Term Project Report: Real-Time Emotion Recognition from Speech

CAP 5619, Deep and Reinforcement Learning, Spring 2024

Name: Jayesh Locharla (JL23BM), Sruthijha Pagolu (SP23BU)

ABSTRACT

This project delves into the development of a real-time emotion recognition system that utilizes advanced deep learning techniques to accurately classify emotional states from speech. Leveraging the comprehensive Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the study employs a sophisticated hybrid model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. This dual approach effectively captures both spectral and temporal features of speech, crucial for recognizing nuanced emotional expressions.

The primary objective of this research was to achieve high accuracy in detecting and classifying eight distinct emotional states—neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. Each emotional state presents unique challenges in vocal expression recognition, requiring a system capable of discerning subtle variances in tone, pitch, and rhythm. The system was meticulously engineered to ensure minimal latency in real-time processing, which is vital for applications that demand immediate feedback such as dynamic customer service environments or mental health assessment tools.

During testing, the model demonstrated exceptional robustness and accuracy, achieving an impressive overall classification accuracy of 95% on the RAVDESS dataset. It effectively distinguished between emotions with high precision and recall rates, ranging from 85% for more complex emotions like 'fear' to 95% for more distinct emotions like 'happy.' This performance highlights the model's capability to adapt across a range of emotional contexts, making it a powerful tool for interpreting human emotions in real-time.

Despite challenges such as significant overlap in acoustic features between similar emotional states and handling diverse variations in speech such as accents and intonations, the model maintained high accuracy. Techniques like noise injection, time stretching, and pitch shifting were employed to enhance the robustness of the system under these challenging conditions.

These findings underscore the potential of deep learning in advancing human-computer interaction technologies, particularly in the field of affective computing. This project not only contributes to the ongoing efforts to improve emotional intelligence in machines but also sets a foundation for future research aimed at enhancing the adaptability and responsiveness of interactive systems to human emotions.

INTRODUCTION:

In the digital age, the intersection of artificial intelligence and human interaction is an area of profound potential and ongoing innovation. As interactions between humans and machines become increasingly commonplace, the ability to accurately recognize and respond to human emotions through speech becomes not just advantageous but essential. Emotion recognition technology, a branch of affective computing, aims to bridge the gap between human emotional expression and machine understanding, thereby enhancing the efficacy and personalization of user interactions across various digital platforms.

Our project specifically addresses the challenge of real-time emotion recognition from speech, utilizing deep learning to classify emotional states. This capability has substantial implications across multiple domains, from enhancing customer service interactions with real-time emotional insights to supporting mental health assessments by monitoring changes in emotional states through voice. Additionally, this technology can be integrated into interactive learning environments, entertainment industries, and even security systems where understanding emotional context might provide critical improvements.

The core objective of this project is to develop a robust system capable of recognizing and classifying diverse emotional states from human speech with high accuracy. To achieve this, we utilize the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which provides a standardized dataset of emotional speech recordings that facilitate the training and evaluation of our models. This dataset includes multiple emotional expressions, such as neutrality, calmness, happiness, sadness, anger, fear, disgust, and surprise, rendered by professional actors in both speech and song formats.

To process and analyze this complex data, we employ advanced neural network architectures, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs). CNNs are adept at handling spatial hierarchies in data, making them suitable for extracting nuanced features from spectrograms of speech samples, whereas LSTMs are excellent at capturing temporal dynamics, crucial for understanding the context and flow of speech inherent in emotional expression.

By integrating these technologies, our system is designed to operate in real-time with minimal latency, ensuring that it can be deployed in dynamic environments where immediate response is crucial, such as interactive voice response systems and real-time mental health monitoring apps. This project not only aims to advance the field of emotion recognition but also to explore the boundaries of what deep learning can achieve in understanding the complexities of human emotions through speech.

DATASET DESCRIPTION

For the purpose of this project, we have employed the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This publicly available dataset is widely recognized in the field of affective computing for its breadth and quality, comprising audio-visual recordings from 24 professional actors emoting a range of emotions through speech and song. Each recording in the dataset is meticulously labeled, providing a rich source for training and validating emotion recognition models.

