Multimodal Speech Emotion Recognition using Audio and Text Fusion

Abstract

This project implements a multimodal emotion recognition system that predicts human emotions from speech using both acoustic and textual information. Audio features such as MFCC, pitch, and energy capture prosodic emotional cues, while text embeddings provide semantic context. Three models were developed: speech-only, text-only, and fusion. Experiments on the TESS dataset show that speech features are the dominant indicator of emotion. The speech-only model achieved 90.71% accuracy, while the text-only model performed at chance level due to identical sentence content across emotions. The fusion model achieved 67.86% accuracy. The results demonstrate that multimodal learning improves performance only when both modalities carry meaningful emotional information.

Architecture Decisions

Speech Pipeline: Feed-forward neural network for summarized acoustic descriptors.

Text Pipeline: Dense classifier on DistilBERT embeddings.

Fusion Pipeline: Concatenation of audio and text representations followed by classifier.

Experiments

Dataset: Toronto Emotional Speech Set (TESS)

Speech Only Accuracy — 90.71%

Text Only Accuracy — 13.57%

Fusion Accuracy — 67.86%

Analysis

Angry and Happy are easiest due to strong pitch and energy variations.

Neutral and Sad are hardest due to similar acoustic patterns.

Text model fails because sentences remain identical across emotions.

Fusion reduces performance due to noisy modality mixing.

Conclusion

Speech carries primary emotional information in TESS dataset. Multimodal learning is useful only when modalities contain meaningful signals.