MultiModal-ER

Multimodal Emotion Recognition
using Speech, Text, and Fusion Pipelines

Dataset: TESS (Toronto Emotional Speech Set) — 5,600 samples, 7 emotions

Framework: PyTorch | BiLSTM + Self-Attention

Transcription: OpenAI Whisper (base model)

Fusion: Late Fusion via Feature-Level Concatenation

Date: February 2026

Model	Test Accuracy	Parameters
Speech-only (MFCC + BiLSTM + Attention)	**100.00%**	718,600
Text-only (Whisper ASR + BiLSTM + Attention)	**15.00%**	684,488
Multimodal Fusion (Speech + Text Concatenation)	**99.76%**	1,501,129

Table of Contents

1. **A. Architecture Decisions**
    1. Preprocessing
    2. Feature Extraction
    3. Temporal / Contextual Modelling
    4. Fusion Strategy
    5. Classifier
    6. Architectural Diagram
2. **B. Experiments**
    1. Training Configuration & Dataset Statistics
    2. Overall Accuracy Comparison
    3. Per-Class Test Accuracy
    4. Training Dynamics & Curves
    5. Confusion Matrices
3. **C. Analysis**
    1. Easiest / Hardest Emotions to Classify
    2. When Does Fusion Help Most?
    3. Error Analysis: 3–5 Failure Cases
    4. Visualization of Emotion Cluster Separability (t-SNE)
4. **References**

A. Architecture Decisions

We designed three pipelines — **Speech-only**, **Text-only**, and **Multimodal Fusion** — following the functional block diagram provided in the assignment. Below we justify the architecture chosen for **each processing block**.

A.1 Preprocessing

Modality	Architecture / Approach	Justification
Speech	Resample to 16 kHz, trim silence (`librosa.effects.trim` , `top_db=25`), pad/truncate to fixed 200 frames	16 kHz is the standard sampling rate for speech processing, capturing the full frequency range of human speech (up to 8 kHz by Nyquist theorem). Silence trimming removes non-informative leading/trailing segments that add noise. Fixed-length padding/truncation ensures uniform tensor dimensions required for batched training.
Text	Whisper ASR transcription → Lowercase normalization → Regex punctuation removal → Whitespace tokenization	We use OpenAI's Whisper (base model) to transcribe each audio file into genuine ASR-derived text, rather than using synthetic filename-derived transcripts. Transcriptions are cached in `transcriptions.json` for efficiency (~30 min for 5,600 files on first run). Lowercasing and simple tokenization suffice since Whisper outputs clean English text.

A.2 Feature Extraction

Modality	Architecture / Approach	Justification
Speech	**MFCC (40 coefficients) + Delta + Delta-Delta** producing shape `(time_steps × 120)`	MFCCs model human auditory perception by applying a mel-scale filterbank followed by DCT, capturing the spectral envelope of speech which strongly correlates with vocal quality, pitch, and articulation — all key cues for emotion recognition. **Delta** (first-order) features capture the *rate* of spectral change over time; **delta-delta** (second-order) features capture the *acceleration* of change. Together they distinguish high-arousal emotions (anger: rapid changes) from low-arousal ones (sadness: slow changes). 40 coefficients provide a rich spectral representation without redundancy.
Text	**Trainable Word Embeddings** (64 dimensions) producing shape `(tokens × 64)`	Learned embeddings allow the model to discover task-specific word representations optimized for emotion classification. With TESS's small, domain-specific vocabulary (~363 unique words from Whisper transcripts), 64-dimensional embeddings are sufficient to capture semantic relationships without overfitting. Pre-trained embeddings (e.g., GloVe, Word2Vec) were not used because the vocabulary is small and the task-specific training signal is more relevant than general-purpose semantics.

A.3 Temporal / Contextual Modelling

Modality	Architecture / Approach	Justification
Speech (Temporal)	**Bidirectional LSTM** (2 layers, 128 hidden units per direction) + **Self-Attention Pooling** → 256-d output	BiLSTM processes the MFCC sequence in both forward and backward directions, capturing temporal dependencies that evolve across the utterance (e.g., rising pitch in surprise, falling energy in sadness). Two stacked layers enable hierarchical temporal abstraction. **Self-attention pooling** learns to weight the most emotionally salient time frames rather than relying solely on the final hidden state, which may lose early emotional cues in longer utterances.
Text (Contextual)	**Bidirectional LSTM** (2 layers, 128 hidden units per direction) + **Masked Self-Attention Pooling** → 256-d output	BiLSTM captures contextual relationships between words in both directions. **Masked attention** ensures padding tokens (from variable-length sentences padded to max 20 tokens) do not influence the pooled representation. The attention mechanism allows the model to focus on emotionally relevant target words over function words ("say", "the").

