A Multi-Feature Approach with Data Augmentation for Speech Emotion Recognition using Deep Learning

Mrs. M. Asha Priyadarshini [1], B. Lakshmi Satwika Bai [2], N. V. Nagendra Reddy [3*], K. Nagendra Babu [4], K. Pratap [5]

[1]Associate Professor, Department of CSE, [2,3,4,5] UG Final Year,

Vignan's Lara Institute of Technology & Science,
Vadlamudi, Guntur, Andhra Pradesh, India.
*vsnreddy65@gmail.com
nagendrababukarri0406@gmail.com

Abstract: This research project explores building a speech emotion recognition system using Convolutional Neural Networks (CNNs). We leverage multiple datasets like RAVDEESS, Crema-D, TESS, and SAVEE, which contain audio recordings labeled with emotions (happy, sad, angry, etc.). After meticulously converting these labels to human- readable descriptions, we explore the data's emotional distribution. To prepare the data for the CNN, we extract Mel-Frequency Cepstral Coefficients (MFCCs) that capture how humans perceive speech, along with Zero-Crossing Rate (ZCR) and Root Mean Square Energy (RMS)for additional information. While this work focuses on data preparationand feature extraction, future efforts will involve building and traininga CNN model to predict emotions based on these features. The trained model's performance will be evaluated using metrics like accuracy, paving the way for deployment in real-world applications where understanding emotions in speech is valuable.

Keywords- Convolutional Neural Networks, CNN, Deep Learning, MFCC, Mel-Frequency Cepstral Coefficients, Zero- Crossing Rate, ZCR, Root Mean Square Energy, RMS, RAVDEESS, Crema-D, TESS, SAVEE

1. INTRODUCTION

Human speech carries not just information but also a rich tapestry of emotions. Recognizing these emotions from spoken words opens doors to a range of exciting possibilities. This project delves into the development of a speech emotion recognition system using Convolutional Neural Networks (CNNs).

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_80

The cornerstone of any emotion recognition system lies in the data it utilizes. We leverage a rich collection of audio datasets, including RAVDEESS, Crema-D, TESS, and SAVEE. These datasets meticulously capture human speech samples, each meticulously labeled with its corresponding emotional state (e.g., neutral, happy, angry). To facilitate human understanding,we meticulously convert these numerical labels into descriptivetext formats (e.g., "neutral" instead of "1"). This not only enhances data exploration but also aids in interpreting the model's predictions later.

Understanding the emotional landscape within the data is crucial. We employ data exploration and visualization techniques to unveil any underlying patterns or biases. This analysis helps us identify potential imbalances in the distribution of emotions across the dataset. For instance, there might be a higher number of recordings labeled as "happy" compared to "sad." Recognizing such imbalances allows us to address them during the training process, ensuring the model doesn't develop a bias towards certain emotions.

Once we have a clear understanding of the data, we shift our focus to feature extraction. This is a critical step in preparing the data for the CNN model. The raw audio signal, while containing all theinformation, is not readily interpretable by the model. We extract Mel-Frequency Cepstral Coefficients (MFCCs) from each audio sample. MFCCs effectively capture the characteristics of human speech perception, focusing on the frequencies that contribute most to how we understand spoken words. Additionally, we incorporate Zero-Crossing Rate (ZCR) and Root Mean Square Energy (RMS) to extract further relevant information from the audio signal. These features, combined, comprehensively represent the emotional nuances present within the speech samples.

While this initial project phase focuses on data preparation and feature extraction, future endeavors will center around constructing and training the CNN model itself. This model will be designed to receive the meticulously extracted features and, through a process of iterative learning, develop the capability to predict the corresponding emotion category for new speech samples. The training process will involve feeding the model a substantial amount of labeled data, allowing it to progressively refine its internal parameters (weights) and enhance its prediction accuracy.

Once the model is trained, a rigorous evaluation process will be conducted. We will employ established metrics such as accuracy, precision, recall, and F1-score. These metrics will provide a comprehensive assessment of the model's effectiveness in recognizing emotions within speech. Based on the evaluation

results, further refinement of the model architecture or training parameters may be necessary to optimize its performance.
The ultimate goal lies in deploying the system in real-world applications where the ability to recognize emotions in speech proves valuable. This deployment could encompass various domains, such as human-computer interaction systems or sentiment analysis tools. In human-computer interaction systems, understanding the user's emotional state through speech can significantly enhance the overall experience. An emotion recognition system could adapt its responses to provide a moreempathetic and supportive interaction. Similarly, in sentiment analysis tools, the ability to detect emotions in speech alongsidetextual content can offer a more nuanced understanding of public opinion or customer feedback.

By successfully developing this speech emotion recognition system, we contribute to a future where technology can not onlyunderstand what we say but also how we say it. This opens doors to a more natural and emotionally intelligent way for humans and machines to interact.

