Name : Varsha Sajjanavar

Multimodal Emotion Recognition using Speech and Text

1. Introduction

Emotion recognition from human communication is a central problem in affective computing. Emotions are expressed through multiple channels, primarily speech prosody and linguistic content. Relying on a single modality can limit performance because emotional cues are often distributed across both acoustic and textual signals.

This project implements a complete multimodal emotion recognition system using the Toronto Emotional Speech Set (TESS). Three model variants are developed and compared:

- A speech-only model
- A text-only model
- A multimodal fusion model

The objective is to design appropriate architectures for preprocessing, feature extraction, temporal/contextual modeling, fusion, and classification, and to analyze how combining modalities influences performance.

---

2. Dataset Description

The experiments use the **Toronto Emotional Speech Set (TESS)**, which contains acted emotional speech recordings labeled with seven emotions:

- angry
- disgust
- fear
- happy
- neutral
- sad
- surprise

Each audio file contains a spoken word with a corresponding transcript derived from the filename. The dataset is balanced across emotion classes, which allows stratified splitting into training, validation, and test sets.

A stratified split is used:

- 72% training
- 8% validation
- 20% test

This ensures equal emotion distribution across all splits.

---

3. System Architecture

The system is organized into modular pipelines: speech, text, and fusion. Each pipeline follows the same functional blocks required by the assignment.

3.1 Preprocessing

Speech preprocessing

Speech signals are standardized to ensure consistent model input:

- Resampling to 16 kHz
- Silence trimming using energy-based detection
- Padding or truncation to a fixed 3-second duration
- Conversion to mel-spectrograms with 128 mel bands

Mel-spectrograms provide a time–frequency representation that preserves emotional prosodic cues such as pitch, energy, and spectral shape.

Text preprocessing

Text transcripts are tokenized using a pretrained tokenizer. Each sentence is:

- Lowercased automatically by the tokenizer
- Tokenized into subword units
- Padded or truncated to a fixed sequence length

This creates consistent token sequences suitable for transformer-based modeling.

3.2 Feature Extraction

Speech features

Mel-spectrograms serve as input features with shape:

time steps × mel features

These features encode short-term spectral energy patterns that correlate with emotional expression.

Text features

Token embeddings are extracted from a pretrained transformer language model. These embeddings capture semantic and contextual emotional information.

3.3 Temporal and Contextual Modeling

Speech temporal modeling

A bidirectional LSTM processes mel-spectrogram sequences. The bidirectional design allows the model to learn emotional patterns across time in both forward and backward directions.

The final utterance representation is obtained by mean pooling over time.

This architecture is chosen because:

- LSTMs model temporal dependencies effectively
- Bidirectional processing captures full context
- The structure is lightweight and stable for audio sequences

Text contextual modeling

The text pipeline uses BERT base uncased as a contextual encoder. The [CLS] token embedding represents the entire sentence.

BERT is selected because it provides strong contextual representations and is effective for classification tasks with limited training data.

3.4 Fusion Strategy

The fusion model combines:

- A 128-dimensional speech embedding
- A 128-dimensional projected text embedding

These vectors are concatenated to form a unified multimodal representation. A fully connected classifier predicts emotion labels.

Early fusion at the embedding level is chosen because it:

- Preserves modality-specific structure
- Allows joint learning of cross-modal interactions
- Keeps the architecture simple and interpretable

3.5 Classifier

All pipelines use a linear classification layer followed by softmax. Cross-entropy loss is optimized using Adam or AdamW optimizers. Early stopping prevents overfitting based on validation loss.

---

4. Experimental Setup

All models were trained using PyTorch with GPU acceleration enabled when available. The same train/validation/test split was used across all pipelines to ensure fair comparison.

Key hyperparameters were selected empirically and kept consistent across experiments:

- **Speech model:** Bidirectional LSTM with hidden size 128 (per direction)
- **Batch size:** 16 for speech, 8 for text and fusion models
- **Learning rate:** 1e-4 for speech and fusion, 2e-5 for the text model
- **Early stopping:** patience of 3 epochs based on validation loss

Model performance was evaluated using multiple complementary metrics:

- **Accuracy** for overall classification performance

- **Weighted F1 score** to account for class balance
- **Confusion matrices** to analyze class-wise errors
- **t-SNE visualization** to inspect the structure of learned embeddings

These evaluation tools provide both quantitative and qualitative insight into how each model separates emotional categories.

---

5. Results

5.1 Quantitative Performance

Model	Accuracy	F1 Score
Speech	0.9821	0.9821
Text	1.0000	1.0000
Fusion	1.0000	1.0000

The text and fusion models achieve perfect classification on the test set, while the speech model performs slightly lower but still very strongly.