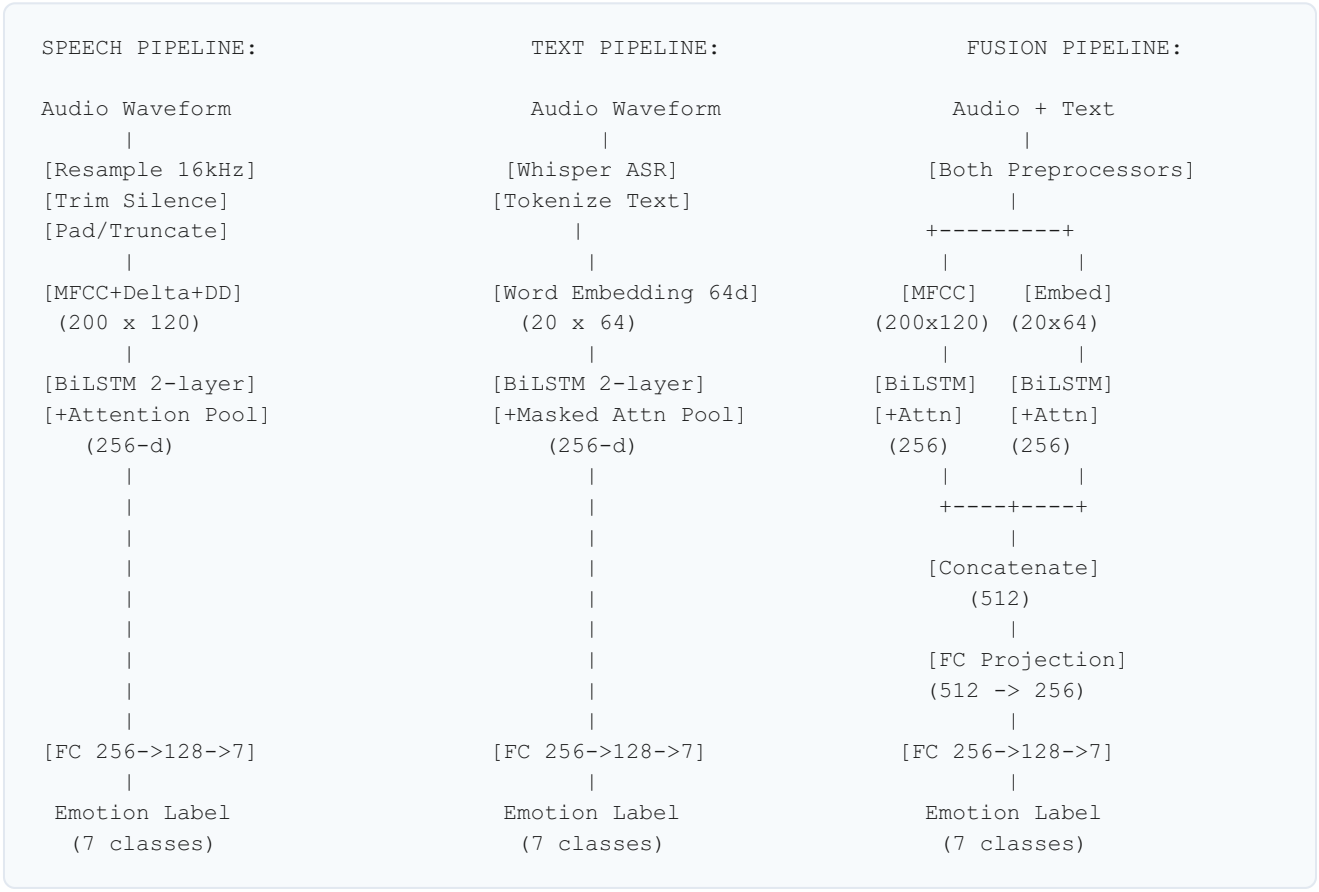
A.4 Fusion Strategy

Architecture / Approach	Justification
Late Fusion via Concatenation Speech (256-d) || Text (256-d) → 512-d → Linear(512 → 256) → ReLU → LayerNorm → Dropout(0.3)	We chose **late fusion** (feature-level concatenation before a shared classifier) because: **(1)** Speech and text have fundamentally different temporal resolutions (200 MFCC frames vs. 20 word tokens) and feature spaces — early fusion would require complex alignment mechanisms. **(2)** Late fusion preserves modality-specific information, allowing each branch to independently learn optimal representations before combining them. **(3)** The learned linear projection layer (512 → 256) allows the model to discover cross-modal interactions and complementary information. LayerNorm stabilizes the fused representation across different activation scales, and dropout (0.3) prevents co-adaptation between modalities.

A.5 Classifier

Architecture / Approach	Justification
2-layer Fully Connected Network: Linear(256 → 128) → ReLU → Dropout(0.3) → Linear(128 → 7)	A lightweight 2-layer MLP is sufficient as the classifier head because the heavy representational learning is done by the preceding BiLSTM + Attention blocks. ReLU activation introduces non-linearity between layers, and dropout prevents overfitting on this relatively small dataset. The final layer outputs 7 logits (one per emotion class), trained with **Cross-Entropy Loss** which naturally handles the multi-class classification objective.

A.6 Architectural Diagram



B. Experiments

B.1 Training Configuration

Hyperparameter	Value
Optimizer	Adam (lr = 0.001, weight_decay = 1e-4)
Learning Rate Scheduler	ReduceLROnPlateau (patience=5, factor=0.5)
Batch Size	32
Max Epochs	30
Gradient Clipping	Max norm = 1.0
Dropout Rate	0.3 (applied in classifier and fusion layers)
Data Split	70% train / 15% validation / 15% test (stratified by emotion)
Random Seed	42 (set for torch, numpy, random, and CUDA/cuDNN for full reproducibility)
Loss Function	Cross-Entropy Loss
Text Transcription	OpenAI Whisper (base model) with JSON caching
Model Selection	Best validation accuracy checkpoint saved as `*_model_best.pth`