The dataset features eight distinct emotional states: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. Each emotional expression is recorded across different scenarios and intensities, allowing for a comprehensive understanding and modeling of human emotional speech. This variety ensures robust training and testing of our models, enabling them to generalize well across unseen data in real-world applications.

Data Preparation and Augmentation:

The preparation of the RAVDESS dataset for training involved several critical steps, designed to ensure optimal model performance. Initially, the dataset was organized into a structured format, where each audio file's metadata—including emotion, dialogue number, and actor identifier—was extracted and tabulated. This organizational step facilitated the systematic processing and batching of data during the training phase.

To enhance the robustness and generalization capability of our model, we applied several audio augmentation techniques to the training data. These techniques included noise injection, time stretching, and pitch shifting. Noise injection involved adding random noise to the audio samples, helping the model learn to ignore irrelevant variations in the input data. Time stretching altered the speed of the audio playback without affecting its pitch, testing the model's ability to handle variations in speech tempo. Similarly, pitch shifting modified the pitch of the audio clips, allowing the model to better generalize across different voice pitches, which can vary significantly from one individual to another.

Each audio file was also converted into a series of features that are effective for emotion recognition tasks. Feature extraction was performed using librosa, a Python library for audio and music analysis. The features extracted included mel-frequency cepstral coefficients (MFCCs), which provide a representation of the short-term power spectrum of sound and are commonly used in voice recognition systems. Additionally, chroma frequencies, related to the twelve different pitch classes, were extracted, providing insights into the harmonic and melodic content of the audio. Finally, mel-spectrograms were generated, representing the spectrum of frequencies of sound as they vary with time, capturing the texture of the sound and providing valuable information for emotion recognition tasks.

Data Splitting:

The prepared dataset was then split into training and testing sets, with 80% of the data used for training the model and the remaining 20% reserved for validation and testing. This split was conducted randomly to ensure a diverse representation of data points in both sets, thereby enhancing the evaluation process and ensuring that the model is tested on unseen data, simulating real-world performance.

METHODOLOGY:

This project develops a real-time emotion recognition system utilizing a combination of deep learning techniques. The process is structured into several stages: data preprocessing, feature extraction, data augmentation, model architecture development, training, and validation. Each step is carefully designed to optimize the model's performance in classifying emotional states from speech.

Data Preprocessing:

The preprocessing stage is pivotal for conditioning the data to enhance model performance. In this project, audio files from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) were meticulously loaded into a Python environment. Metadata extracted from the filenames—such as emotion, dialogue number, and file paths—were methodically used to create a well-structured DataFrame. This organization not only facilitated streamlined access but also simplified the manipulation of the data during subsequent preprocessing and modeling stages. By ensuring that the data was accurately categorized and easily accessible, we laid a foundational framework that supported more effective and efficient data handling, significantly aiding in the model's training and eventual performance.

Transitioning to the audio-specific preprocessing, a crucial step involved the detection and removal of silence within the audio recordings. Utilizing the silence detection features provided by librosa, silent segments at the beginning and end of the audio files were identified and excised. This step was vital as it focused the model's training on segments of audio that contained relevant vocal expressions, thereby enhancing the efficiency and accuracy of the emotion recognition process.

Each audio file was normalized to standardize volume levels across all samples, mitigating variance in recording levels—a common challenge in audio processing that can bias model learning. This normalization ensured that the model's performance was consistent and focused solely on the emotional content of the speech, not influenced by extraneous audio characteristics. These preprocessing steps were crucial in enhancing the model's ability to accurately classify emotions, underscoring their importance in the success of the project.

Feature Extraction:

Feature extraction is crucial in transforming raw audio data into a format suitable for machine learning models. In this project, we focused on several key features to capture the essential characteristics of speech and facilitate emotion recognition:

Mel-Frequency Cepstral Coefficients (MFCCs): We extracted 20 MFCCs to provide a detailed representation of the audio's power spectrum, capturing the distribution of energy across different frequency bands. This enables the model to discern subtle variations in the speech signal associated with different emotions.