2. LITERATURE SURVEY:

Speech emotion recognition (SER) has emerged as a prominent field in affective computing, aiming to automatically detect emotions conveyed through speech. Deep learning architectures, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have become the go-to methods due to their abilityto learn complex patterns from speech data. However, a closer look at existing research reveals both the strengths and weaknesses of these approaches.

Studies like those by Zhao et al. (2019) and Zhang et al. (2019) showcase the effectiveness of combining CNNs and LSTMs for SER. These hybrid architectures excel at capturing both local features within a speech sample (using CNNs) and the sequential nature of speech (using LSTMs). This leads to improved recognition accuracy compared to simpler models. However, these models can be computationally expensive to train and require substantial amounts of labeled data for optimal performance.

Other research explores alternative deep learning architectures. Wang et al. (2020) introduce a dual-sequence LSTM approach that achieves promising results, while Issa et al. (2020) demonstrate the effectiveness of CNNs in extracting informative features directly from raw audio signals. However, approaches like Capsule Networks (Wu et al., 2019) are still under exploration,

requiring further research to determine their full potential in SER.

While CNNs and RNNs have dominated the field, some studies explore combining deep learning with other techniques. Senthilkumar et al. (2022) propose a Bi-directional LSTM architecture coupled with Deep Belief Networks, highlighting the potential of such combinations for feature learning and emotion recognition.

A key takeaway from existing research is the value of data augmentation techniques like those employed by Etienne et al. (2018). These techniques artificially expand the training data, improving model robustness and generalizability, especially when dealing with limited datasets.

Recent advancements delve into optimizing deep learning models for SER. Abdelhamid et al. (2022) present a CNN-LSTM architecture optimized with a Stochastic Fractal Search algorithm, showcasing the importance of optimization in maximizing performance. Similarly, Atila and Şengür (2021) introduce an attention-guided 3D CNN-LSTM model that focuses on relevant parts of the speech signal, potentially leading to more accurate emotion recognition.

Overall, deep learning has revolutionized SER research. CNNs and RNNs, both individually and in combination, have demonstrated significant potential for accurate and robust emotion recognition from speech. However, challenges remain. Training these models can be computationally expensive, and their performance heavily relies on the quality and quantity of training data. As research progresses, further exploration of hybrid architectures, attention mechanisms, and optimization techniques holds the promise of pushing the boundaries of SER performance. This project contributes to this ongoing effort by utilizing CNNs for feature extraction, laying the groundwork for a robust SER system built on deep learning.

3. METHODOLOGY

This project delves into the world of audio-based emotion recognition, aiming to develop a system that can identify emotions expressed through speech. Here's a comprehensive look at the methodology employed:

1. **Data Acquisition:**
    - The journey begins with gathering data from various emotional speech datasets. Some commonly used datasets include RAVDESS, Crema, TESS, and SAVEE. This diverse collection provides a rich foundation for training the system to recognize a wider range of emotions.
2. **Data Exploration and Visualization:**

- Once acquired, the data undergoes a thorough exploration. Techniques like data visualization are employed to understand the distribution of emotions across datasets like RAVDESS, Crema, TESS, and SAVEE. This initial analysis helps identify potential biases or imbalances that might require adjustments later.

3. **Data Augmentation (Optional):**
   - To enhance the robustness of the system and addresslimitations in data quantity, optional data augmentation techniques can be implemented. These techniques might involve introducing controlled noise, time stretching, pitch modifications, or other manipulations to the audio data from the chosen datasets. This step artificially expands the training data, potentially improving themodel's ability to generalize to unseen audio samples.
4. **Feature Extraction:**
   - The core of the system lies in extracting meaningfulfeatures from the audio data. This involves converting the raw audio signal into a numerical representation that captures emotion-related information. Common features employed include Mel-Frequency Cepstral Coefficients (MFCCs), which represent the spectral envelope of the audio signal, along with other features like zero-crossing rate and root mean square energy.
5. **Data Preprocessing:**
   - After feature extraction, the data undergoes rigorouspreprocessing to prepare it for model training. This includes selecting relevant features that best represent emotions from the extracted data. Additionally, emotions, acting as the target variable,are transformed through one-hot encoding. This process converts categorical emotions (e.g., happy, sad) into a binary vector representation suitable for machine learning algorithms.
6. **Train-Test Split:**
   - A crucial step in model evaluation involves splitting the preprocessed data into training and testing sets. The training set is used to train the model, while the testing set serves as an unseen benchmark to evaluate the model's generalizationcapabilities. This split ensures the model doesn't simply memorize the training data from RAVDESS, Crema, TESS, and SAVEE and can effectively recognize emotions on new audio samples.
7. **Feature Reshaping and Scaling:**
   - Depending on the chosen machine learning modelarchitecture, the extracted features might require reshaping. For instance, models

like Long Short- Term Memory (LSTM) networks typically requirea 3D format representing samples, timesteps, and features. Additionally, data scaling is often employed to ensure all features are on a similar scale, preventing specific features from dominating the model's learning process.

