Multimodal Emotion Recognition

Student Details

Name: Katta Bhavana
Email: kattabhavana34@gmail.com
Institute Email: 22211a6757@bvrit.ac.in
GitHub Repository:
https://github.com/kattabhavana9/Multimodal-Emotion-Recognition

1. Introduction

Emotion recognition is an important task in human-computer interaction, speech analytics, and affective computing. This assignment covers the design and implementation of a multimodal emotion recognition system utilizing the Toronto Emotional Speech Set dataset.

It was implemented by three different approaches:

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

It was not only to create these models but also to compare the performance and test how the combination of modalities influences classification accuracy.

2. Dataset Description

The Toronto Emotional Speech Set (TESS) is an audio data set with spoken words delivered under varying emotional conditions. The dataset is comprised of:

- 8 emotion classes
- Speech recordings from different speakers
- Corresponding text (spoken word)

These words are neutral in and of themselves ("name," "search," etc.), and the emotion is contained in tone of voice.

The following characteristic of the dataset also turned out to be important for the results obtained:

3. System Architecture

The system comprises of five functional blocks:

1. Preprocessing
2. Feature Extraction
3. Temporal/Contextual Modelling
4. Fusion
5. Classification

The structure for each of the pipelines (speech-only, text-only, fusion).

4. Architecture Decisions

4.1 Speech Pipeline

Preprocessing

- All audio clips were resampled to 16 kHz.
- Each clip was trimmed or padded out to a constant length of 3 seconds.
- MFCC features, which consist of 40 coefficients, are Silence trimming and normalization were used.

Reason:
MFCC are indeed those features that capture important frequency-domain characteristics of speech, reflecting emotional cues related to pitch and energy.

Feature Extraction

MFCC features were used as input representations in the format:

(time steps × 40 features)

Reason:
MFCC is commonly used in speech processing because it models human auditory perception very well.

Temporal Modelling

The architecture used:

- Convolutional Neural Network (CNN)
- Bidirectional LSTM (BiLSTM)

CNN learns short-term local acoustic patterns.

By using BiLSTM, it can effectively extract the long-term temporal relationships in both directions.

Reason:
Emotions in verbal communication have an evolutionary base, and temporal modeling is used to track such patterns.

Classifier

- Fully connected layer
- Softmax activation
- CrossEntropy loss

4.2 Text Pipeline

Preprocessing

- Spoken word extracted from filename.
- Tokenized using BERT tokenizer.

Contextual Modelling

Model used:

- BERT (bert-base-uncased)

BERT generates contextual embeddings of size 768.

Reason:

BERT provides strong contextual representations and is effective for text classification tasks.

Classifier

- Dropout layer
- Fully connected layer
- Softmax activation

4.3 Fusion Pipeline

Fusion was implemented using late fusion:

- Speech embedding extracted from CNN + BiLSTM
- Text embedding extracted from BERT
- Concatenation of both embeddings
- Fully connected layers for classification

Reason:
Late fusion allows independent learning of modality-specific representations before combining them.

5. Experiments

Three experiments were conducted:

Model	Test Accuracy
Speech-only	86.07%
Text-only	13.21%
Fusion	57.50%

6. Analysis

6.1 Easiest Emotions to Classify

Based on the confusion matrix:

- Happy
- Disgust
- Angry

These emotions have strong acoustic characteristics such as higher pitch, intensity, and dynamic variations. These distinctive features make them easier to classify.

6.2 Hardest Emotions to Classify

- Neutral
- Fear

These emotions are more subtle and share overlapping acoustic features with other classes. For example:

- Neutral and sad both have lower energy.
- Fear and angry share higher pitch characteristics.

This leads to misclassification.

6.3 When Does Fusion Help?

Fusion improved performance compared to the text-only model. However, it did not outperform the speech-only model.

This is mainly because:

- The TESS dataset contains neutral words.
- The text does not carry emotional information.
- Emotion is expressed only through speech tone.

If emotional sentences were used instead of neutral words, fusion would likely provide better performance.

6.4 Error Analysis

Some observed failure cases:

1. Fear misclassified as Angry due to similar pitch and energy.
2. Neutral confused with Sad because of low-energy patterns.
3. Certain classes were predicted incorrectly due to overlapping acoustic features.
4. Fusion model misclassified some samples because text introduced noise instead of meaningful context.
5. Minor class imbalance also influenced predictions.

