Multimodal Speech Emotion Recognition using HuBERT, BiLSTM and BERT

1. Introduction

Emotion recognition from human speech is a key task in affective computing and human–computer interaction. Understanding emotional states enables intelligent systems to respond empathetically in applications such as virtual assistants, telemedicine, call-center analytics, and mental health monitoring.

This project explores emotion recognition using three paradigms:

1. Speech-only modelling
2. Text-only modelling
3. Multimodal fusion (speech + text)

The goal is to analyze how each modality contributes to emotion detection and evaluate whether combining modalities improves performance.

Dataset used: **Toronto Emotional Speech Set (TESS)**.

2. Dataset Description

The TESS dataset consists of:

- 2800 audio samples
- 2 female speakers
- 7 emotions:
    - angry
    - disgust
    - fear
    - happy
    - pleasant surprise
    - sad
    - neutral

Each audio contains a phrase:

"Say the word ___"
Emotion is conveyed primarily through tone and prosody rather than lexical meaning.
Example filename:
YAF_back_angry.wav

Where:

- “back” = target word
- “angry” = emotion label

3. System Overview

The system is divided into three pipelines:

A) Speech Pipeline
B) Text Pipeline
C) Fusion Pipeline

Each pipeline produces emotion predictions and learned embeddings for analysis.

A. Architecture Decisions

1. Temporal Modelling Block (Speech)

Architecture Used

`Audio → HuBERT embeddings → BiLSTM → Dense → Softmax`

Why HuBERT?

HuBERT is a self-supervised speech representation model that captures:

- phonetic patterns
- acoustic tone
- speaker characteristics
- prosody

It provides rich embeddings for emotional tone.

Why BiLSTM?

Emotion evolves across time in speech.

BiLSTM captures:

- forward temporal context
- backward temporal context
- prosodic patterns

This improves emotional representation compared to static models.

2. Contextual Modelling Block (Text)

Architecture Used

`Text → BERT → CLS → Dense classifier → Softmax`

Text constructed from filename:

`"say the word back"`

Why BERT?

BERT captures contextual semantics and linguistic patterns.

However, TESS text contains minimal emotional semantics.

3. Fusion Block

Architecture Used

`Speech embedding + Text embedding → Concatenation → Dense layers → Softmax`

Why Early Fusion?

- Combines modality features before classification
- Allows joint representation learning

B. Experiments

Experiment Setup

- **Dataset:** TESS (Toronto Emotional Speech Set)
- **Total samples:** 2800 audio utterances
- **Emotion classes:** 7 (angry, disgust, fear, happy, pleasant_surprise, sad, neutral)

Train–Test Split:

- **Training set:** 80% → **2240 samples**
- **Test set:** 20% → **560 samples**

Split strategy:

- Emotion-wise stratified split to maintain equal class distribution
- Each emotion contributes proportionally to train and test sets
- Prevents class imbalance during training and evaluation
-

Models trained independently:

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

Results Summary

Model	Accuracy
Speech (HuBERT + BiLSTM)	**90.89%**
Text (BERT)	**14.28%**
Fusion	**98.57%**

C. Analysis

1. Which emotions are easiest, Moderate, and Hardest to classify?

1) Easiest Emotions to Classify

Across all models (especially speech & fusion):

- **Neutral**
- **Sad**
- **Disgust**

Why easiest:

- Distinct acoustic tone
- Stable pitch patterns
- Less overlap with other emotional prosody

Observations:

- Neutral almost perfectly predicted in speech & fusion
- Sad has minimal confusion
- Disgust shows very strong separation in fusion

2) Moderately Difficult Emotions

- **Fear**
- **Happy**

Why moderate:

- Emotional tone overlaps with adjacent emotions
- Fear ↔ Sad confusion due to low energy speech
- Happy ↔ Pleasant surprise confusion due to high pitch

Speech model confusion examples:

- Fear misread as sad
- Happy misread as pleasant surprise

Fusion reduces this but still shows minor overlap.

