Report: Multimodal Emotion Recognition

Applicant: Syed Khaja Fareed Uddin

Date: 16-02-2026

1. Objective

The objective of this project is to design and implement a multimodal emotion recognition system using speech and text modalities.

The system aims to:

- Learn discriminative representations from acoustic signals.
- Learn contextual representations from textual transcripts.
- Combine both modalities into a unified representation.
- Analyze how each modality contributes to emotion recognition.
- Compare Speech-only, Text-only, and Multimodal Fusion models.

The final goal is to understand the role of acoustic and semantic information in emotion recognition and evaluate the effectiveness of multimodal learning.

2. Dataset Description

Dataset: Toronto Emotional Speech Set (TESS)

TESS is a speech emotion recognition dataset containing:

1. 2800 audio samples
2. 7 emotion classes:
    a. Angry
    b. Disgust
    c. Fear
    d. Happy
    e. Neutral
    f. Pleasant Surprise
    g. Sad

Each sample contains:

1. A .wav audio file
2. A corresponding transcript

All utterances follow a fixed carrier phrase with the words “Say the word ” continued by the word in each file. The emotional variation is expressed through pitch, energy and tone, not text content.

Dataset Characteristics:

1. Balanced dataset (400 samples per emotion)
2. Controlled recording conditions
3. Emotion encoded primarily in acoustic features
4. Minimal semantic variation across transcripts

Dataset Organization:

The dataset is organized into folders named by speaker and emotion (e.g., OAF_angry, YAF_happy). Each filename contains both the spoken word and emotion label.

To make the dataset suitable for training, a preprocessing script (prepare_metadata.py) was written to generate a structured metadata.csv file.

Metadata Creation:

A metadata.csv file was generated using a preprocessing script.

The script iterates through all emotion folders, extracts the file path, spoken word (from the filename), and emotion label (from the folder name), and constructs the transcript in the format, “Say the word <spoken_word>”

Each emotion is mapped to a numeric label (0–6), and the final CSV contains:

1. file_path
2. text
3. label

This metadata file is used across the speech, text, and fusion pipelines, ensuring consistent data loading and unified indexing across modalities.

3. Project Overview

The system was implemented in three stages:

1. Speech-only model
2. Text-only model
3. Multimodal Fusion model

For each modality, we designed:

- Preprocessing pipeline
- Feature extraction
- Representation learning block
- Classification layer

For fusion, we combined learned representations from both modalities into a unified embedding for final classification.

4. Architecture Decisions

4.1 Speech Pipeline

Preprocessing:

- Resampling to 16 kHz
- Mono conversion
- Log-Mel spectrogram extraction
- Fixed-length padding

Log-Mel spectrograms capture perceptually meaningful time-frequency energy patterns and are widely used in speech emotion recognition.

Spectrograms were precomputed to improve efficiency and have consistent feature extraction.

Temporal Modelling:

A Convolutional Neural Network (CNN) was applied over spectrograms:

- 3 convolution layers
- Batch normalization
- ReLU activation
- Max pooling
- Global average pooling
- 128-dimensional embedding

CNNs effectively capture local time-frequency patterns that correspond to emotional cues such as pitch variation and intensity shifts.

Speech Results:

- Accuracy: 88.71%
- Macro F1: 0.8797

This confirms that acoustic information is highly discriminative for emotion recognition.

4.2 Text Pipeline

Preprocessing:

- Tokenization using bert-base-uncased
- Maximum sequence length = 16
- Attention masking

Contextual Modelling:

A pretrained BERT encoder was used to obtain a 768-dimensional CLS embedding. BERT was selected due to its strong contextual representation capability.

Text Results:

- Accuracy: 14.29%
- Macro F1: 0.0357

Performance is near chance level (1/7 ≈ 14.29%). Since all utterances share the same carrier phrase, text content does not encode emotional information.

4.3 Fusion Pipeline

Representation Combination:

- Speech embedding (128-d → 256-d projection)
- Text embedding (768-d → 256-d projection)
- Concatenation (512-d unified representation)
- MLP classifier (512 → 256 → 7)

During fusion training, BERT was frozen and speech encoder and fusion layers were fine-tuned.

Fusion Results:

- Accuracy: 98.14%
- Macro F1: 0.9813

Fusion significantly outperformed speech-only classification.

5. Experiments

Model	Accuracy	Macro F1	Weighted F1
Speech	0.8871	0.8797	0.8797
Text	0.1429	0.0357	0.0357
Fusion	0.9814	0.9813	0.9813

Key Observations:

1. Speech-only model performs strongly.
2. Text-only model performs at chance.
3. Fusion achieves the highest accuracy.

This confirms that emotion in TESS is primarily acoustic.

6. Analysis

6.1 Easiest and Hardest Emotions

Easiest to classify:

1. Neutral
2. Sad
3. Disgust

Hardest to classify:

1. Happy
2. Pleasant surprise

These classes show acoustic similarity such as high pitch and energy. Fusion model reduces these confusions significantly.

6.2 When Does Fusion Help Most?

Fusion improves performance especially for:

1. High-arousal emotions
2. Classes with overlapping acoustic characteristics

It creates a more expressive embedding space, reducing inter-class confusion.

