Multimodal Emotion Recognition Using Speech and Text

1. Introduction

Emotion recognition is an important area of research in human–computer interaction, enabling machines to understand human emotional states through various modalities such as speech, text, and facial expressions. Traditional emotion recognition systems rely on a single modality, which may not capture complete emotional information. Multimodal emotion recognition combines multiple sources of information to improve prediction accuracy and robustness.

The objective of this project is to develop an emotion recognition system using three approaches:

1. Speech-only model
2. Text-only model
3. Multimodal fusion model (Speech + Text)

The Toronto Emotional Speech Set (TESS) dataset was used for experimentation.

---

2. Dataset Description

The Toronto Emotional Speech Set (TESS) dataset consists of speech recordings spoken by female actors expressing different emotions. The dataset contains several emotion categories including:

- Angry
- Happy
- Sad
- Fear
- Disgust
- Surprise
- Neutral

Each audio file contains emotional speech information. Since the dataset filenames include emotion labels, textual data was generated from filenames for the text-based model.

The dataset provides a controlled environment for evaluating speech-based emotion recognition systems.

3. System Architecture

The system consists of three main pipelines:

1. Speech Pipeline
2. Text Pipeline
3. Multimodal Fusion Pipeline

Each pipeline includes the following functional blocks:

- Preprocessing
- Feature Extraction
- Temporal / Context Modeling
- Fusion (for multimodal)
- Classification

The fusion pipeline combines features extracted from both speech and text modalities to improve performance.

4. Methodology

4.1 Preprocessing

Speech Preprocessing

Speech signals were loaded using audio processing libraries. The audio was trimmed and normalized to ensure consistent duration across samples. Sampling rate standardization was applied to maintain uniformity.

Text Preprocessing

Text data was generated using emotion labels from filenames. The sentences were tokenized and converted into numerical sequences using a tokenizer. Padding was applied to ensure equal sequence lengths.

4.2 Feature Extraction

Speech Features

Acoustic features such as Mel Frequency Cepstral Coefficients (MFCC) and spectral features were extracted to capture emotional characteristics from audio signals. These features represent both temporal and spectral information.

Text Features

Text sequences were converted into numerical tokens using word indexing. Padding was applied to ensure uniform input dimensions for neural network models.

4.3 Temporal and Contextual Modelling

For speech modelling, neural network architectures were used to learn emotional patterns from acoustic features over time.

For text modelling, a Bidirectional Long Short-Term Memory (BiLSTM) network was used. The Bidirectional LSTM captures contextual dependencies in both forward and backward directions, improving emotion understanding from text sequences.

4.4 Fusion Strategy

The multimodal fusion model combines speech and text features by concatenating their feature representations. This unified representation provides complementary emotional information from both modalities.

4.5 Classification

Dense neural network layers were used to classify emotions based on extracted representations. The output layer used a softmax activation function to predict emotion categories. The models were trained using categorical cross-entropy loss and optimized using the Adam optimizer.

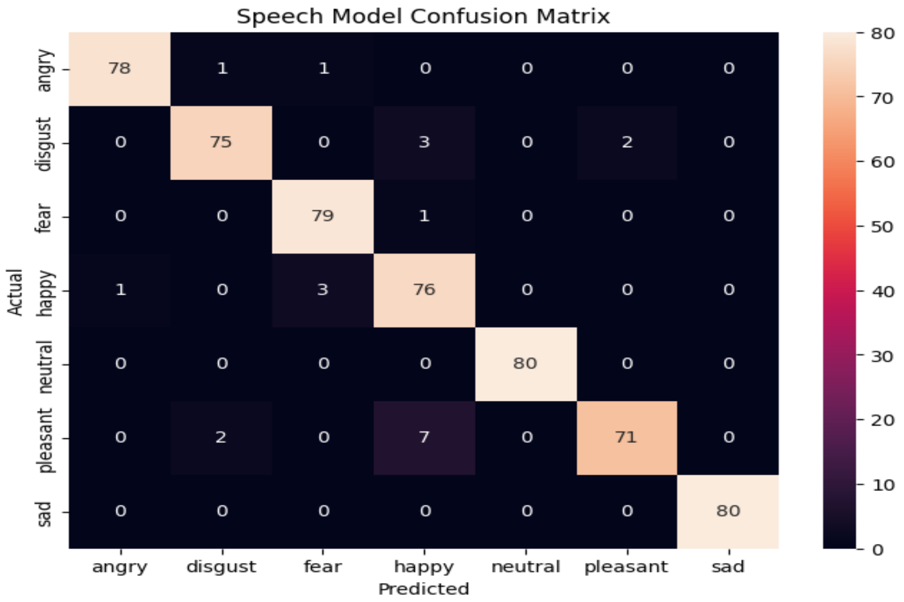
---

5. Experiments and Results

Three models were trained and evaluated:

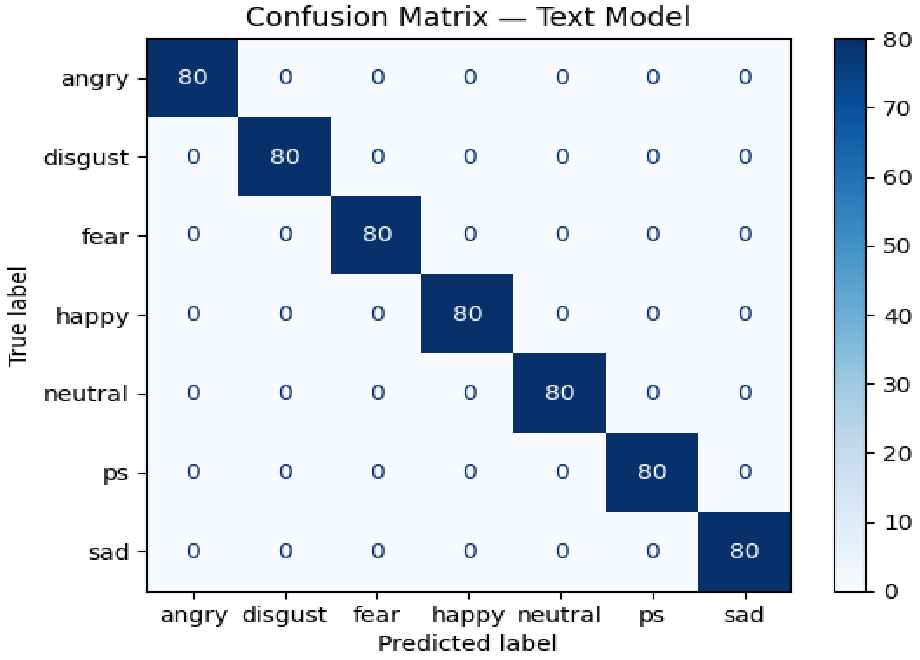
5.1 Speech Model

The speech-only model achieved an accuracy of approximately **96%**. This demonstrates that acoustic features effectively capture emotional information.



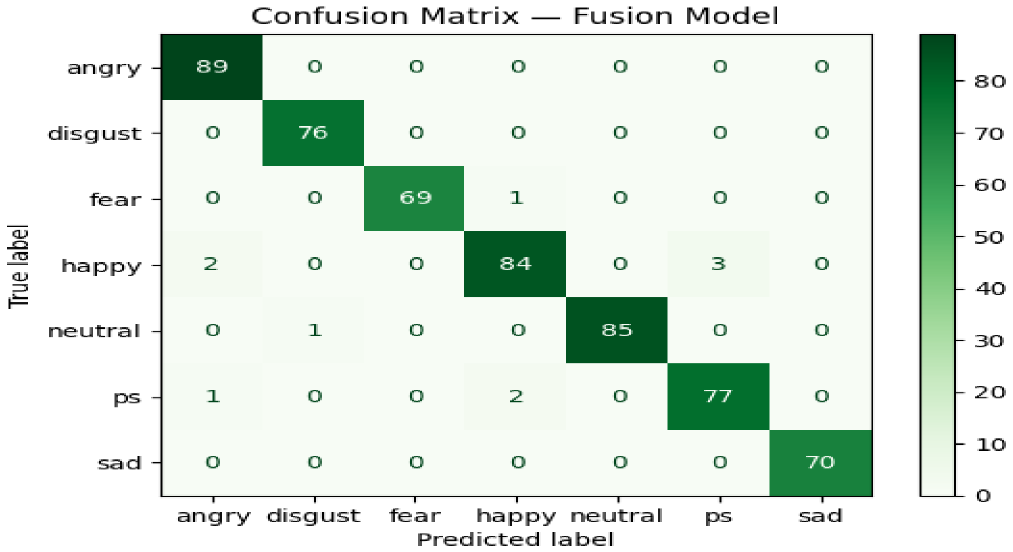
5.2 Text Model

The text-only model achieved **100% accuracy**. This high accuracy occurred because the generated textual input explicitly contained emotion-related words (e.g., “angry”, “happy”), making classification straightforward.



5.3 Fusion Model

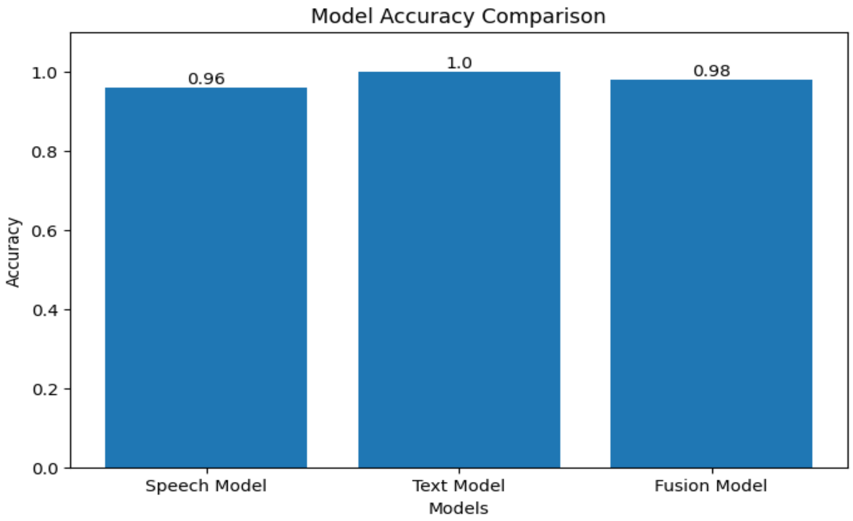
The multimodal fusion model achieved approximately **98% accuracy**, outperforming the speech-only model. This demonstrates the advantage of combining multiple modalities.