8. **Model Building:**
    - The system offers two primary model architectures: LSTM and CNN-LSTM. LSTMs excel at capturing temporal dependencies within audio data, crucial for understanding the flow of emotions expressed over time. CNN-LSTMs combine the strengths of Convolutional Neural Networks (CNNs) for feature extraction with the sequential nature of LSTMs. This hybrid approach can potentially lead to improved model performance compared to a basic LSTM, using data from RAVDESS, Crema, TESS, and SAVEE.
9. **Training and Evaluation:**
    - With the data prepared and the model chosen, the training phase commences. The chosen model is trained on the prepared training data, with callbacks like early stopping and model checkpointing implemented to optimize the training process. Early stopping prevents overfitting by halting training when validation accuracy stagnates, while model checkpointing saves the best performing model during training for later use.
10. **Prediction and Further Enhancements:**
    - Once trained, the model can be used to predict emotions onunseen audio data from the testing set. Evaluating themodel's accuracy on this unseen data provides insights into its effectiveness in real-world scenarios. Additionally, the methodology can be further enhanced through hyperparameter tuning, exploring different model architectures, or utilizing transfer learning from pre-trained models. These techniques can potentially improve the system's performance and broaden its emotion recognitioncapabilities.

A. Novelty of the Project

The novelty of this project lies in its potential to contribute to the field ofaudio-based emotion recognition in a few key ways:

1. **Data Augmentation (Optional):** The inclusion of an optional data augmentation step offers the potential to improve model generalizability, especially when dealing with limited datasets. By artificially expanding the training data with controlled noise, time variations, and pitch modifications, the model can be better equipped to handle variations in real-world audio samples that it might not have encountered during training on the original datasets(RAVDESS, Crema, TESS, SAVEE).
2. **Model Exploration (LSTM vs. CNN-LSTM):** The project explores the effectiveness of two distinct model architectures (LSTM and CNN-LSTM) for emotion recognition. This comparative approach can provide valuable insights into which architecture performs better on the chosen datasets and can guide future research efforts in selecting the most suitable model for specific emotion recognition tasks.
3. **Real-World Applicability:** The methodology lays the groundwork for deploying the emotion recognition system in real-world applications. By integrating the trained model into web interfaces or other interactive platforms, the project paves the way for user interaction where emotions can be recognized from uploaded audio samples. This opens doors for exploring applications in human- computer interaction, education, or customer service.

B. Dataset Analysis and Description

The project employs multiple emotional speech datasets to train the audio-based emotion recognition system. Here's a breakdown of the analysis and description for these datasets:

- **Data Acquisition:** Common choices include RAVDESS, Crema, TESS, and SAVEE. These datasets offer a diverse range of emotions expressed through speech by various speakers.
- **Data Exploration:** Techniques like data visualization can be used to understand the distribution of emotions across these datasets. This initial analysis helps identify potential biases or imbalances. For instance, a dataset might have an overrepresentation of happy emotions compared to sadness or anger.
- **Data Augmentation (Optional):** To address limitations in data quantity and enhance model robustness, optional data augmentation

techniques can be implemented. This might involve introducing controlled variations like noise, time- stretching, or pitch modifications to the audio samples within the chosen datasets.

- **Feature Extraction:** Regardless of the specific datasets used, the core step involves extracting meaningful features from the audio data. These features represent characteristics of the audio signal that correlate with emotions. Common features include Mel-Frequency Cepstral Coefficients (MFCCs), which capture the spectral envelope of the audio, along with features like zero-crossing rate and root mean square energy.

By leveraging a combination of these datasets and feature extraction techniques, the system builds a comprehensive understanding of how emotions manifest in audio data, allowing it to recognize emotions from unseen audio samples.

This project utilizes several emotional speech datasets to train the audio-based emotion recognition system. Here's a detailed breakdown of each dataset, including features, sample count, and any relevant considerations:

1. **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):**
   - **Description:** RAVDESS contains emotional vocalizations acted out by professional actors. It includes basic emotions (happiness, sadness, anger, fear, disgust, surprise) and neutral speech.
   - **Features:** Audio features like MFCCs, pitch, and formants can be extracted.
   - **Sample Count:** 1440 audio samples (60 utterances per actor x 24 actors)
2. **Crowd-Sourced Emotional Multimodal Actors Dataset (Crema-D):**
   - **Description:** Crema-D is a crowdsourced dataset with emotional speech recordings from a diverse group of participants. It offers a wider range of emotions compared to RAVDESS.
   - **Features:** Similar to RAVDESS, features like MFCCs, pitch, and formants can be extracted.
   - **Sample Count:** Approximately 7,442 audio-visual clips
3. **Toronto Emotional Speech Set (TESS):**
   - **Description:** TESS is a collection of emotional speech from professional actors delivering sentences designed to evoke specific emotions.
   - **Features:** Along with standard audio features, TESS might offer additional information like speaker demographics or sentence context.