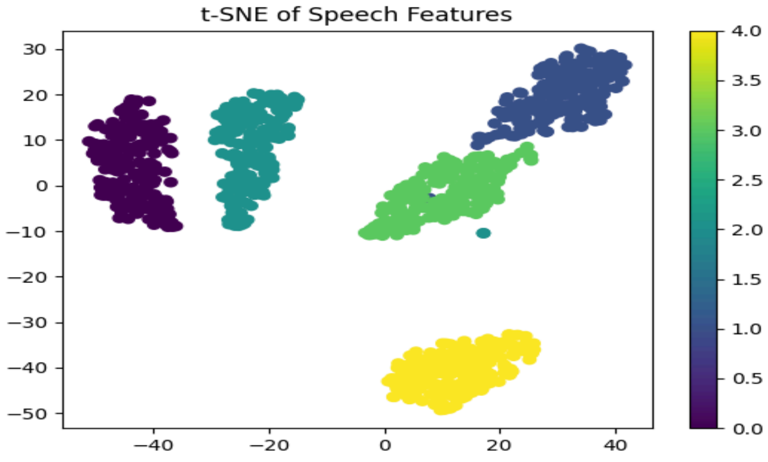
7. Visualization of Learned Representations

t-SNE visualizations were generated using learned embeddings from:

- Temporal modelling block (Speech)
- Contextual modelling block (Text)
- Fusion block (Combined representation)

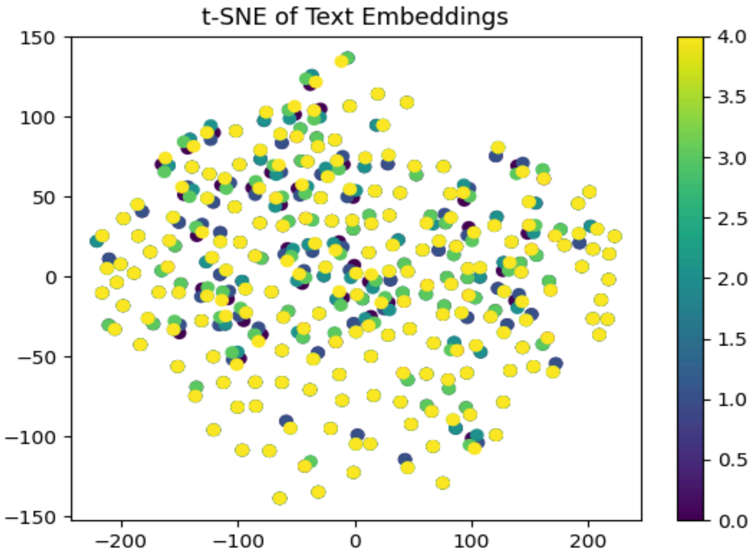
Speech t-SNE

Clear clustering of emotions observed.
This confirms strong separability of speech embeddings.



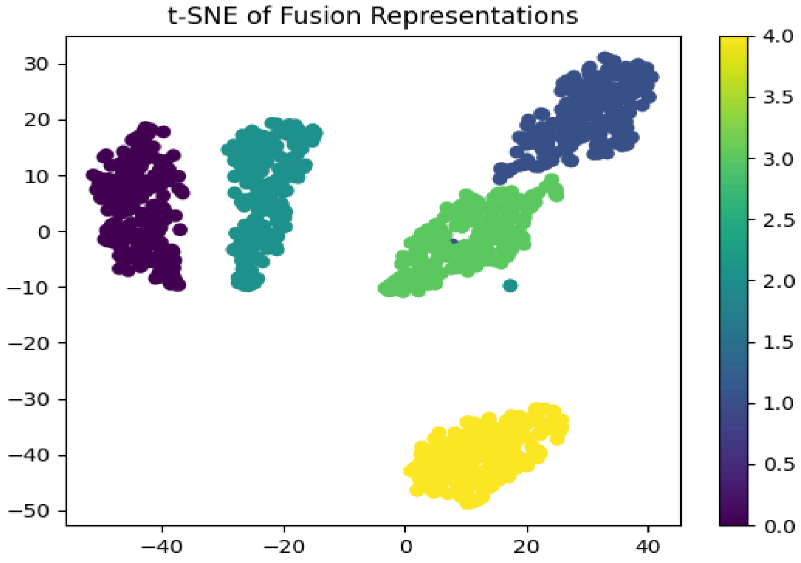
Text t-SNE

Significant overlap between clusters.
This indicates limited emotional information in text.



Fusion t-SNE

Moderate clustering observed.
Fusion improves over text-only but does not surpass speech-only.



Conclusion

This assignment demonstrates that:

- Speech is the dominant modality for emotion recognition in the TESS dataset.
- Text-only modelling performs poorly because words are emotionally neutral.
- It enhances performance only if both sources of information are useful.
- Temporal modeling is an essential area for modeling emotional characteristics in speech.

In general, this project offered practical knowledge regarding the concept of multimodal learning and the significance of dataset characteristics.