Speech Emotion Recognition

Yahia Ibrahim AlKaranshawy, Hossam Osama Iraqi, Ali Hassan Ali

May 9, 2025

# Problem statement

It’s required to Create the Feature Space in the time domain, or in the frequency domain and Convert the audio waveform to a mel spectrogram and compare the performance of the CNN models in the two approaches using [Speech Emotion Recognition (en)](https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en)[.](https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction)

## Dataset Loading and visualize

The CREMA-D dataset contains audio files labeled with six emotion classes: **Angry, Disgust, Fear, Happy, Neutral, and Sad**. Each filename follows a specific format that encodes the emotion (e.g., 1001\_DFA\_ANG\_XX.wav where ANG = Angry).

**Key Steps in Data Loading:**

1. **File Parsing**
   * Scan the dataset directory for .wav files.
   * Extract emotion labels from filenames using the embedded emotion codes (e.g., ANG, DIS).
2. **Audio Loading**
   * Use librosa.load() to read audio files at a standard sampling rate (e.g., 22050 Hz).
   * Ensure uniform duration by padding/trimming signals to 3 seconds (66150 samples).
3. **Emotion Mapping**
   * Convert emotion codes (ANG, DIS, etc.) to full names (Angry, Disgust, etc.).

A screenshot of a computer screen

AI-generated content may be incorrect.

## Dataset Splitting

* 1. Split the data into 70% training and validation and 30% testing.
* 2. Use 5% of the training and validation data for validation.

# MEL