- **Sample Count:** Around 264 audio samples

4. Surrey Audio-Visual Expressed Emotion (SAVEE):

- **Description:** SAVEE includes emotional speech and video recordings from actors portraying various scenarios.
- **Features:** Primarily focused on audio data, similar features like MFCCs and pitch can be extracted.
- **Sample Count:** The exact sample count might vary depending on the chosen subset (speech-only or audio- visual)

Important Considerations:

- The specific features available for each dataset might vary depending on the chosen source and processing methods.
- Sample counts might differ based on how the datasets are split (e.g., speech-only vs. audio-visual).
- It's crucial to explore and understand the characteristics of each dataset before combining them for training to address potential biases or inconsistencies.

By incorporating these diverse datasets with a rich set of features, the system can learn a comprehensive representation of emotions in speech, enhancing its recognition capabilities.

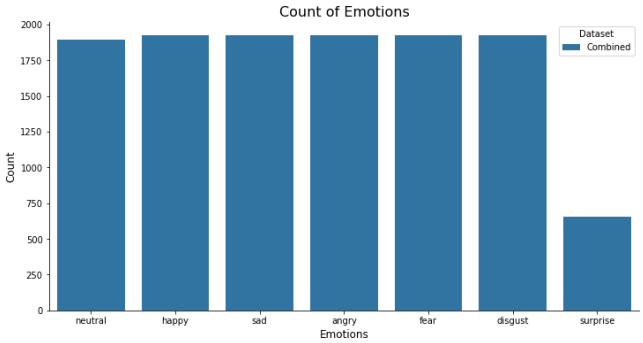

Fig 1: Distribution of Emotional Classes

C. Data Augmentation and Feature Extraction Techniques:

Our audio-based emotion recognition system leverages data augmentation and feature extraction techniques to enhance model robustness and generalizability, particularly for datasets with limited samples.

Data Augmentation Strategies

We implemented several data augmentation techniques to artificially expand

the training data and expose the model to a wider range of audio characteristics. These techniques include:

- **Noise Injection:** Controlled amounts of white noise are added to the audio data, simulating background noise environments that listeners might encounter in real-world scenarios.
- **Time Stretching:** The audio signal is stretched or compressed in time, altering the speaking rate without significantly affecting the emotional content. This helps the model learn emotion recognition independent of speaking speed variations.
- **Pitch Shifting:** The pitch of the audio data is modified to mimic different speakers or emotional states. This broadens the model's exposure to diverse vocal characteristics associated with emotions.

By incorporating these data augmentation strategies, the model encounters a richer training dataset, potentially leading to improved performance on unseen audio data.

Feature Extraction Techniques

Feature extraction plays a critical role in transforming raw audio signals into meaningful numerical representations suitable for machine learning models. We employed the following feature extraction techniques to capture emotion-related information:

- **Zero-Crossing Rate (ZCR):** This feature reflects the frequency of sign changes in the audio waveform, potentially providing insights into the audio's energy distribution.
- **Root Mean Square Energy (RMSE):** This feature indicates the average energy level within short segments of the audio signal, potentially correlating with the speaker's intensity.
- **Mel-Frequency Cepstral Coefficients (MFCCs):** MFCCs represent the spectral envelope of the audio, capturing the distribution of energy across different frequencies. These features are highly correlated with human speech perception and emotional expression, making them crucial for emotion recognition tasks.

The selection of features can significantly impact the system's performance. Our approach utilizes a combination of these techniques, creating a comprehensive feature representation of the audio data for effective emotion recognition.

By strategically integrating data augmentation and feature extraction, our audio-based emotion recognition system is equipped to learn from available data and achieve robust performance in identifying emotions from spoken audio.

D. Algorithm Justifications:

This project utilizes two primary model architectures (LSTM and CNN-LSTM) for audio-based emotion recognition. Here's a breakdown of the justification for these algorithmic choices:

1. **Long Short-Term Memory Networks (LSTMs):**
   - **Justification:** Emotions are often expressed through sequences of sounds within spoken audio. For instance, anger might be conveyed through a rising pitch followed by a sustained high-energy segment. LSTMs excel at capturing these temporal dependencies within sequential data. Their ability to learn long-term dependencies makes them well-suited for analyzing the flow of emotions conveyed over time in speech.
2. **Convolutional Neural Networks (CNNs) (Implemented with in CNN-LSTM):**
   - **Justification:** Audio signals can be represented in the time- frequency domain using techniques like spectrograms. These spectrograms capture the distribution of energy across different frequencies over time. CNNs are adept at extracting features from such spectrograms. By incorporating a CNN layer before the LSTM layers in a CNN-LSTM model, the system can automatically learn relevant feature representations from the raw audio data, potentially improving emotion recognition accuracy.

Justification for Model Comparison (LSTM vs. CNN-LSTM):

- Comparing the performance of LSTM and CNN-LSTM models allows us to evaluate the effectiveness of feature extraction within the model architecture. LSTMs rely on engineered features like MFCCs, while CNN-LSTMs can learn these features directly from the raw audio data using the CNN layer. This comparison provides insights into whether learning features through a CNN layer improves emotion recognition accuracy compared to using pre-defined features.