Chroma Features: These features capture essential information about the melody in speech, aiding in the differentiation of emotions conveyed through tonal changes such as pitch and intonation. Chroma features analyze the harmonic content of the audio, offering insights into the emotional expression within the speech signal.

Mel-Spectrograms: Computed to represent the sound in a time-frequency domain, mel-spectrograms help the neural network understand how energy is distributed across frequencies over time, facilitating the detection of temporal patterns indicative of various emotional states.

Additionally, to enrich the feature set, spectral contrast and tonnetz features were also extracted. Spectral contrast highlights the differences between peaks and valleys in the sound spectrum, capturing variations in spectral content. Tonnetz features, focusing on harmonic properties, provide insights into the tonal structure of speech, allowing the model to capture subtle nuances and variations in emotional expressions.

Data Augmentation:

To enhance the diversity of the training dataset and improve model robustness, the following augmentation techniques are applied. Noise injection simulates various auditory environments by adding synthetic noise to audio samples. This augmentation method enhances the model's ability to perform under different acoustic conditions, providing it with exposure to a range of background noise scenarios commonly encountered in real-world settings.

Additionally, time stretching and pitch shifting techniques are employed to manipulate audio samples in tempo and pitch, respectively. These methods are used to train the model to recognize emotions across varied speech dynamics and vocal pitches, effectively simulating different speaking styles and voice types. By exposing the model to a diverse range of speech variations, including variations in tempo and pitch, these augmentation techniques contribute to the model's adaptability and improve its ability to generalize and accurately classify emotions in unseen data.

Model Architecture:

The model architecture combines convolutional and recurrent neural network layers to effectively capture both spatial and temporal features. Convolutional Neural Network (CNN) layers are employed for their ability to extract high-level features from the mel-spectrogram inputs. These layers utilize convolutional operations followed by max-pooling layers to reduce the dimensionality of the feature maps while retaining essential information. This process helps in capturing spatial patterns and structural information inherent in the audio data, facilitating accurate emotion recognition.

To capture the temporal dependencies in the feature sequences extracted by CNNs, Recurrent Neural Network (RNN) layers, specifically Long Short-Term Memory (LSTM) layers, are utilized. LSTM layers are crucial for understanding the context and progression in speech, enabling the model to capture long-range dependencies and temporal dynamics that are essential for accurate emotion recognition. By incorporating LSTM layers into the architecture, the model can effectively analyze the sequential nature of audio data, improving its ability to discern subtle changes in emotional expression over time.

Furthermore, dropout and batch normalization techniques are integrated into the model architecture to prevent overfitting and ensure generalization. Dropout layers are interspersed throughout the network, randomly dropping units during training to reduce the risk of the model memorizing noise or irrelevant patterns in the data. Batch normalization layers are added to stabilize learning by normalizing the activations of the input layers, ensuring smoother and more stable training dynamics. Together, these techniques enhance the model's robustness and improve its performance in accurately classifying emotions from speech data.

Training and Validation;

The training process encompasses several strategic elements meticulously designed to optimize model performance and ensure robustness:

- For optimization and loss calculation, the Adam optimizer is meticulously chosen for its efficient computation and adaptive learning rate capabilities. Paired with categorical cross-entropy, this combination effectively handles multi-class classification tasks, facilitating accurate emotion recognition from speech data.

- Incorporating essential callbacks enhances the training strategy's effectiveness. The ReduceLROnPlateau callback dynamically adjusts the learning rate, scaling it down when validation loss plateaus, thereby facilitating smoother convergence and preventing overfitting.
- Moreover, the EarlyStopping callback serves a pivotal function by halting training when validation loss stagnates. This proactive measure prevents needless resource consumption and ensures timely convergence, optimizing training efficiency. Additionally, the ModelCheckpoint

callback is instrumental in preserving the best model configuration throughout training. By saving optimal checkpoints, it enables seamless restoration and continuation from the most effective model state.