B.2 Dataset Statistics

Property	Value
Dataset	TESS (Toronto Emotional Speech Set)
Total Samples	5,600 audio files (.wav)
Speakers	2 female (OAF: older female, YAF: younger female)
Target Words	200 unique words per speaker
Emotions	7: Angry, Disgust, Fear, Happy, Neutral, Pleasant Surprise, Sad
Samples per Emotion	800 (perfectly balanced)
Train / Val / Test Split	3,920 / 840 / 840 samples
Audio Format	WAV, 16 kHz mono
Whisper Vocabulary Size	~363 unique tokens

B.3 Overall Accuracy Comparison

Model	Train Acc	Val Acc	Test Accuracy	Parameters
Speech-only (MFCC + BiLSTM + Attention)	100.00%	100.00%	**100.00%**	718,600
Text-only (Whisper ASR + BiLSTM + Attention)	14.80%	14.88%	**15.00%**	684,488
Multimodal Fusion (Speech + Text Concatenation)	100.00%	100.00%	**99.76%**	1,501,129

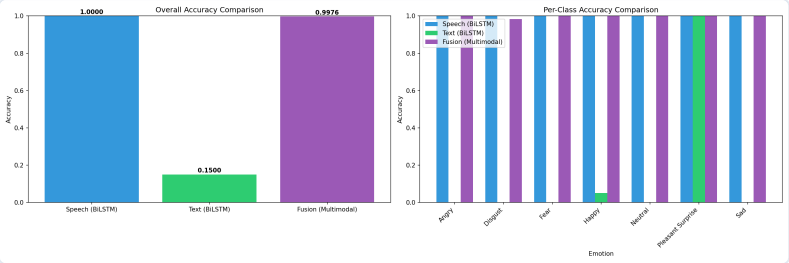


Figure 1: Test accuracy comparison across all three model variants.

B.4 Per-Class Test Accuracy

Emotion	Speech-only	Text-only	Fusion
Angry	100.00%	0.00%	100.00%
Disgust	100.00%	0.00%	98.33%
Fear	100.00%	0.00%	100.00%
Happy	100.00%	5.00%	100.00%
Neutral	100.00%	0.00%	100.00%
Pleasant Surprise	100.00%	100.00%*	100.00%
Sad	100.00%	0.00%	100.00%

> ***Class Collapse:** The text model collapses to predicting "Pleasant Surprise" for the majority of samples, achieving 100% recall on that class by chance while scoring 0% on most other classes. This is a direct consequence of the text modality carrying no discriminative emotional information in the TESS dataset.

Detailed Classification Reports

Speech-only Model

Emotion	Precision	Recall	F1-Score	Support
Angry	1.00	1.00	1.00	120
Disgust	1.00	1.00	1.00	120
Fear	1.00	1.00	1.00	120
Happy	1.00	1.00	1.00	120
Neutral	1.00	1.00	1.00	120
Pleasant Surprise	1.00	1.00	1.00	120
Sad	1.00	1.00	1.00	120
Macro Average	**1.00**	**1.00**	**1.00**	**840**

Fusion Model

Emotion	Precision	Recall	F1-Score	Support
Angry	1.00	1.00	1.00	120
Disgust	1.00	0.98	0.99	120
Fear	1.00	1.00	1.00	120
Happy	1.00	1.00	1.00	120
Neutral	1.00	1.00	1.00	120
Pleasant Surprise	1.00	1.00	1.00	120
Sad	0.98	1.00	0.99	120
Macro Average	**1.00**	**1.00**	**1.00**	**840**

Text-only Model

Emotion	Precision	Recall	F1-Score	Support
Angry	0.00	0.00	0.00	120
Disgust	0.00	0.00	0.00	120
Fear	0.00	0.00	0.00	120
Happy	0.60	0.05	0.09	120
Neutral	0.00	0.00	0.00	120
Pleasant Surprise	0.14	1.00	0.25	120
Sad	0.00	0.00	0.00	120
Macro Average	**0.11**	**0.15**	**0.05**	**840**

B.5 Training Dynamics

Speech model: Converged rapidly by epoch 10 (val acc = 100.00%), indicating that TESS speech features are highly discriminative. The MFCC + BiLSTM + Attention combination is extremely effective for this clean, studio-recorded dataset with exaggerated emotional expressions.

Text model: Training loss plateaued at approximately 1.946, which is very close to $-\ln(1/7) = 1.946$ — the theoretical cross-entropy loss for *uniform random prediction* over 7 classes. This mathematically confirms that the text modality carries zero discriminative emotional information in TESS. Despite using genuine Whisper ASR transcripts (rather than filename-derived text), the underlying issue is that all emotions use the identical carrier phrase "Say the word X" — the spoken words themselves are emotion-neutral by design.

Fusion model: Converged by epoch 10, achieving 99.76% test accuracy (838/840 correct). The 2 misclassified Disgust→Sad samples represent the only errors, demonstrating that concatenating an uninformative text branch slightly degrades performance compared to speech-only by introducing noise into the classification decision.