An important observation is that the text model achieves perfect accuracy because the transcripts in the Toronto emotional speech set explicitly contain the emotion-bearing word (for example, “happy” or “sad”). This allows the model to directly associate specific keywords with emotion labels. Therefore, the text results mainly reflect lexical matching rather than deep emotional language understanding. This characteristic is a property of the dataset and should be considered when interpreting the performance of the text and fusion models.

5.2 Confusion Matrix Analysis

The speech model shows minor confusion between acoustically similar emotions, particularly neutral, sad, and surprise. The text and fusion models exhibit near-perfect diagonals, indicating clean separation.

These results suggest that semantic information from text strongly disambiguates emotional categories.

---

6. Error Analysis

Five representative speech model errors were examined:

(5, 6 → 3), (47, 6 → 3), (49, 6 → 4), (335, 3 → 6), (367, 6 → 1)

We examined five representative misclassified samples from the speech model to understand its limitations (each tuple is: index, true label → predicted label):

- **Sample 5:** Surprise → Happy
- **Sample 47:** Surprise → Happy

- **Sample 49:** Surprise → Neutral
- **Sample 335:** Happy → Surprise
- **Sample 367:** Surprise → Disgust

Most errors involve surprise being confused with other emotions, especially happiness. Surprise and happiness share high pitch and strong energy patterns, which can make them acoustically similar. In some cases (e.g., Sample 49), lower emotional intensity leads the model to predict neutral instead.

The confusion between happy and surprise (Sample 335) suggests overlap in expressive features such as elevated pitch and tempo. The rare confusion with disgust (Sample 367) may result from abrupt spectral changes that resemble harsher vocal qualities.

Overall, these errors indicate that the model performs well on broad emotional categories but struggles with fine-grained distinctions between closely related high-arousal emotions. This trend is consistent with the confusion matrices and the partial overlap observed in the t-SNE emotion clusters.

---

7. Visualization of Learned Representations

t-SNE is used to project high-dimensional embeddings into two dimensions.

7.1 Speech embeddings

Speech embeddings form clear clusters with slight overlap between neutral, sad, and surprise. This explains the small number of speech-only errors.

7.2 Text embeddings

Text embeddings show strong class separation. Each emotion occupies a distinct region, consistent with perfect classification.

7.3 Fusion embeddings

Fusion embeddings are the most compact and well separated. Combining modalities produces highly discriminative representations with minimal overlap.

---

8. Comparative Analysis

Easiest emotions

Happy and angry are easiest to classify because they exhibit strong acoustic and semantic signals.

Hardest emotions

Neutral and surprise are harder in speech-only settings due to subtle prosodic differences.

When fusion helps most

Fusion is most beneficial when acoustic cues are ambiguous. Text provides stabilizing semantic information that resolves uncertainty.

---

9. Discussion

The experiments demonstrate that multimodal fusion improves robustness by combining complementary information sources. Speech captures expressive prosody, while text encodes semantic intent.

The strong performance of the text model suggests that the structured transcripts in TESS contain highly predictive emotional cues. In more natural conversational datasets, fusion would likely provide even greater benefits.

The t-SNE visualizations confirm that internal representations align with classification accuracy: better-separated clusters correspond to stronger performance.

9.1 Title: *Dataset Limitations*

Although the models achieve very high performance, the dataset has several limitations that affect generalization. The recordings are acted in a controlled studio environment and contain a limited vocabulary with clearly expressed emotions. Real-world emotional speech is often noisier, more spontaneous, and linguistically diverse. Because of this, performance on natural conversational data may be lower than what is observed in these experiments. The strong results should therefore be interpreted as performance on a clean benchmark dataset rather than a guarantee of real-world robustness.

---

10. Conclusion

This project implements a complete multimodal emotion recognition system using speech, text, and fusion pipelines. The architectures effectively model temporal and contextual information and demonstrate the value of multimodal integration.

Key findings:

- Speech alone provides strong emotional discrimination
- Text offers highly separable semantic representations
- Fusion produces the most stable and robust embeddings

Future work could explore attention-based fusion, larger speech encoders, and evaluation on spontaneous conversational datasets.

---

11. Reproducibility

All code is organized into modular pipelines with training and testing scripts. The project includes:

- Structured directory organization
- Requirements file
- Automated dataset download
- Saved model checkpoints
- Evaluation scripts and visualizations

This ensures that experiments can be reproduced by running the provided notebook sequentially.