Multimodal Speech Emotion Recognition

Final Project for the Course Natural Language Processing - Theory and Applications

By Shrey Chhabra

1. Introduction

Emotion recognition from human speech is a fundamental challenge in affective computing and has wide-ranging applications in human-computer interaction, virtual assistants, healthcare diagnostics, and more. While many approaches use either acoustic features or linguistic content in isolation, this project explores a multimodal approach, combining both audio signals and their textual transcriptions to improve emotion classification accuracy.

The novelty of this work lies in its fusion of real-time transcription from OpenAI's Whisper model with deep learning models trained on both audio and text inputs, followed by late fusion of their independent emotion predictions. This dual-channel processing provides robustness to noisy speech, ambiguity in tone, or lexically neutral phrases.

2. Project Objectives

- Build an emotion classifier using audio features extracted from speech.
- Build a text-based emotion classifier using pre-trained transformers.
- Use OpenAI Whisper for real-time speech-to-text conversion.
- Perform late fusion on audio and text outputs to make a final emotion prediction.
- Enable an interactive interface for users to record speech and receive emotion feedback.

3. Datasets and Label Mapping

3.1 CREMA-D Dataset (Audio Supervision)

The CREMA-D dataset (Crowd-sourced Emotional Multimodal Actors Dataset) contains over 7,400 recordings from 91 actors portraying six basic emotions: Anger, Disgust, Fear, Happy, Neutral, Sad. Audio files were processed into Mel spectrograms and used to train a CNN classifier.

3.2 GoEmotions Model (Text Supervision)

The text classifier uses a pre-trained transformer model `j-hartmann/emotion-english-distilroberta-base`, based on the GoEmotions dataset. GoEmotions provides 28 fine-grained emotion categories. Since CREMA-D uses only six, a mapping was created:

GoEmotions → CREMA-D:
- anger, annoyance → Anger

- disgust → Disgust
- fear, nervousness → Fear
- joy, excitement → Happy
- sadness, grief → Sad
- neutral → Neutral

4. System Architecture

The system is composed of four main modules:

4.1 Speech Transcription: Whisper
OpenAI's Whisper model is used to transcribe raw `.wav` or `.m4a` audio. It is robust to background noise, accents, and poor enunciation. Whisper supports automatic language detection and outputs highly accurate text from speech.

4.2 Audio-Based Emotion Classifier (CNN)
Audio is converted to a Mel spectrogram with 256 mel bands, resized to 256×256, and normalized. The CNN model consists of Conv2D + MaxPool layers followed by Dense layers and a softmax output.

4.3 Text-Based Emotion Classifier (Transformer)
A DistilRoBERTa transformer model outputs a probability distribution over GoEmotions, which are mapped to CREMA-D categories and normalized using softmax.

4.4 Late Fusion
A linear weighted average is applied: Final = α * Audio + $(1 - \alpha)$ * Text. Default $\alpha = 0.5$ for equal contribution. Final prediction is based on the maximum probability class.

5. Interactive UI: Record and Predict

A notebook-based UI using `ipywidgets` was implemented. It allows users to:
1. Record 5 seconds of live audio
2. Transcribe it using Whisper
3. Predict emotions from audio and text
4. Fuse predictions for final output

6. Novelty and Contributions

This project introduces several novel elements:
- Multimodal late fusion of deep audio and text classifiers
- Use of Whisper for robust speech-to-text in noisy conditions
- Practical interface for live recording and emotion prediction
- CREMA-D alignment with GoEmotions using a custom label mapping

Unlike unimodal systems, this pipeline avoids blind spots of each modality—text classifiers can't detect tone, while audio-only models miss semantic clues.

7. Future Work

- Explore early/intermediate fusion models
- Expand to multilingual emotion recognition
- Deploy as a mobile/web app
- Include speaker traits for personalized emotion detection