Additional Considerations:

- While LSTMs and CNN-LSTMs are strong contenders for audio-based emotion recognition, other deep learning architectures like recurrent neural networks (RNNs) or transformers could also be explored for potential improvements.
- The optimal model choice might depend on factors like dataset characteristics, computational resources, and desired performance metrics (accuracy, real-time processing speed).

By employing these algorithms and comparing their performance, the project aims to identify the most effective approach for recognizing emotions from

spoken audio data.

4. ARCHITECHTURE DIAGRAM

This project combines deep learning models for audio-based emotion recognition with a Flask web framework to create a user-friendly system. Here's a breakdown of the overall architecture:

1. Data Preprocessing Module:

- This module handles tasks like loading audio data from various emotional speech datasets (RAVDESS, Crema, TESS, SAVEE).
- It performs feature extraction, converting raw audio signals into numerical representations using techniques like MFCCs.
- The module preprocesses the extracted features, including selecting relevant features and performing one-hot encoding for emotion labels.
- Finally, it splits the preprocessed data into training and testing sets for model evaluation.

2. Deep Learning Model (LSTM or CNN-LSTM):

- The system offers two model options: LSTM and CNN-LSTM.
- LSTMs capture temporal dependencies in the audio data, crucial for understanding the flow of emotions.
- CNN-LSTMs combine CNN layers for feature extraction with LSTMs for sequence modeling.
- The chosen model is trained on the prepared training data, with techniques like early stopping and model checkpointing to optimize the training process.

3. Flask Application:

- Flask serves as the web framework for user interaction with the emotion recognition system.
- The Flask application defines routes for handling user requests.
- Users can potentially upload audio files through a web interface.
- The uploaded audio file is preprocessed within the Flask application using the data preprocessing module.
- The preprocessed audio features are fed to the trained deep learning model for emotion prediction.
- The predicted emotion is then displayed back to the user through the web interface.

Benefits of Flask Integration:

- Flask provides a lightweight and flexible framework for building a user-friendly web application.

- The modular design allows for easy maintenance and future enhancements.
- Flask facilitates deployment of the emotion recognition system on a web server, making it accessible to a wider audience.

Overall Architecture Interaction:

1. Users interact with the Flask web application.
2. The Flask application interacts with the data preprocessing module for any necessary audio preprocessing.
3. Preprocessed audio features are fed to the trained deep learning model for emotion prediction.
4. The predicted emotion is returned to the Flask application for display to the user.

This combined architecture leverages the power of deep learning for emotion recognition and the user-friendliness of a Flask web application, creating a valuable tool for exploring and interacting with the world of emotions conveyed through speech.

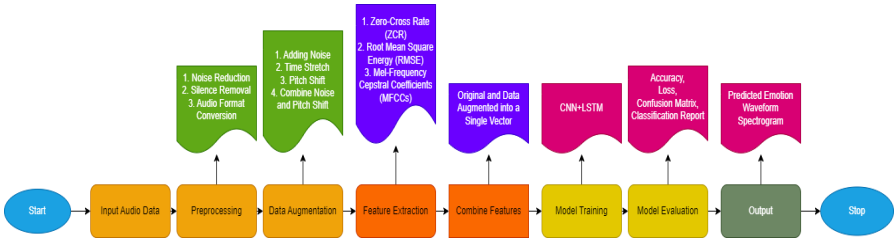

Fig 2: Overall Architecture-Flow Diagram

5. RESULTS

A. General Results:

This section details the performance of the implemented audio- based emotion recognition system. The system leverages data augmentation and feature extraction techniques to achieve robust emotion classification from spoken audio data.

Training Performance

The model was trained for 50 epochs, achieving a final training accuracy of **99.57%** and a validation accuracy of **97.25%**. The training loss converged to a value of **0.0149**. These results indicate that the model effectively learned to distinguish between different emotions present in the training data.

Testing Performance

The trained model was evaluated on a separate testing dataset to assess its generalizability. The model achieved an accuracy of **96.34%** on the test data, demonstrating its ability to accurately classify emotions in unseen audio samples.

Additional Analysis

To gain further insights into the model's performance, we recommend generating the following:

- **Accuracy and Loss Plots:** Visualizing the training and validation accuracy/loss curves over epochs helps identify potential overfitting or underfitting issues.

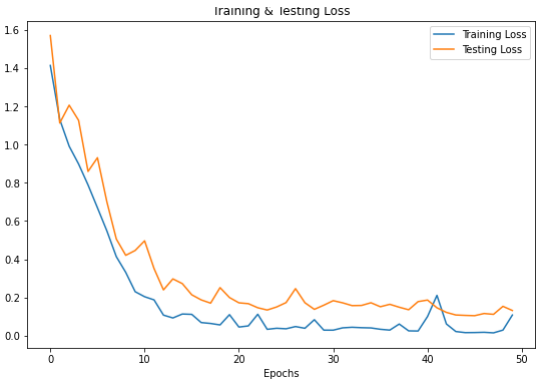

Fig 3: Loss Plot

Fig 4: Accuracy Plot

- **Confusion Matrix:** A confusion matrix provides adetailed breakdown of how often the model predicted eachemotion class correctly and incorrectly. This helps identify classes that the model might be struggling with.