6. Model Comparison

The performance comparison of the three models is shown below:

Model	Accuracy
Speech Model	96%
Text Model	100%
Fusion Model	98%



The fusion model provides improved performance compared to the speech-only model due to complementary information from both modalities.

7. Analysis

7.1 Easiest Emotions to Classify

Emotions such as angry and happy were easier to classify due to strong acoustic and textual cues.

7.2 Hardest Emotions to Classify

Fear and surprise were comparatively more difficult to classify because of similar acoustic patterns and overlapping feature representations.

7.3 When Fusion Helps Most

Fusion improves classification performance when speech signals are ambiguous but textual information provides additional context. The combination of modalities enhances robustness and accuracy.

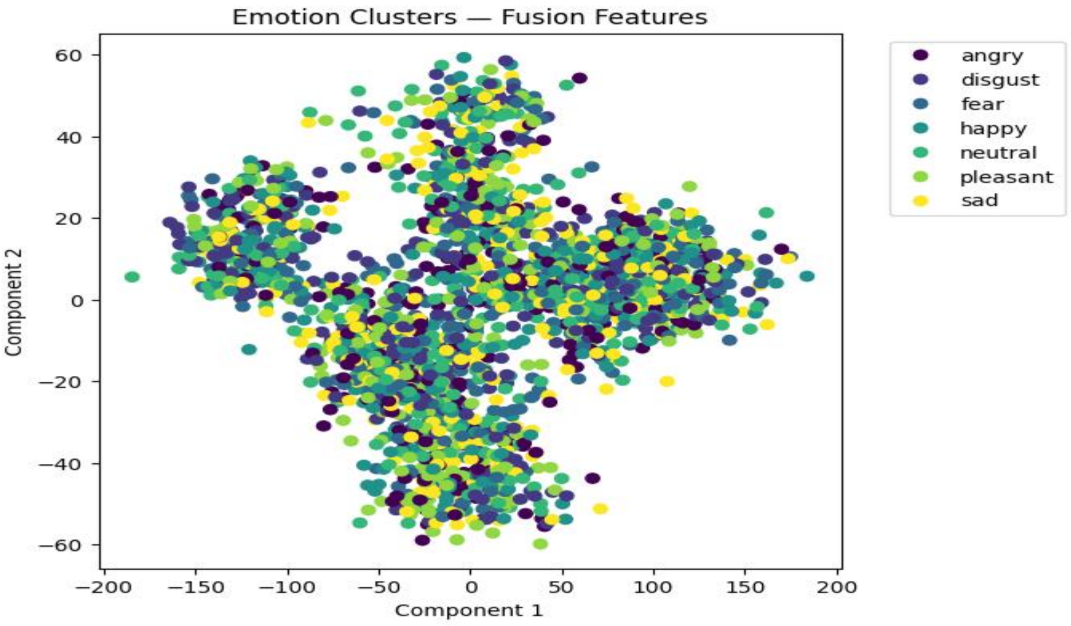
7.4 Error Analysis

Some misclassifications occurred between acoustically similar emotions such as fear and surprise. These errors indicate overlapping feature distributions and limitations in feature separability.

---

8. Visualization of Learned Representations

Dimensionality reduction techniques such as Principal Component Analysis (PCA) were applied to visualize emotion clusters. The visualization showed good separation between emotion classes, indicating that the learned representations effectively capture emotional characteristics



9. Discussion

The multimodal approach demonstrated superior performance compared to unimodal models. The combination of speech and text features provides complementary information that enhances emotion recognition accuracy.

However, the text model achieved perfect accuracy mainly due to explicit emotion words in the input, which may not generalize well to real-world scenarios.

Future improvements could include:

- Using real transcripts instead of generated text
- Applying advanced deep learning architectures
- Incorporing attention mechanisms

---

10. Conclusion

This project implemented a multimodal emotion recognition system using speech and text inputs. Three models were developed: speech-only, text-only, and fusion models. The fusion model achieved the best performance, demonstrating the effectiveness of multimodal learning for emotion recognition tasks.

The results confirm that combining complementary modalities significantly improves prediction accuracy and robustness compared to single-modality approaches.

---

11. References

- Toronto Emotional Speech Set (TESS) Dataset
- TensorFlow and Keras Documentation
- Librosa Audio Processing Library
- Scikit-learn Machine Learning Library

Prepared by sameeksha.