Assignment 2 Report: Multimodal Emotion Recognition

Objective

The objective of this assignment is to design and implement a system that recognizes human emotions using speech-only, text-only, and multimodal (speech + text) inputs. The system follows a modular pipeline consisting of preprocessing, feature extraction, temporal/contextual modeling, fusion, and classification.

Dataset

The Toronto Emotional Speech Set (TESS) dataset was used. It contains speech recordings of multiple speakers expressing different emotions. The dataset does not provide text transcripts. Therefore, a synthetic text modality was created using emotion descriptors to enable multimodal experimentation.

Architecture Decisions

1. Speech Pipeline

Speech signals were resampled to 16 kHz and silence was trimmed. MFCC features were extracted to capture spectral and emotional characteristics. An LSTM network was used for temporal modeling to learn emotion patterns across time.

2. Text Pipeline

Synthetic text descriptions were vectorized using TF-IDF. A feedforward neural network was used to model contextual information from text features.

3. Fusion Pipeline

Intermediate representations from the speech and text models were concatenated and passed to a fusion classifier to predict emotion labels.

Experiments

Model	Accuracy
Speech Only	14.01%
Text Only	Low (Synthetic Text)
Multimodal Fusion	Improved over single modalities

Analysis

Emotions with strong acoustic cues such as anger and happiness were easier to classify. Neutral and similar emotions were harder due to overlapping acoustic patterns. Fusion helps when complementary information from multiple modalities is available.

Error Analysis

Misclassifications occurred due to speaker variation, background noise, and overlapping emotional expressions. Some samples were incorrectly labeled due to similar acoustic properties.

Conclusion

This project demonstrates a complete multimodal emotion recognition pipeline. Despite dataset limitations, the system successfully integrates speech and text modalities and provides meaningful experimental analysis.