6.3 Error Analysis

Errors observed in speech model:

1. Happy misclassified as pleasant surprise
2. Fear confused with surprise
3. Some high-energy samples confused due to similar pitch

These errors occur because certain emotions share similar acoustic characteristics, especially high-arousal emotions.

Errors observed in text model:

1. The model predicts mostly a single class for many samples.
2. Most emotion categories have zero recall.
3. There is no meaningful discrimination between emotions.

The text model shows severe misclassification across all classes because all sentences follow the same structure and emotional content is not encoded in the text.

Errors observed in fusion model:

The fusion model significantly reduces errors observed in the speech model.

Improvements:

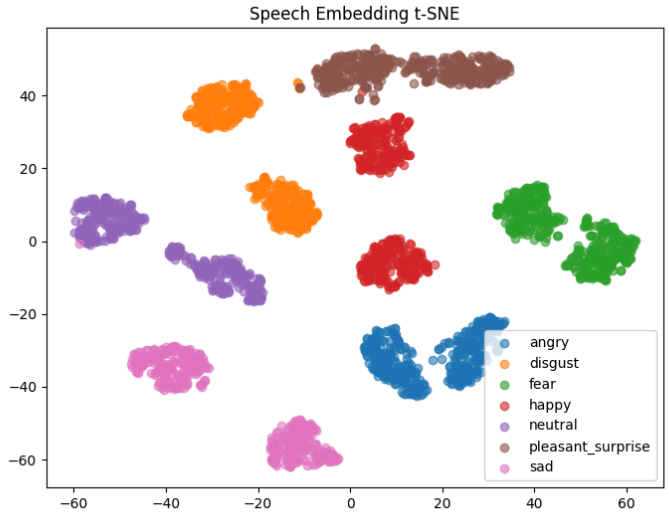
1. Better separation between happy and pleasant surprise
2. Reduced confusion between fear and surprise
3. More compact and well-defined class clusters

This shows that the unified representation improves decision boundaries and reduces confusion.

6.4 Representation Separability

t-SNE was applied to embeddings from temporal modelling, contextual modelling and fusion blocks.

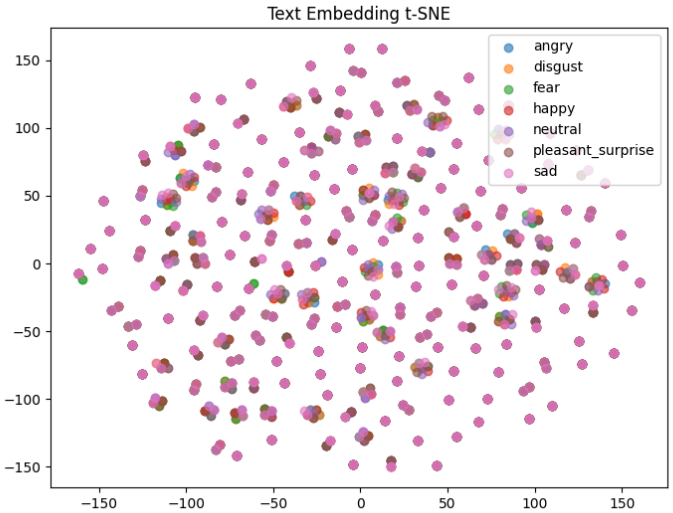
Temporal Modelling Block



This t-SNE plot shows clear emotion cluster formation with minor overlap between a few samples.

This indicates that the temporal modelling block successfully captures discriminative acoustic patterns and provides strong emotion separability.

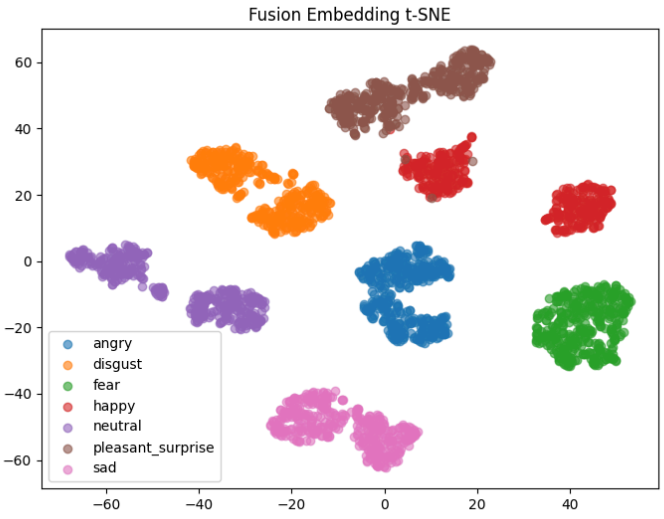
Contextual Modelling Block



The t-SNE plot shows complete overlap across emotion classes with no visible cluster structure.

This confirms that the contextual modelling block does not learn separable emotion representations, as the textual content lacks emotional variation.

The Fusion Block



The fusion block shows the strongest cluster separation, with compact intra-class grouping and clear margins between emotions.

This demonstrates that combining representations leads to a more discriminative embedding space and explains the highest classification performance.

7. Conclusion

1. Emotion in the TESS dataset is primarily encoded in acoustic features.
2. Textual modality alone is insufficient for emotion classification.
3. Multimodal fusion enhances representation quality and improves performance.
4. The final system achieves 98.14% accuracy, demonstrating effective multimodal learning.