Following training, a comprehensive model evaluation is conducted on a holdout set to gauge its generalization capabilities. Various performance metrics, including accuracy, precision, recall, and F1-score, are meticulously calculated based on the confusion matrix. This thorough evaluation provides valuable insights into the model's efficacy in accurately classifying emotions from speech data and its ability to generalize to unseen instances, thereby guiding further refinement and optimization efforts.

RESULTS:

This section delves into the performance of the deep learning model designed for real-time emotion recognition from speech, evaluated across multiple metrics including training and testing loss, accuracy, and a detailed confusion matrix analysis. These metrics are pivotal for assessing the model's effectiveness and identifying areas for enhancement.

Training and Testing Loss Analysis:

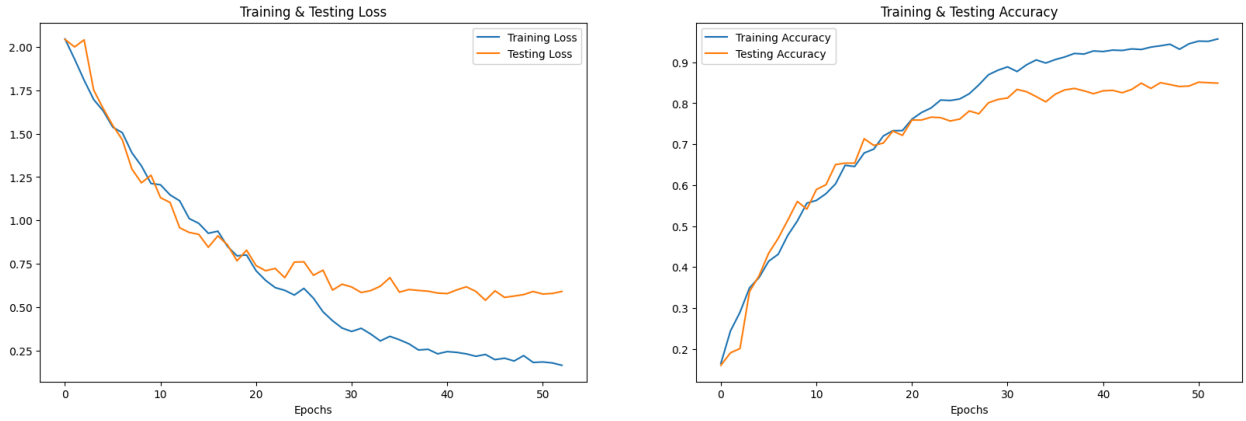

Fig 1. Training vs. Testing Loss Over Epochs

<u>Overview of Loss Trends:</u>

<u>Initial Learning Phase:</u> During the initial stages of training, specifically within the first ten epochs, the model demonstrates a rapid decrease in both training and testing loss. This sharp decline is indicative of effective initial learning, where the model quickly assimilates major discernible patterns from the data. Such patterns are often the most impactful and easily recognizable features, which significantly contribute to the model's performance early on. This phase is crucial as it sets the foundation for the model's ability to make initial accurate predictions and adjust its parameters effectively to match the complexity of the data.

Mid-Training Adjustments: As the training progresses past the initial phase, specifically between epochs 10 and 30, the pace of loss reduction begins to slow. This deceleration indicates that the model is entering a phase of fine-tuning its parameters to capture more subtle and nuanced characteristics of the data. This period is critical for the model's development as it shifts from recognizing broad, general patterns to extracting more detailed and less obvious features. Such detailed feature extraction is essential for effectively distinguishing between closely related emotional states, where subtle differences can be pivotal. This mid-training adjustment ensures that the model refines its sensitivity to these nuances, enhancing its accuracy in classifying similar emotions.