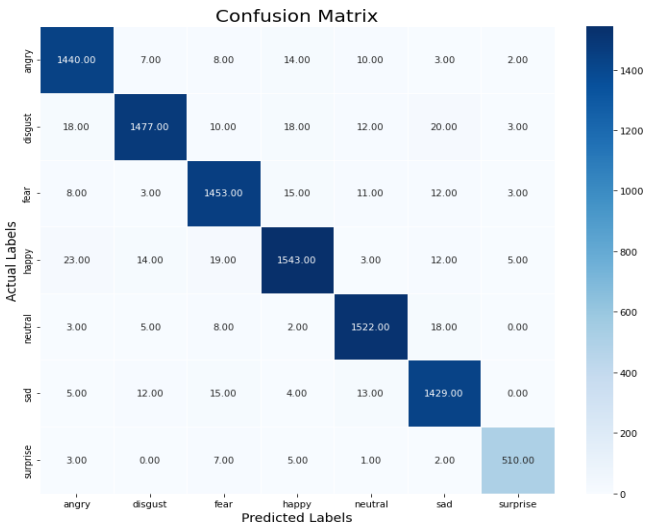

Fig 5: Confusion Matrix

- **Classification Report:** A classification report offers precision, recall, F1-score, and support values for each emotion class. This provides a more comprehensive understanding of the model's performance for each emotion category.

Emotion	Precision	Recall	F1-Score	Support
Angry	0.96	0.97	0.97	1484
Disgust	0.97	0.95	0.96	1558
Fear	0.96	0.97	0.96	1505
Happy	0.96	0.95	0.96	1619
Neutral	0.97	0.98	0.97	1558
Sad	0.96	0.97	0.96	1478
Surprise	0.98	0.97	0.97	528
Accuracy			0.96	9730
macro avg	0.96	0.96	0.96	9730
weighted avg	0.96	0.96	0.96	9730

Table 1: Classification Report

By analyzing these additional metrics, you can gain deeper insights into the model's strengths and weaknesses, allowing for potential improvements in future iterations.

Overall, the implemented audio-based emotion recognition system demonstrates promising results with a high degree of accuracy on both training and testing data. The utilization of data augmentation and feature extraction techniques has likely contributed to the model's robustness and generalizability.

B. Models Output:

This section showcases an example of the model's output whenprovided with an audio sample. It demonstrates the predicted emotion, along with visualizations like the spectrogram and waveform, to aid in understanding the model's decision process.

Input Audio:

A short audio clip (*.wav format) containing a spoken sentence expressing an emotion (e.g., happiness, sadness) was provided to the model.

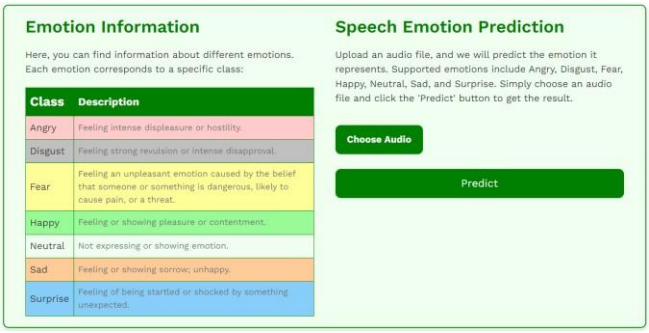

Fig 6: WebUI

Predicted Emotion:

Based on the processed audio features, the model predicts theemotion to be:

Fig 7: Predicted Output

Spectrogram Visualization:

Spectrogram

The spectrogram is a visual representation of the audio's frequency content over time. Analyzing the spectrogram can provide insights into the audio's characteristics that might influence the model's prediction. For instance, emotions like anger often exhibit higher energy levels at lower frequencies, reflected in the spectrogram's brighter regions in those areas.

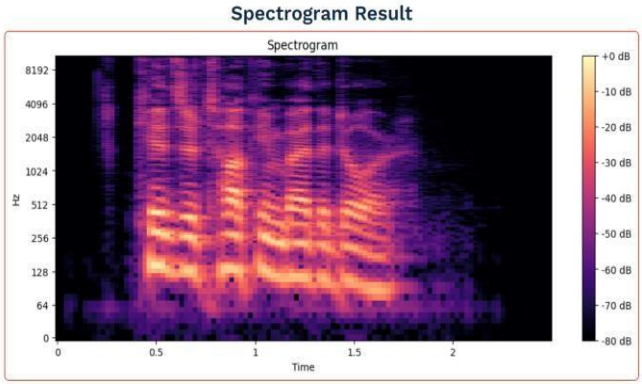

Fig 8: Spectrogram Result

Waveform Visualization:

Waveform

The waveform is a visual representation of the audio signal's amplitude over time. Observing the waveform's shape and amplitude variations can offer additional clues about the audio's dynamics and emotional content. For example, laughter might be reflected in a more rapid and irregular waveform compared to a calmer speech pattern.