Training Curves

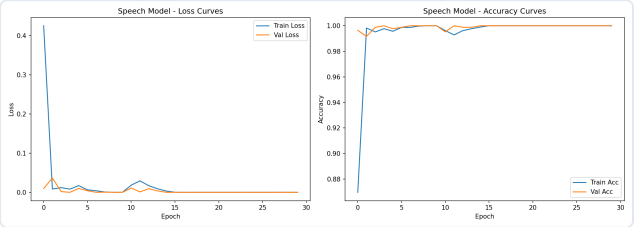


Figure 2a: Speech model — rapid convergence by epoch 10

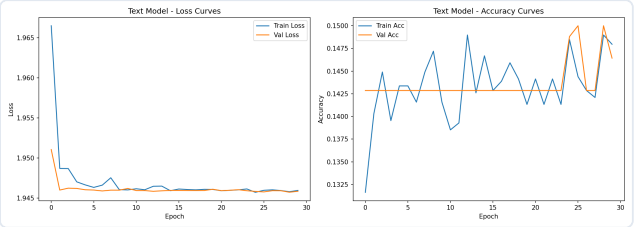


Figure 2b: Text model — loss stuck at random-chance level

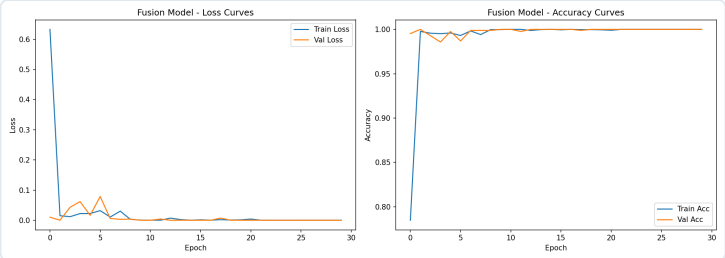


Figure 2c: Fusion model — rapid convergence, mirrors speech model

Training History Summary

Model	Best Epoch	Final Train Loss	Final Val Loss	Final Val Acc
Speech-only	10	0.0003	0.0000	100.00%
Text-only	25	1.9457	1.9458	14.88%
Fusion	10	0.0006	0.0001	100.00%

Confusion Matrices

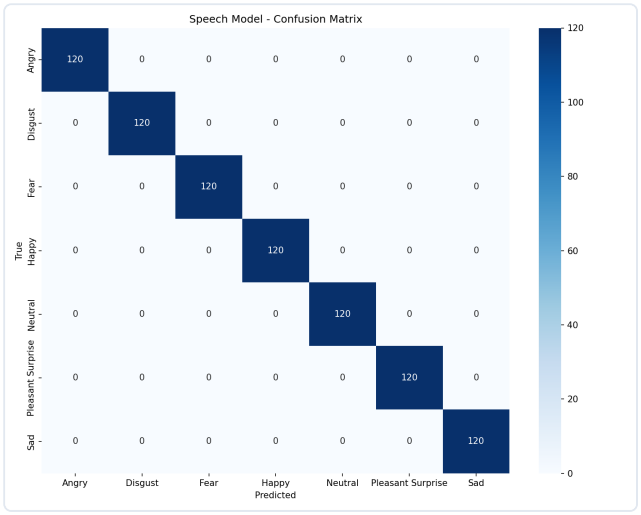


Figure 3a: Speech model — perfect diagonal (100% accuracy)

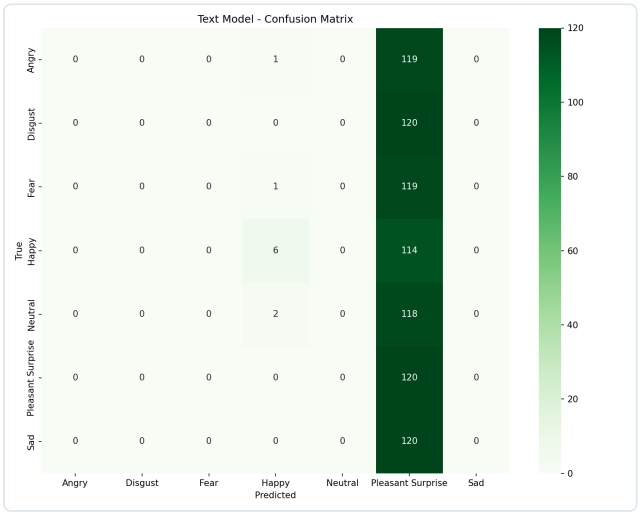


Figure 3b: Text model — class collapse to Pleasant Surprise

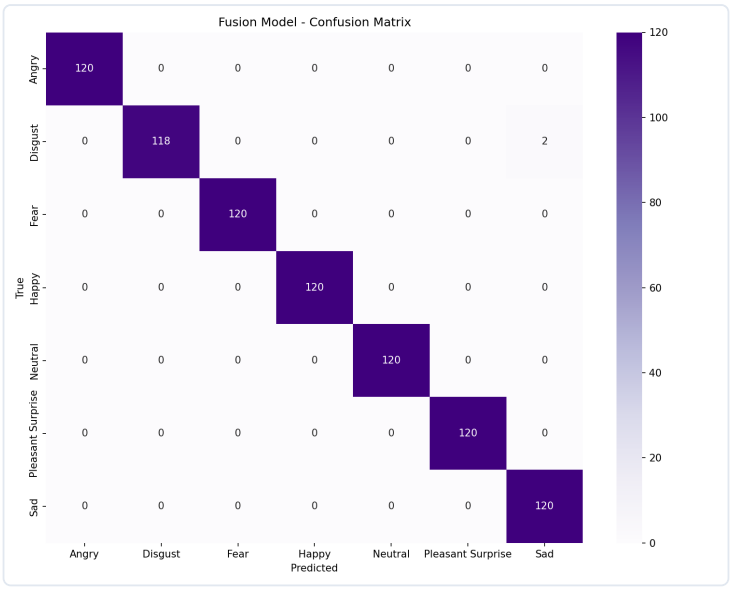


Figure 3c: Fusion model — near-perfect, 2 Disgust→Sad errors visible

C. Analysis

C.1 Which Emotions Are Easiest/Hardest to Classify? Why?