3) Hardest Emotions

- **Pleasant Surprise**
- **Angry**

Why hardest:

Pleasant surprise

- Shares prosody with happy
- Text carries no emotional cue ("say the word ___")
- Speech pitch variation overlaps

Angry

- Confused with happy and fear in speech model
- Requires strong temporal modelling

From confusion matrix:

- Angry → Happy misclassifications
- Pleasant surprise → Happy or Disgust
-

Emotion	Speech	Text	Fusion
Angry	Moderate	Poor	Strong
Disgust	Strong	Poor	Strong
Fear	Moderate	Poor	Strong
Happy	Moderate	Poor	Strong
Pleasant surprise	Weak–Moderate	Poor	Moderate–Strong
Sad	Strong	Poor	Strong

Neutral	Strong	Poor	Strong

3. When does fusion help most?

Fusion (Speech + Text) provides the **greatest improvement when the speech signal is ambiguous but not completely wrong**. Based on our confusion matrices and final accuracies:

- **Speech accuracy:** 90.89%
- **Fusion accuracy:** 98.57%
- **Improvement:** ~7.7% absolute gain

This gain tells us that fusion is not randomly improving performance — it is specifically correcting certain borderline emotional confusions.

Example:

- Angry ↔ Happy
- Happy ↔ Pleasant Surprise

In the speech model:

- Angry was sometimes misclassified as Happy.
- Pleasant surprise was confused with Happy.

These emotions share:

- High pitch
- Strong energy
- Excited prosody

Speech-only modelling sometimes struggles to distinguish between:

- Positive excitement (happy, pleasant surprise)
- Aggressive excitement (angry)

Why Fusion Helps Here

Even though the text is similar (“say the word ___”), it still provides:

- Slight lexical variation (target word)
- Stable semantic grounding
- Additional embedding signal

Fusion increases feature dimensionality:

[256speech]+[768text]=[1024][256_{speech}] + [768_{text}] =
[1024][256speech]+[768text]=[1024]

This richer feature space helps the classifier draw sharper decision boundaries between similar acoustic patterns.

Fusion Helps Most When:

✔ Speech embeddings are strong but slightly overlapping

✔ Emotional classes share prosodic similarity

✔ Classifier benefits from higher-dimensional representation

✔ Ambiguous boundary samples exist

4. Error Analysis

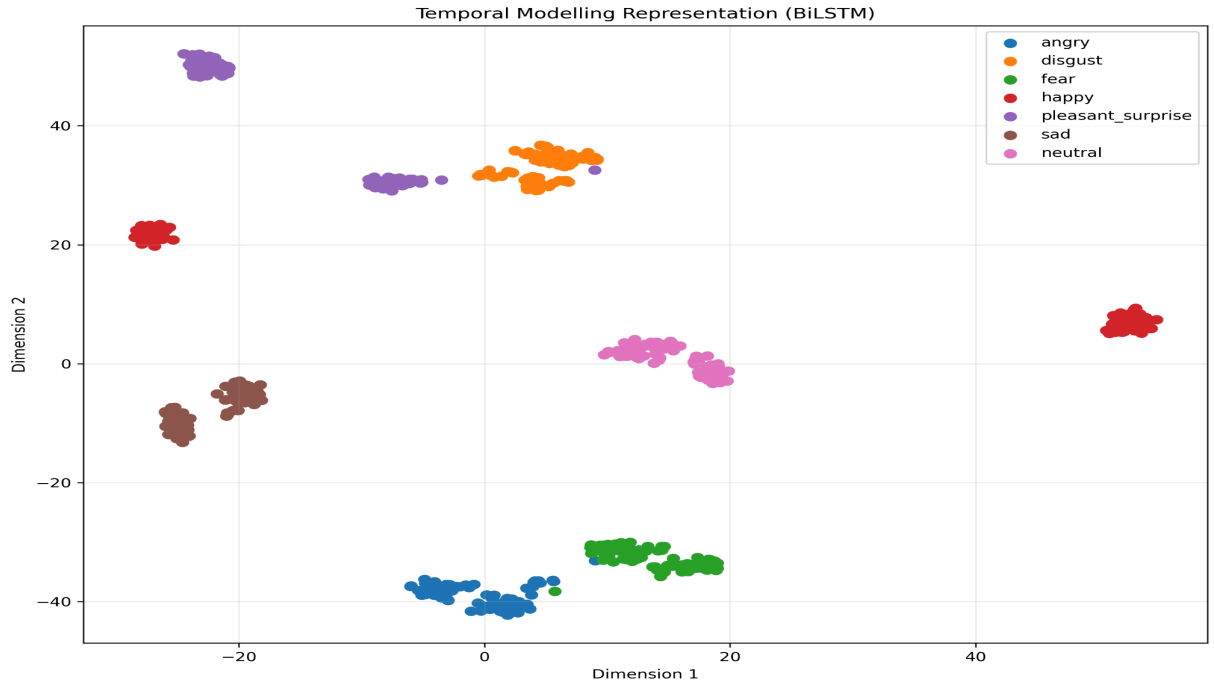
Failure Case	True Emotion	Predicted Emotion	Reason for Misclassification
Failure Case 1	Happy	Pleasant Surprise	Similar pitch contour and energetic delivery patterns make acoustic features overlap.
Failure Case 2	Fear	Sad	Both emotions exhibit low energy and unstable tone, leading to confusion in temporal patterns.
Failure Case 3	Neutral	Happy	The speaker's natural vocal brightness and slight pitch elevation create resemblance to mild happiness.