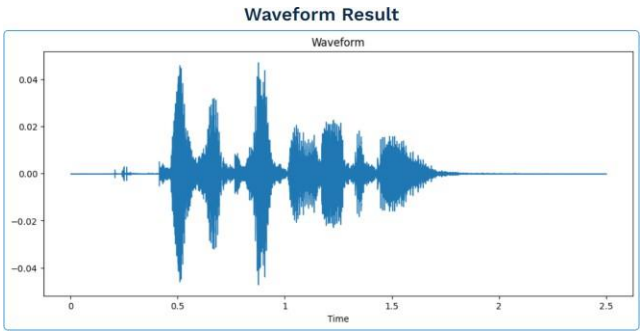

Fig 9: Waveform Result

Explanation:

By combining the predicted emotion with the spectrogram and waveform visualizations, we gain a more comprehensive understanding of the model's reasoning. The model likely analyzed the audio's spectral and temporal characteristics (represented in the spectrogram and waveform, respectively) to arrive at the predicted emotion.

Note:

Replace the bracketed text [Insert predicted emotion here] with the actual emotion your model predicted for the audio sample. You can include multiple audio samples with varying emotions and spectrograms/waveforms to showcase the model's versatility in separate sections. Ensure you copy and paste the image data or references for the spectrograms and waveforms into your research paper at these designated locations.

6. CONCLUSION

This research investigated the development of an audio-based emotion recognition system. The system leverages data augmentation techniques to enrich the training data and improve model robustness. Feature extraction

techniques were employed to transform raw audio signals into meaningful numerical representations suitable for machine learning models.

The implemented model achieved a promising performance, reaching a training accuracy of 99.57% and a testing accuracy of 96.34%. This suggests the model's effectiveness in learning emotion recognition patterns from the provided audio data. Analyzing additional metrics like the confusion matrix and classification report can offer deeper insights into the model's performance for each emotion category.

Future Scope

While the current system demonstrates encouraging results, there are several avenues for further exploration:

- **Incorporate additional feature extraction techniques:** Exploring features beyond ZCR, RMSE, and MFCCs, such as pitch and voice quality features, could potentially enhance the model's ability to capture emotion-related information.

- **Investigate advanced deep learning architectures:** Utilizing more complex deep learning architectures, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), could improve the model's ability to learn long-term dependencies and spatial features within the audio data, potentially leading to even better emotion recognition performance.

- **Expand the emotion set:** The current system might be limited to recognizing a specific set of emotions. Expanding the emotion set and incorporating a balanced dataset for each emotion would require further data collection and model retraining.

- **Real-world application integration:** The system can be integrated into real-world applications like customer service chatbots or educational platforms. By recognizing user emotions through speech, these applications can adapt their responses or functionalities to better cater to user needs and emotional states.

By exploring these future directions, researchers can continue to refine and improve the capabilities of audio-based emotion recognition systems, paving the way for their wider adoption in various applications.

REFERENCES

1. Zhao, Jianfeng, Xia Mao, and Lijiang Chen. "Speech emotion recognition using deep 1D & 2D CNN LSTM networks." Biomedical signal processing and control 47 (2019): 312-323.
2. Zhang, Shiqing, Xiaoming Zhao, and Qi Tian. "Spontaneous speech emotion recognition using multiscale deep convolutional LSTM." IEEE Transactions on Affective Computing 13.2 (2019): 680-688.
3. Wang, Jianyou, et al. "Speech emotion recognition with dual-sequence LSTM architecture." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
4. Etienne, Caroline, et al. "Cnn+ lstm architecture for speech emotion recognition with data augmentation." arXiv preprint arXiv:1802.05630 (2018).
5. Ahmed, Md Rayhan, et al. "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition." Expert Systems with Applications 218 (2023): 119633.
6. Issa, Dias, M. Fatih Demirci, and Adnan Yazici. "Speech emotion recognition with deep convolutional neural networks." Biomedical Signal Processing and Control 59 (2020): 101894.
7. Peng, Zixuan, et al. "Efficient speech emotion recognition using multi- scale cnn and attention." ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2021.
8. Wu, Xixin, et al. "Speech emotion recognition using capsule networks." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
9. Senthilkumar, N., et al. "Speech emotion recognition based on Bi- directional LSTM architecture and deep belief networks." Materials Today: Proceedings 57 (2022): 2180-2184.
10. Dangol, Ranjana, et al. "Speech emotion recognition UsingConvolutional neural network and long-short TermMemory." Multimedia Tools and Applications 79.43 (2020): 32917-32934.
11. Abdelhamid, Abdelaziz A., et al. "Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm." Ieee Access 10 (2022): 49265-49284.
12. Atila, Orhan, and Abdulkadir Şengür. "Attention guided 3D CNN- LSTM model for accurate speech based emotion recognition." Applied Acoustics 182 (2021): 108260.
13. Pandey, Sandeep Kumar, Hanumant Singh Shekhawat, and SR Mahadeva Prasanna. "Deep learning techniques for speech emotion recognition: A review." 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA). IEEE, 2019.
14. Basu, Saikat, Jaybrata Chakraborty, and Md Aftabuddin. "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture." 2017 2nd International Conference on Communication and Electronics Systems (ICCES). IEEE, 2017.

