteristics and challenges, and the differences in performance emphasize the complexity of recognizing emotions in speech. Future studies could explore why the CNN model worked better on some datasets than others, like looking into unique dataset features or using techniques that let knowledge from one dataset improve performance on others.

To sum it up, our results emphasize how choosing the right dataset is key when building models for recognizing emotions in speech. Researchers and practitioners need to carefully pick datasets with diverse and representative data to train effective models. Understanding the specifics of each dataset helps interpret results and guides the development of better models for recognizing emotions in speech. section

References

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in International Conference on Machine Learning, 2016, pp. 173–182.

[2] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[3] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in Platform Technology and Service (PlatCon), 2017 International Conference on. IEEE, 2017, pp. 1–5.

[4] Anjali, Tripathi.; Upasana, Singh.; Garima, Bansal.; Rishabh, Gupta.; Ashutosh Kumar, Singh. "A Review on Emotion Detection and Classification using Speech". In Proceedings of the International Conference on Innovative Computing and Communications (ICICC), Online, 15 May 2020

[5] Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol 2018;21:93–120. https://doi.org/10.1007/s10772-018-9491-z.

[6] Google, "Cloud speech-to-text," http://cloud.google.com/speech-to-text/, 2018.

[7] Dong Yu and Li Deng, AUTOMATIC SPEECH RECOGNITION., Springer, 2016.

[8] Dario Bertero and Pascale Fung, "A first look into a convolutional neural network for speech emotion detection," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5115–5119

[9] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," Proc. Interspeech 2017, pp. 1263–1267, 2017.