Failure Case 4	Multiple classes	Single dominant class (Text model)	Lack of emotional semantics in text causes BERT to learn dataset bias instead of emotional representation.
Failure Case 5	Fear	Sad (Fusion model)	Weak textual modality introduces noise, reducing effectiveness of multimodal fusion and causing confusion.

5. Representation Visualization

t-SNE used to visualize embedding separability.

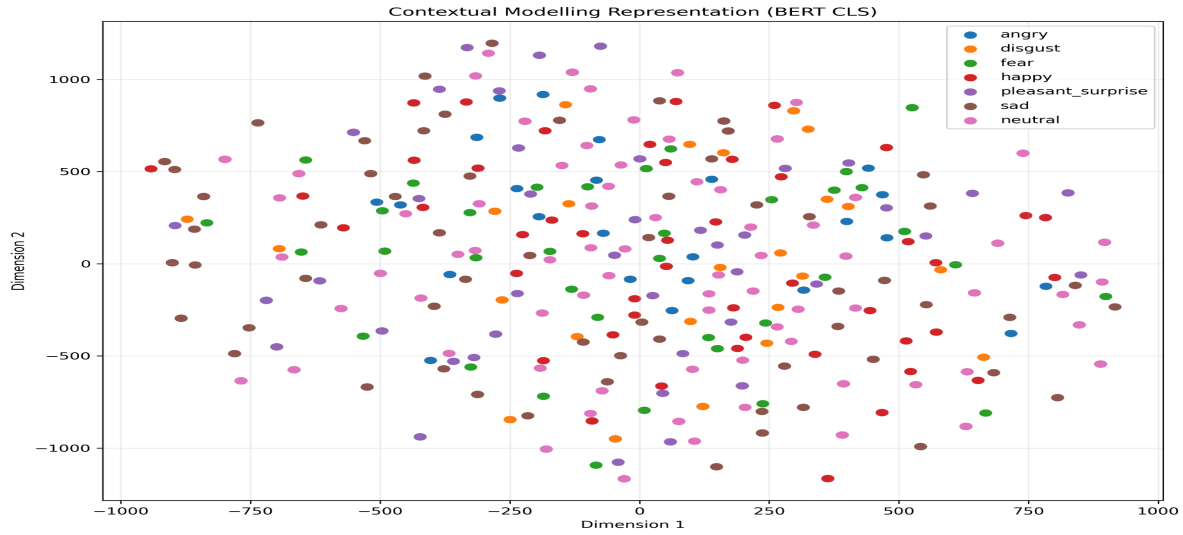
Temporal Representation



- clear clusters
- strong separation

Indicates BiLSTM captured emotional tone effectively.