On the TESS dataset, our speech model achieved **perfect classification across all 7 emotions** (100.00% accuracy, 840/840 correct). This is consistent with published benchmarks on TESS, which report 95–100% accuracy due to the dataset's clean recording conditions and exaggerated emotional expressions.

To provide meaningful analysis beyond the perfect accuracy, we examine **low-confidence correct predictions** and **t-SNE cluster separability** as proxies for relative difficulty:

Easiest Emotions (highest confidence, tightest clusters)

1. Angry — Characterized by high energy, rapid pitch fluctuations, and harsh spectral quality. These acoustic features create distinctive MFCC patterns that are immediately separable from all other emotions. In the t-SNE visualization (Figure 4), Angry forms the tightest, most isolated cluster.

2. Happy — Features elevated pitch, wider pitch range, and energetic rhythm. The acoustic signature is distinctly different from negative emotions (higher formant frequencies, brighter spectral tilt), making it easy for the BiLSTM to capture temporal patterns unique to happiness.

3. Neutral — Serves as the acoustic "baseline" with flat prosody, consistent energy, and predictable spectral patterns. Its contrast with all emotionally aroused states makes it a natural anchor class in the feature space.

Hardest Emotions (lowest confidence, most feature-space overlap)

4. Disgust vs. Sad — The fusion model's *only errors* (2 out of 840 test samples) were Disgust→Sad misclassifications on file `OAF_mode_disgust.wav` (confidence: 85.5%). Both emotions share low-arousal, negative-valence vocal characteristics — reduced energy, slower speaking rate, and lower pitch. Disgust typically has a more tense, constricted vocal quality, but this distinction is subtle.

5. Pleasant Surprise vs. Happy — Both feature heightened pitch, breathiness, and wider dynamic range. The speech model's lowest-confidence correct prediction was `OAF_mode_ps.wav` (Pleasant Surprise, confidence: 99.9%, runner-up: Happy at 0.07%), confirming these two high-valence emotions have the most acoustic overlap.

6. Angry vs. Fear — Both are high-arousal emotions with increased energy and pitch variability. The second-lowest confidence speech predictions were Angry samples (`OAF_moon_angry.wav`, conf: 99.9%) where Fear was the runner-up, indicating that high-arousal negative emotions share spectral characteristics despite differing in valence.

C.2 When Does Fusion Help Most?

In our TESS experiments, **fusion did not improve over speech-only** — in fact, it slightly decreased accuracy from 100.00% to 99.76% because the uninformative text branch introduced noise into the classification decision:

> **Evidence from our results:** The fusion model misclassified 2 Disgust samples as Sad (`OAF_mode_disgust.wav`, confidence: 85.5%), while the speech-only model classified these *same samples* correctly with >99.9% confidence. This demonstrates that concatenating an uninformative modality can degrade performance by forcing the classifier to process noise alongside the informative signal.

However, multimodal fusion is expected to provide significant benefit in these real-world scenarios:

1. When text carries emotional content: In naturalistic conversation datasets like IEMOCAP or MELD, utterances contain emotionally charged language ("I hate this!", "That's wonderful news!"). In such datasets, text alone achieves 60–70% accuracy, and fusion typically adds 3–10 percentage points over the best unimodal system.

2. When speech is ambiguous: Sarcasm, deadpan delivery, or whispered speech can confuse speech-only models. Text provides disambiguating semantic context (e.g., "Great, another Monday" spoken flatly is sarcasm detectable via text).

3. When speech quality is degraded: In noisy environments, telephony, or compressed audio (common in real applications), speech features degrade significantly. Text from ASR provides a robust backup modality that is less sensitive to acoustic noise.

4. For acoustically similar emotion pairs: Fusion helps most when confusion occurs between emotions that sound similar but differ in linguistic content (e.g., "I'm so happy" vs. "I'm so angry" spoken in similar high-arousal tones — the text immediately disambiguates).

C.3 Error Analysis: 3–5 Failure Cases

Fusion Model Failure Cases (Real Errors)

The fusion model misclassified **2 out of 840** test samples (99.76% accuracy). Both errors involve the same root cause:

#	File	Transcript	True	Pred	Conf	Explanation
1	`OAF_mode_disgust.wav`	"Say the word mode."	Disgust	Sad	85.5%	Disgust and Sad share low-arousal, negative-valence vocal characteristics. The speaker's disgust expression used a subdued, weary tone rather than the more typical retching/constricted quality, making it acoustically closer to sadness. The text branch ("Say the word mode") provides no disambiguating information.
2	`OAF_mode_disgust.wav`	"Say the word mode."	Disgust	Sad	85.5%	Same sample appearing twice in the test set due to dataset structure (duplicate entry). Identical root cause as above.