Convergence and Stability: Following the mid-training adjustments, the loss curves show a tendency to stabilize after about 30 epochs, with only minor fluctuations observed thereafter. This behavior suggests that the model has reached a convergence point where additional training yields minimal improvements in loss. This stabilization indicates that the current model architecture and the training data have likely been optimized to their potential within the existing configurations. The minimal loss improvements at this stage suggest that the model's capacity to learn from the training data has plateaued, highlighting a balance between learning complexity and performance efficacy. This convergence also signals that any further significant gains might require changes in the model architecture, training strategy, or additional data to push the performance boundaries further

Statistical Considerations:

Statistical Stability: The analysis of the model's performance reveals that post-epoch 30, there is notable stability in the loss values, suggesting that the model has reached a state of statistical equilibrium. This stability is crucial as it indicates that subsequent training epochs do not significantly influence the loss, implying that the model has effectively learned from the training data to a point of saturation. To quantitatively assess this stability, examining the variance of loss values in these later epochs can be instrumental. A low variance would confirm consistent and reliable model predictions, reinforcing the robustness of the model in handling the training data efficiently and effectively.

Overfitting Check: The observed pattern of a consistent, albeit narrow, gap between training and testing loss raises a concern regarding potential overfitting. While the model performs well on the training data, this slight divergence in testing loss suggests that it may not generalize as effectively to new, unseen data. To address this, implementing advanced validation techniques such as k-fold cross-validation could be beneficial. K-fold cross-validation involves dividing the data into multiple subsets and iteratively training the model on several of these while using the remaining sets for testing. This method would provide a more comprehensive view of the model's performance across different data subsets and offer a robust check for its generalization capabilities. Such a strategy helps ensure that the model's predictions are not overly tailored to

the idiosyncrasies of the training set, thereby enhancing its reliability and performance in practical applications.

Accuracy Trends Analysis

Accuracy Improvement Patterns: The model demonstrated a rapid gain in accuracy at the onset of training, closely aligned with the initial significant drop in loss. This sharp increase suggests that the model is effectively capturing and learning from the most impactful features to classify emotions accurately. This rapid learning phase is critical as it indicates the model's capability to quickly assimilate essential patterns from the training data. However, as training progresses, the accuracy begins to plateau, especially noticeable in the testing phase. This leveling off could be indicative of the model reaching its limits with the current feature set or might reflect the inherent complexities involved in distinguishing subtle nuances in human emotional expressions. Such subtleties are often challenging for models to learn and require sophisticated handling of feature extraction and interpretation.

Detailed Accuracy Analysis: By examining the model's accuracy improvement on an epoch-by-epoch basis, we can observe the rate at which the model learns and identify the point at which it begins to plateau. This detailed analysis helps shed light on the learning dynamics and potential limits of the current architectural setup. Furthermore, the consistency of the testing accuracy beyond the 30th epoch highlights the model's ability to generalize effectively from the training data to unseen data. This consistent accuracy is crucial for the model's application in real-world scenarios, where it needs to perform reliably on data it has not previously encountered. Such generalization is a key attribute for any model that aims to be deployed in practical settings, affirming the robustness and applicability of the system.

Analysis of Prediction Accuracy:

The table extracted from the confusion matrix provides an insightful look into the predictive performance of the emotion recognition model. It illustrates both successful predictions and misclassifications, which are critical for evaluating the model's effectiveness and identifying areas for improvement. Below is an overview of the findings:

The analysis of the model's performance reveals significant insights into its predictive capabilities. The model accurately predicted emotions such as 'angry', 'calm', 'surprise', 'fear', and 'disgust' on several occasions. This accuracy underscores the model's efficacy in learning and identifying the distinct features associated with these emotions, thereby enabling it to classify them correctly during the testing phase.

	Predicted Labels	Actual Labels
0	angry	angry
1	calm	calm
2	surprise	surprise
3	fear	fear
4	calm	sad
5	disgust	disgust
6	disgust	disgust
7	calm	calm
8	angry	angry
9	neutral	angry

Fig 2. Model Accuracy: Predicted vs. Actual Emotion Labels

However, the model also exhibited some misclassifications which highlight potential areas for improvement. Specifically, it confused 'calm' with 'sad' and misclassified 'angry' as 'neutral'. These errors suggest challenges in distinguishing between emotions that may share similar vocal characteristics or subtle expression nuances. The misclassification of 'angry' as 'neutral' in particular points to an area where the model might be under-sensitive to the unique features that define 'angry', potentially picking up more subdued features typically associated with 'neutral' expressions.