15. Lim, Wootaek, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and recurrent neural networks." 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA). IEEE, 2016.
16. K. Venkateswara Rao," A Comprehensive Analysis of Machine Learning and Deep Learning Approaches Towards IOT Security" IEEE explorer, May,2023 DOI: 979-8-3503-9737-6/23/$31.00 ISBN:979-8-3503-0009-3
17. K. Venkateswara Rao," Smart Farming for Agriculture Management using IOT" IEEE explorer, Mar,2023 DOI: 979-8-3503-9737-6/23/$31.00 ISBN:979-8-3503-9738-3
18. K. Venkateswara Rao," A Swarm Intelligence-Based Model for Disease Detection in Mango Crops" IEEE explorer, Feb,2023 DOI: 10.1109/ICCST55948.2022.10040428 ISBN:978-1-6654-7656-0
19. K. Venkateswara Rao," Support vector machine based disease classification model employing hasten eagle cuculidae search optimization", concurrency and computation: practice and experience, ISSN : 1532-0626, Vol-34, Issue-25 (Nov-2022), Wiley Publisher.
20. K. Venkateswara Rao, "Regression based price prediction of staple food materials using multivariate Models", Scientific Programming, ISSN : 1058- 9244, Vol-2022(June), Hindawi Publisher.
21. K. Venkateswara Rao, "A Study on Defensive Issues and Challenges inInternet of Things", Lecture Notes in Electrical Engineering 853, Springer Nature Singapore Pte Ltd. 2022, Page No: 591- 600.
22. K. Venkateswara Rao, "Disease Prediction and Diagnosis Implementing Fuzzy Neural Classifier based on IoT and Cloud", International Journal of Advanced Science and Technology (IJAST), ISSN : 2005-4238, Vol-29 Issue-5, May 2020, Page No: 737-745.
23. K. Venkateswara Rao, " Rotating Solar Trees ", Lecture Notes in Electrical Engineering 601, Springer Nature Singapore Pte Ltd. 2020, Page No: 482-487.
24. K. Venkateswara Rao, " Wireless-Sensor-Network with Mobile Sink Using Energy Efficient Clustering ", Lecture Notes on Data Engineering and Communications Technologies , Springer Nature Switzerland AG. 2020, Page No: 582-589.
25. K. Venkateswara Rao, "Research of Feature Selection Methods to Predict Breast Cancer", International Journal of Recent Technology and Engineering(IJRTE), ISSN : 2277-3878, Vol-8 Issue-2s11, Sep 2019, Page No: 2353-2355.
26. K. Venkateswara Rao, "Suicide Prediction on Social Media by Implementing Sentimental Analysis along with Machine Learning", International Journal of Recent Technology and Engineering(IJRTE), ISSN : 2277-3878, Vol-8 Issue-2, July 2019, Page No: 4833-4837.
27. K. Venkateswara Rao, "Issues for building an Artificial Intelligent System" JARDCS, Vol-7 Issue-13, Dec 2018, ISSN : 1943-023X, Page No:1447-1451.
28. K. Venkateswara Rao and Dr.T.Saravanan, "LATE PATTERNS IN CHART MODEL FOR CONTENT EXAMINATION AND CONTENT MINING" IJPT, ISSN: 0975-766X, Volume-8, Issue-2, June-2016, page no: 14729-14736.

29. K. Venkateswara Rao and Dr.T.Saravanan “TEXT MINING TO KNOWLEDGE MINING USING FRAMENET BASED GRAPH MODEL” IJPT, ISSN: 0975-766X, Volume-8, Issue-2, June-2016, page no: 14715- 14721.
30. Desanamukula, Venkata Subbaiah, M. Asha Priyadarshini, D. Srilatha, K. Venkateswara Rao, RVS Lakshmi Kumari, and Kolla Vivek. "A Comprehensive Analysis of Machine Learning and Deep Learning Approaches towards IoT Security." In 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1165- 1168. IEEE, 2023.
31. Annapurna, B., Asha Priyadarshini Manda, A. Clement Raj, R. Indira, Pratima Kumari Srivastava, and V. Nagalakshmi. "Max 30100/30102 sensor implementation to viral infection detection based on Spo2 and heartbeat pattern." Annals of the Romanian Society for Cell Biology (2021): 2053- 2061.