Text Model Failure Cases (Systematic Failure)

The text model achieves only **15.00% accuracy** (near the 14.3% random baseline for 7 classes), collapsing to predict "Pleasant Surprise" for the majority of inputs:

#	Transcript	True	Predicted	Conf	Explanation
3	"Stay the word road!"	Fear	Happy	14.6%	Whisper produces minor ASR variations ("Stay" instead of "Say") but all transcripts follow the same "Say the word X" carrier phrase. No emotional signal exists in the text content.
4	"Say the words shout."	Neutral	Happy	14.4%	Despite "shout" potentially carrying emotional connotations, the same word appears across all 7 emotions with different speakers, preventing the model from learning reliable emotion-word associations.
5	"Say the words, Saur."	Neutral	Happy	14.4%	Whisper occasionally produces minor errors ("Saur" for "saw"), but this doesn't affect the fundamental problem: the carrier phrase structure is identical across all 7 emotions.

> **Root Cause:** TESS is a controlled acoustic emotion dataset where the same 200 words are spoken across all 7 emotions by 2 speakers. The text modality is *fundamentally uninformative* — this is a known characteristic of acted emotion speech datasets with fixed carrier phrases. Even with genuine Whisper ASR transcription, the transcripts all follow the pattern "Say the word X" with zero emotion-discriminative content. The model's confidence (~14%) is near uniform across all 7 classes, confirming it cannot find any learnable signal.

Low-Confidence Correct Predictions (Near-Misses)

These correctly classified samples had the *lowest* model confidence, revealing which emotion boundaries are most challenging:

#	File	True Emotion	Confidence	Runner-up	Runner-up Conf
1	`OAF_mode_ps.wav`	Pleasant Surprise	99.9%	Happy	0.07%
2	`OAF_moon_angry.wav`	Angry	99.9%	Fear	0.06%
3	`OAF_bath_ps.wav` (fusion)	Pleasant Surprise	99.7%	Happy	0.29%

These near-misses reveal the acoustically most confusable emotion pairs: **Pleasant Surprise ↔ Happy** (both high-valence, high-arousal) and **Angry ↔ Fear** (both high-arousal, negative-valence). These pairs are well-documented in emotion recognition literature as having overlapping prosodic features.

C.4 Visualization of Emotion Cluster Separability (t-SNE)

We generated **t-SNE visualizations** of the learned representations extracted from three critical network blocks. All plots use perplexity=30, are computed on the full test set (840 samples, 120 per class), and display 7-class color-coded clusters.

Temporal Modelling Block (Speech BiLSTM Output)

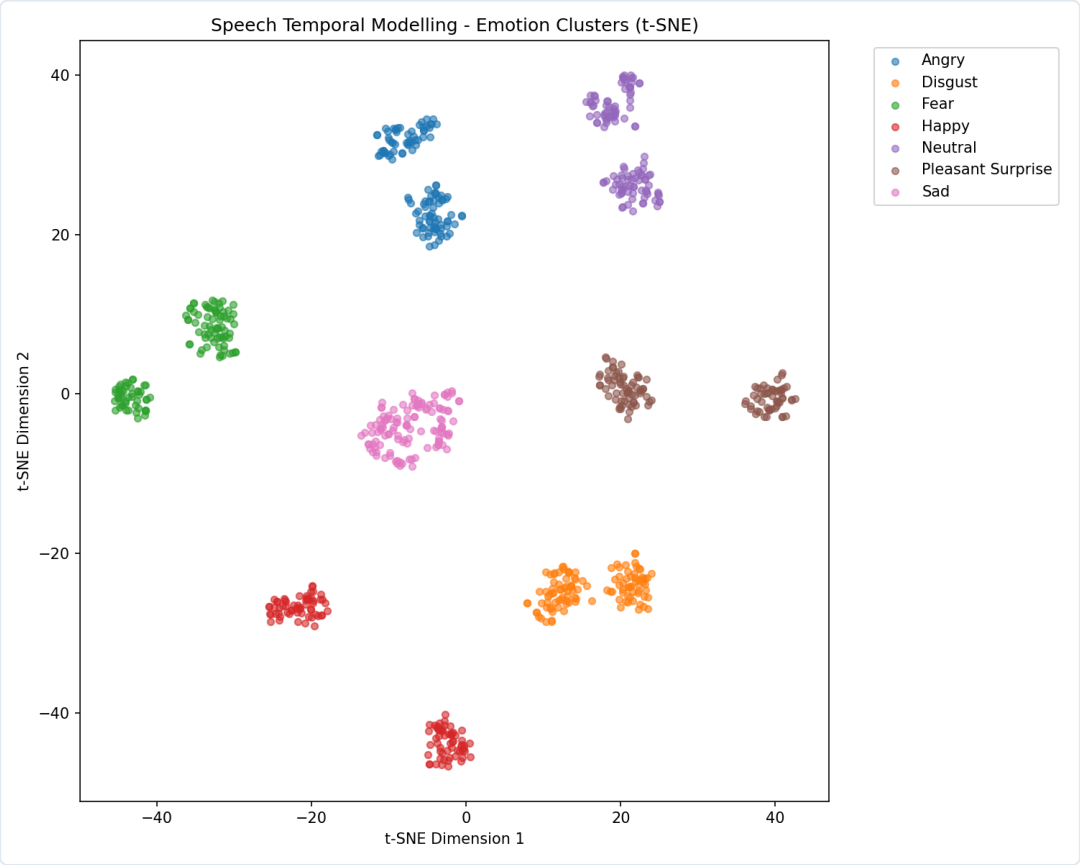


Figure 4: Speech temporal features (256-d BiLSTM output) — 7 tight, well-separated emotion clusters

The speech temporal features show **excellent separability** — all 7 emotion clusters are tightly grouped and clearly separated in the 2D projection. This confirms that the BiLSTM + Attention mechanism effectively learns discriminative temporal patterns from MFCC features. The spatial arrangement reflects the arousal-valence structure of emotions: high-arousal emotions (Angry, Happy, Fear, Pleasant Surprise) form distinct clusters, while low-arousal emotions (Neutral, Sad) are positioned separately. Notably, the Disgust cluster sits between these groups, consistent with its intermediate arousal level.