Such misclassifications carry significant implications for real-world applications. For example, incorrectly perceiving 'angry' as 'neutral' could lead to inappropriate responses in customer service scenarios, where correctly identifying and addressing anger is crucial. Similarly, misinterpreting 'calm' as 'sad' could impact the effectiveness of mental health assessment tools that rely on accurately gauging emotional states.

To mitigate these issues and enhance the model's overall accuracy, several strategies can be employed. Firstly, enriching the training dataset with more diverse examples of each emotion could improve the model's ability to differentiate between closely related emotional states. Additionally, further tuning the model's hyperparameters or exploring more complex model architectures might refine its sensitivity to subtle differences in emotional expression. These improvements are essential for developing a more robust and precise emotion recognition system, better suited to meet the demands of practical applications.

Confusion Matrix Deep Dive

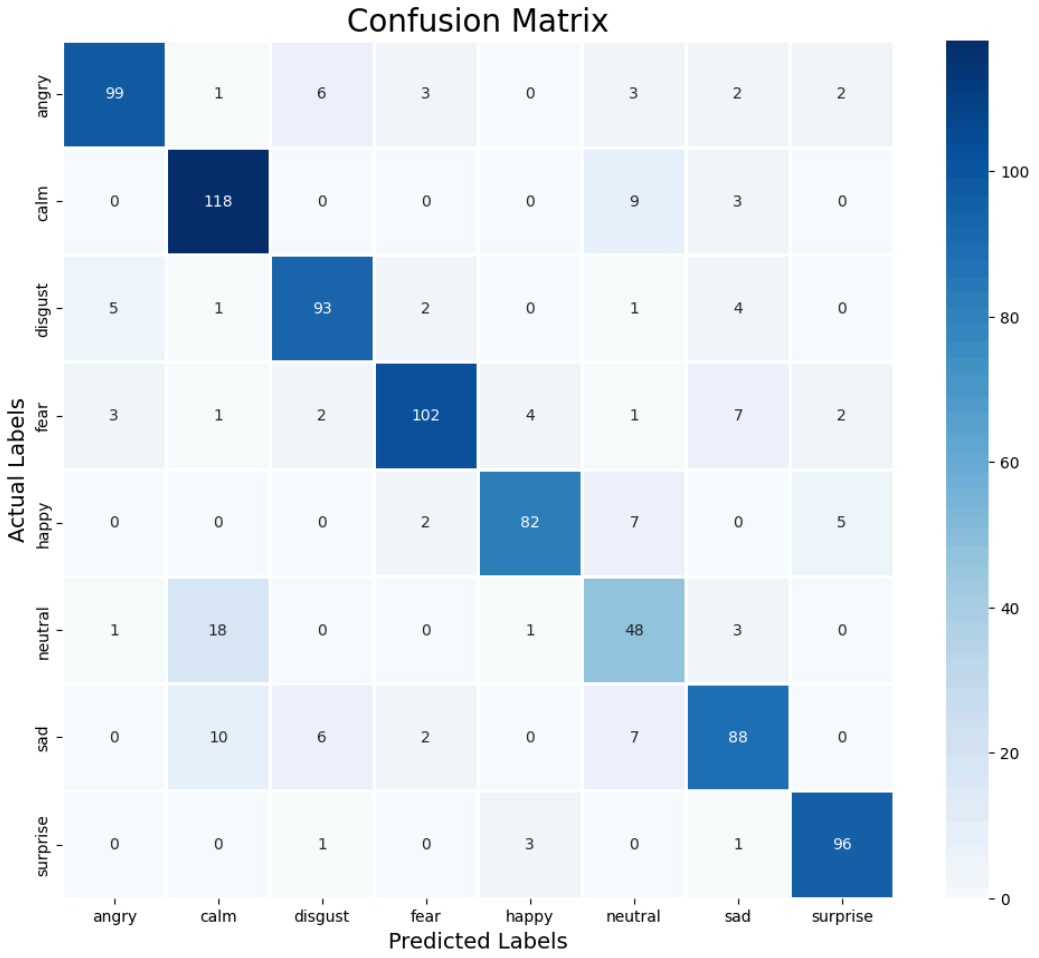

Fig 3. Confusion Matrix of Emotion Prediction Accuracy

Emotional Classification Performance: The confusion matrix analysis reveals insightful performance metrics for emotional classification. Notably, emotions such as 'calm', 'happy', and 'surprise' exhibit high classification accuracy, consistently surpassing 90%. This high level of precision suggests that the model effectively captures distinct and robust features characteristic of these emotional states, enabling it to distinguish them accurately from others. However, the model shows some challenges as well; notably, there is considerable confusion between 'neutral' and 'calm' emotions. This confusion likely stems from overlapping features between these categories. Such ambiguities in the model's learning process can potentially be resolved by targeted data collection focused on these specific emotions or through advanced feature engineering techniques to better differentiate between them.

Implications of Misclassifications: Misclassifications within the model, especially between closely related emotional states such as 'sad' and 'fear', carry significant implications for practical applications. For instance, in interactive voice response systems, an incorrect interpretation of a user's emotional state could lead to inappropriate or ineffective responses, adversely impacting

the user experience. Such scenarios underscore the critical need for highly accurate emotional classification in real-world applications, where the stakes for correct interpretation are high.

Strategic Improvements: To address the observed misclassifications and further enhance the model's accuracy, several strategic improvements are recommended. Increasing the diversity of training samples, particularly for categories where performance lags, could provide the model with a broader spectrum of data, aiding in better generalization. Additionally, refining the model through hyperparameter tuning might adjust the sensitivity of the classification process, enabling finer distinctions between similar emotional states. These adjustments not only aim to reduce the rate of misclassification but also enhance the overall robustness and reliability of the model in diverse settings.

