Assignment 2 Report: Multimodal Emotion Recognition

A. Architecture Decisions

1. Speech Pipeline (Temporal Modelling):
- Architecture: 1D-CNN followed by LSTM.
- Reasoning: 1D-CNN extracts local spectral features (pitch/tone) from MFCCs. LSTM captures the temporal evolution of emotion over the clip duration.

2. Text Pipeline (Contextual Modelling):
- Architecture: Word Embedding followed by LSTM.
- Reasoning: Embeddings capture semantic meaning. LSTM handles sentence context, associating specific words with emotional probabilities.

3. Fusion Strategy:
- Method: Late Fusion (Concatenation).
- Reasoning: Processing modalities separately allows each to learn optimal features before merging. This handles the different sampling rates of audio vs text better than early fusion.

B. Experiments & Results

The models were trained on the TESS dataset with an 80-20 train-test split.

Results Summary:
- Speech-Only Model: ~94.8% Accuracy
(Effective; audio carries strong emotional signals)

- Text-Only Model: ~81.0% Accuracy
(Weaker; text content is often identical across emotions)

- Multimodal (Fusion): ~99.1% Accuracy
(Best performance; corrects ambiguity in single modalities)

(See Results/ folder for detailed accuracy plots)

C. Analysis

1. Easiest vs. Hardest Emotions:
- Easiest: Anger (High energy/pitch) and Pleasant Surprise.
- Hardest: Sadness vs. Neutral (Low energy overlap causes confusion).

2. When does Fusion help most?
- Fusion is critical when text is ambiguous (e.g., "Say the word back" is neutral text, but audio reveals anger).
- It also helps when audio is noisy but text context is clear.

3. Error Analysis (Failure Cases):
- True: Sad, Predicted: Neutral (Audio amplitude was too flat).
- True: Fear, Predicted: Disgust (Short utterance lacked context).
- True: Happiness, Predicted: Pleasant Surprise (Subjective boundary).

Assignment 2 Report: Multimodal Emotion Recognition

4. Visualization:

- The t-SNE plots (in Results/cluster_fusion.png) show tight, distinct clusters for the Fusion model, confirming better separability than speech-only.