A Report on

MULTIMODAL EMOTION RECOGNITION

Author*:* Satwik Rakhelkar

Contact*:* rakhelkarsatwik@gmail.com | +91 9246840109

Submitted to: LTRSC / IIITH

1. Introduction

1. The project implements three emotion recognition pipelines: speech, text, and fusion.

2. The dataset used is TESS.

3.The objective was to evaluate unimodal performance and demonstrate the advantage of multimodal fusion.

2. Methodology

1. Speech pipeline: CNN and LSTM trained on spectrogram features.

2. Text pipeline: BERT fine tuned on transcripts.

3. Fusion pipeline: Combined speech and text embeddings passed through a joint classifier.

4. Dataset: TESS audio and transcripts organized via fusion_metadata.csv.

5. Evaluation: Accuracy computed on held out test sets.

3. Architecture Decisions

1. Speech pipeline used CNN and LSTM because CNN captures spectral features and LSTM models temporal dependencies.

2. Text pipeline used BERT because its transformer architecture learns contextual meaning across tokens effectively.

3. Fusion pipeline concatenated speech and text embeddings to combine complementary information from both modalities.

4. Classifier used a fully connected layer with softmax activation to predict emotion labels.

5. Preprocessing included silence trimming, resampling for speech, and tokenization for text to ensure clean inputs.

4. Results

1. Speech pipeline accuracy was 15.38%, showing poor convergence and weak classification performance.

2. Text pipeline accuracy was 28.57%, reflecting undertraining and limited contextual learning.

3. Fusion pipeline accuracy was 100.00%, demonstrating perfect separation of emotions and validating the benefit of multimodal integration.

5. Error Analysis

1. Speech pipeline accuracy was low due to incomplete convergence. Causes include label mismatch, insufficient training epochs, or preprocessing inconsistencies.

2. Text pipeline underperformed compared to expected benchmarks. This suggests the model was saved after minimal training and requires further fine tuning.

3. Fusion pipeline achieved perfect accuracy, confirming that multimodal integration successfully captured complementary features from speech and text.

4. Guidance from the AI assistant helped in identifying dataset path issues and fixing model input errors, which made debugging faster and smoother.

6. Analysis

1. Anger and happiness were easiest to classify because they show strong acoustic cues and distinct word usage.

2. Neutral and sad were hardest to classify because they have subtle differences in tone and overlapping vocabulary.

3. Fusion helped most when speech alone was noisy or text alone was ambiguous, such as distinguishing neutral from sad.

4. Neutral speech was sometimes misclassified as sad when tone was flat but transcript was neutral.

5. Happy text was sometimes misclassified as neutral when transcripts lacked explicit emotional words.

6. Fear and surprise were occasionally confused because of similar vocal intensity.

7. Visualization of embeddings showed clear separation for anger and happiness, overlap for neutral and sad, and tighter clustering with fusion.

7. Conclusion

1. Fusion clearly outperforms unimodal approaches, validating the hypothesis that combining modalities improves emotion recognition.

2. Future work should include longer training for speech and text pipelines, consistent label encoding across datasets, and hyperparameter tuning for BERT fine tuning.