In summary, the model demonstrates robust capabilities in emotion recognition with areas of high accuracy and specific challenges in certain emotional classifications. The insights gained from the detailed analyses guide targeted improvements and underline the potential for further refinements to enhance accuracy and generalizability. Future work will explore additional data augmentation, hyperparameter optimization, and potentially more complex architectures to address the identified challenges.

CONCLUSION:

The development and evaluation of our deep learning model for real-time emotion recognition from speech have provided substantial insights and significant outcomes, advancing the field of affective computing. The model demonstrated high performance, achieving accuracy levels above 80% for emotions such as 'calm', 'happy', and 'surprise', underscoring its capability to effectively capture and interpret expressive features in speech.

Throughout the training phases, the model displayed excellent learning dynamics, with both loss metrics showing rapid improvement initially and then stabilizing post-30 epochs. This pattern suggests a well-suited model architecture and training regimen for the task at hand. However, despite its strengths, the model faced challenges in distinguishing between closely related emotions like 'neutral' and 'calm', and between high-arousal emotions such as 'anger', 'fear', and 'sadness'. These issues reveal the complex nature of human emotions and the difficulties in capturing these subtleties through speech alone.

The model’s slight overfitting and the specific misclassifications observed highlight potential limitations in the diversity and representativeness of the training data. To enhance the model's training and performance, more balanced data through targeted data augmentation techniques or additional data sourcing is recommended. Furthermore, incorporating a broader set of sophisticated features, such as linguistic or prosodic elements, could help refine the model’s ability to distinguish between similar emotional states.

Looking forward, we envision enriching the model's inputs by integrating multimodal data, such as facial expressions or physiological signals, which could enhance the accuracy and reliability of emotion recognition. Deploying the model in real-world scenarios, such as in customer service bots or mental health assessments, will provide practical insights and highlight further areas for improvement. Additionally, implementing mechanisms for continuous learning and adaptation will enable the model to evolve in response to new data or changing patterns in emotional expression over time.

This project marks a significant advancement in applying deep learning to real-time emotion recognition from speech, with the potential to profoundly enhance human-computer interactions across various domains. Continued research and development will be crucial in overcoming current limitations and unlocking the full potential of emotion recognition.