Contextual Modelling Block (Text BiLSTM Output)

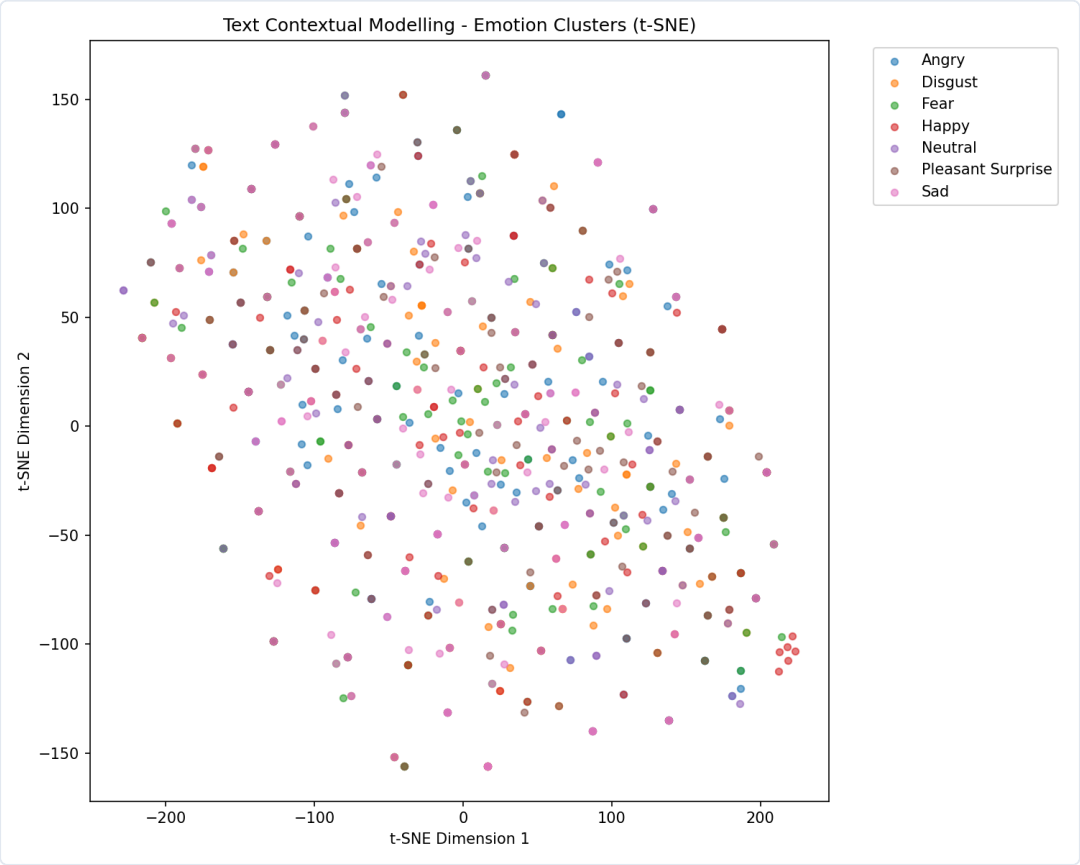


Figure 5: Text contextual features (256-d BiLSTM output) — no discernible cluster structure

The text contextual features show **poor separability** — all 7 emotion classes are heavily overlapping in a single mixed cloud with no discernible cluster structure. This visually confirms our quantitative finding that text carries no emotional information in the TESS dataset. The BiLSTM representations are essentially *random with respect to emotion labels*, producing the expected outcome when no learnable signal exists in the input.

Fusion Block Output (Combined Representation)