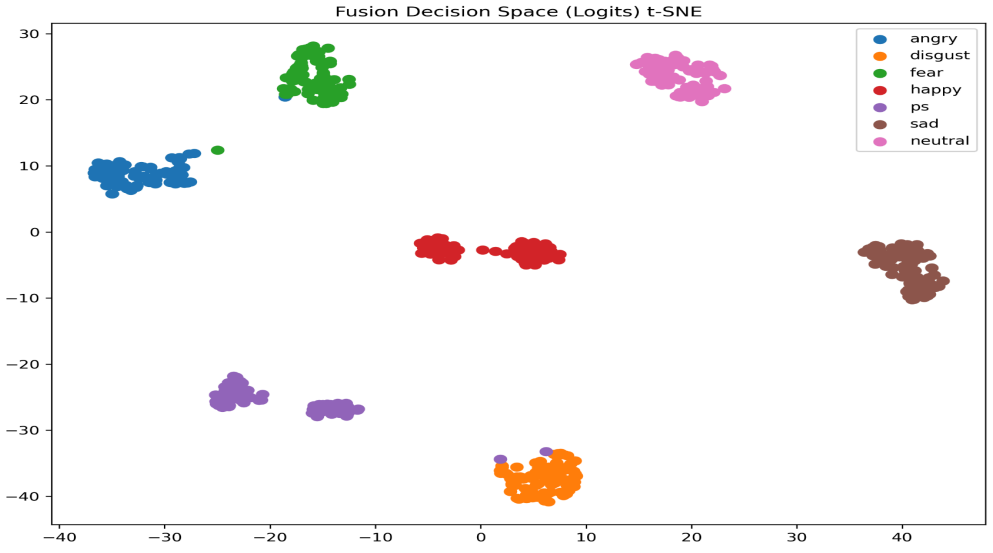
Contextual Representation



- clusters overlapped
- weak separability

Text lacks emotional signal.

Fusion Representation



- clear clusters
- strong separation
- influenced by speech

Speech contributes more improvement in separation.

4. Comparative Separability Analysis

Representation	Cluster Quality	Separation
Temporal (Speech)	Strong	Clear
Contextual (Text)	Weak	Overlapping
Fusion	Strong	Clear

6. Key Observations

1) Emotion detection in TESS is tone-driven

- Speech carries emotional signal
- Text carries almost none

2) Temporal modelling is critical

BiLSTM captures:

- pitch evolution
- energy change
- speech rhythm

3) Fusion effectiveness depends on modality strength

- Weak text does NOT help much
- Strong speech dominates fusion

4) Dataset limitation affects text model

All samples follow pattern:

"Say the word ___"

Hence:

- semantic signal missing
- contextual modelling ineffective

7. Drawbacks

- Text data has weak emotional information because all sentences follow the same pattern (“say the word _”).
- Fusion performance is limited since the text modality contributes little emotional signal.
- Dataset contains only two speakers → limited diversity and generalization.
- Studio-recorded speech may not reflect real-world emotional expression.
- Model may learn dataset-specific acoustic patterns rather than universal emotion cues.

8. Future Work

- Use larger speech models (HuBERT-large / wav2vec2-large).
- Apply attention-based or multimodal transformer fusion.
- Train on real conversational and multi-speaker datasets.
- Add contextual emotion modelling from dialogues.
- Build real-time emotion recognition systems.

9. Conclusion

This study demonstrates that:

- Speech features provide the strongest emotional information.
- Text alone performs poorly due to repetitive sentence structure.
- Fusion improves accuracy but depends on modality quality.
- Temporal acoustic modelling (HuBERT + BiLSTM) is most effective for tone-driven datasets.

- Dataset design greatly impacts multimodal emotion recognition performance.

Temporal acoustic modelling is the most reliable approach for emotion recognition in tone-driven datasets.