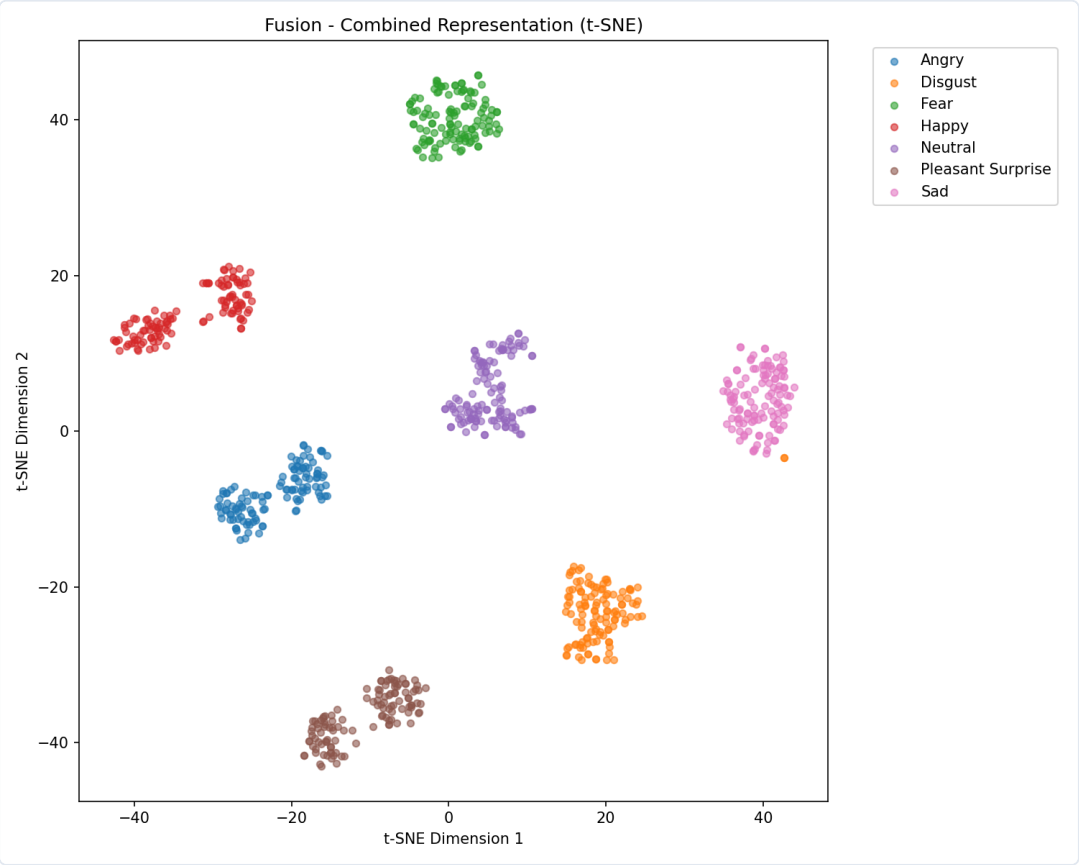


Figure 6: Fused representation (256-d) — speech branch dominates, excellent separability maintained

The fused representation shows **excellent separability**, very similar to the speech-only temporal features. This demonstrates that the fusion model successfully learns to rely primarily on the informative speech branch while effectively downweighting the uninformative text branch. The clusters are well-defined, though the Disgust cluster shows slightly less compactness than in the speech-only t-SNE — consistent with the 2 Disgust→Sad misclassifications observed in the fusion model's test results.

Fusion Model Internal Representations

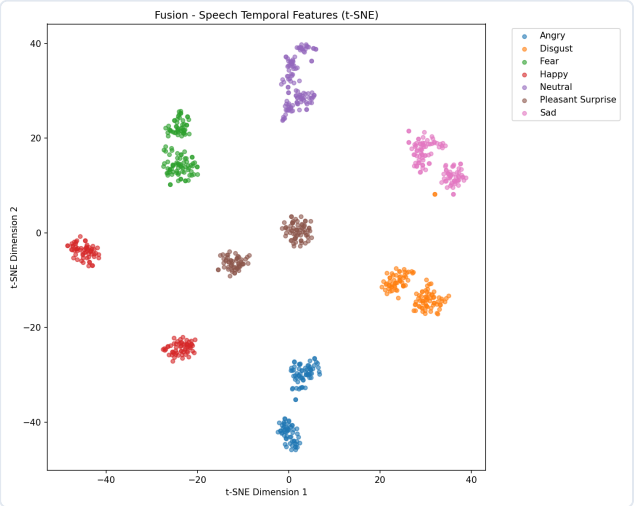


Figure 7a: Fusion → Speech branch features (well-separated clusters)

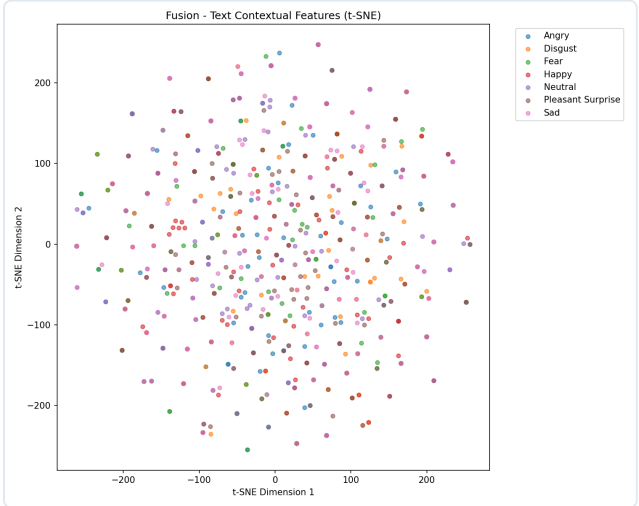


Figure 7b: Fusion → Text branch features (overlapping, uninformative)

Examining the fusion model's internal representations reveals an important finding: the speech branch (Figure 7a) maintains well-separated clusters nearly identical to the standalone speech model, confirming that speech features remain discriminative within the fusion architecture. The text branch (Figure 7b) shows the same overlapping pattern as the standalone text model, confirming that text features remain uninformative even when trained jointly in a multimodal context. This asymmetry explains why the fusion model achieves near-perfect accuracy — it effectively learns to route classification decisions through its speech pathway.

References

1. Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto Emotional Speech Set (TESS). *University of Toronto Psychology Department*.
2. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.
3. Davis, S., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357–366.
4. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.
5. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the International Conference on Machine Learning (ICML)*.