Speech Emotion Recognition

~ Utsav

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects the underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion.

Dataset

- **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)**
  - https://paperswithcode.com/dataset/ravdess
  - Paper ⇒https://zenodo.org/records/1188976#.YFZuJ0j7SL8
  - Readily Available Data ⇒ https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio (holds only portion of all data mentioned in the paper)
    - This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.
- CREMA-D ⇒ Crowd Sourced Emotional Multimodal Actors Dataset
  - https://github.com/CheyneyComputerScience/CREMA-D
  - CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74

coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified).

    - Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral and Sad) and four different emotion levels (Low, Medium, High and Unspecified).

- [ ] Explore other dataset for Over-engineering Use case 🙂

Dataset Viz:

We are utilizing RAVDESS Dataset for the experiments further

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB) only 1440 files are used in our case

The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong),

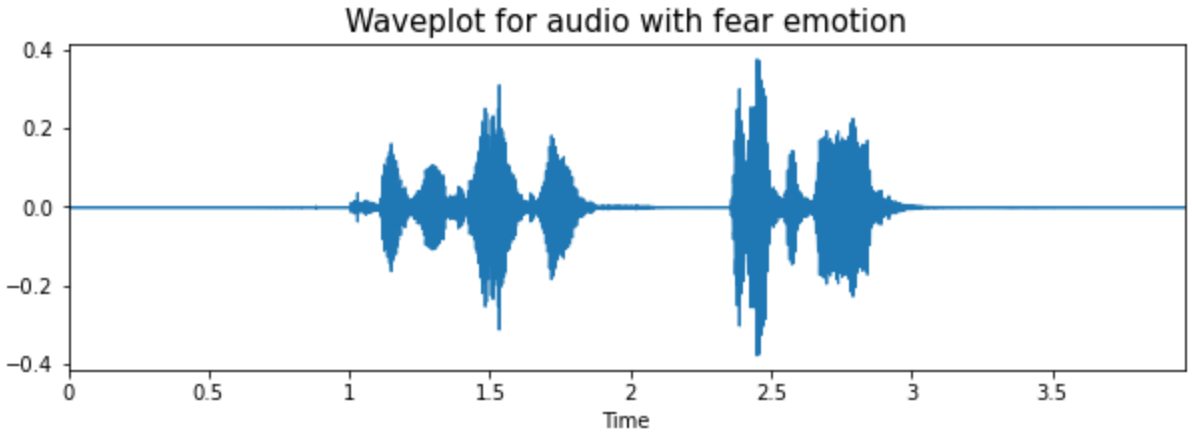
The filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-12.mp4). These identifiers define the stimulus characteristics:

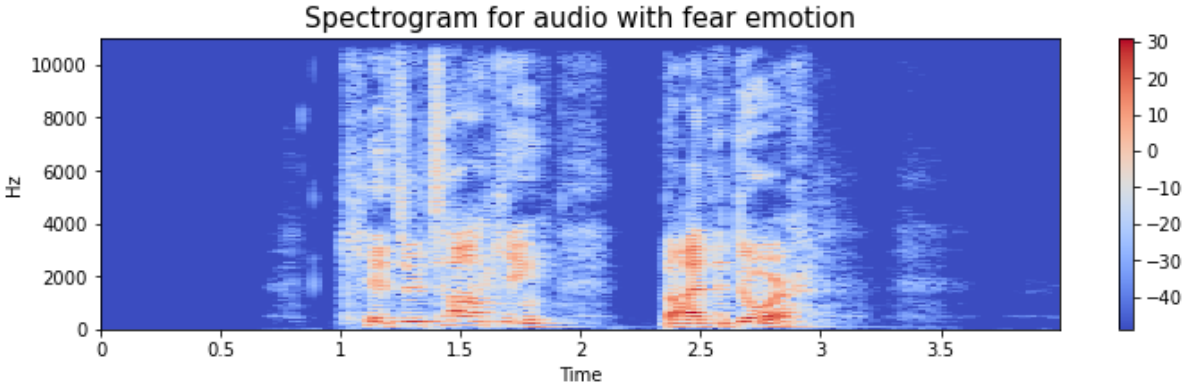
Filename identifiers

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).

- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

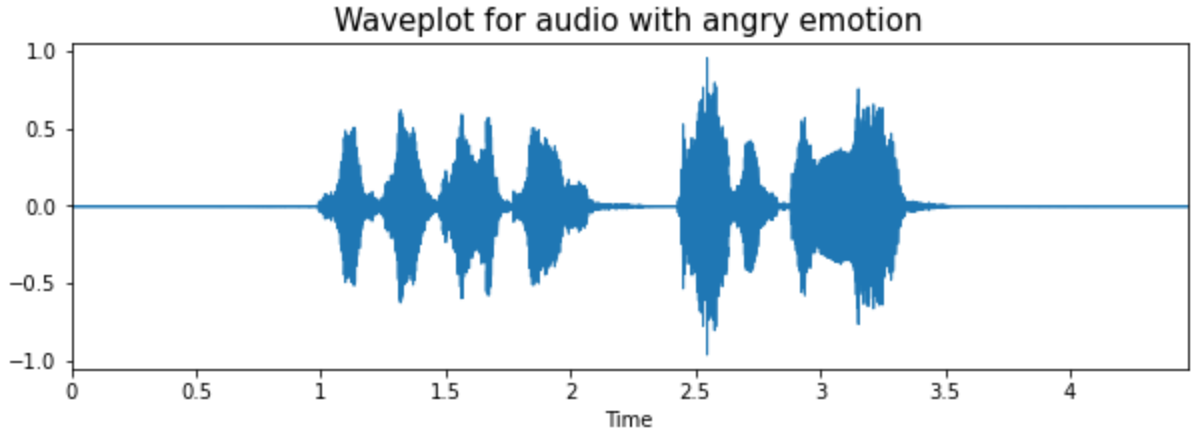
Fear

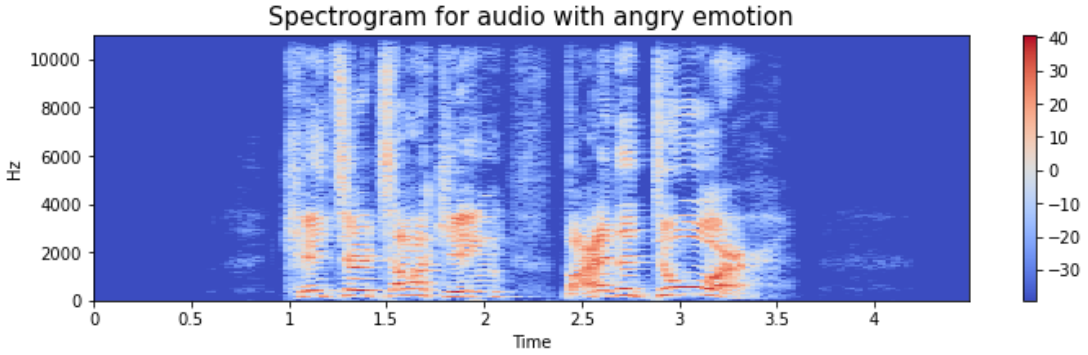




https://prod-files-secure.s3.us-west-2.amazonaws.com/4cc20139-7260-4713-8add-ece438cde12e/f9382944-6d4d-4281-8cd0-ed1aa50d9c51/03-01-06-02-01-02-02.wav

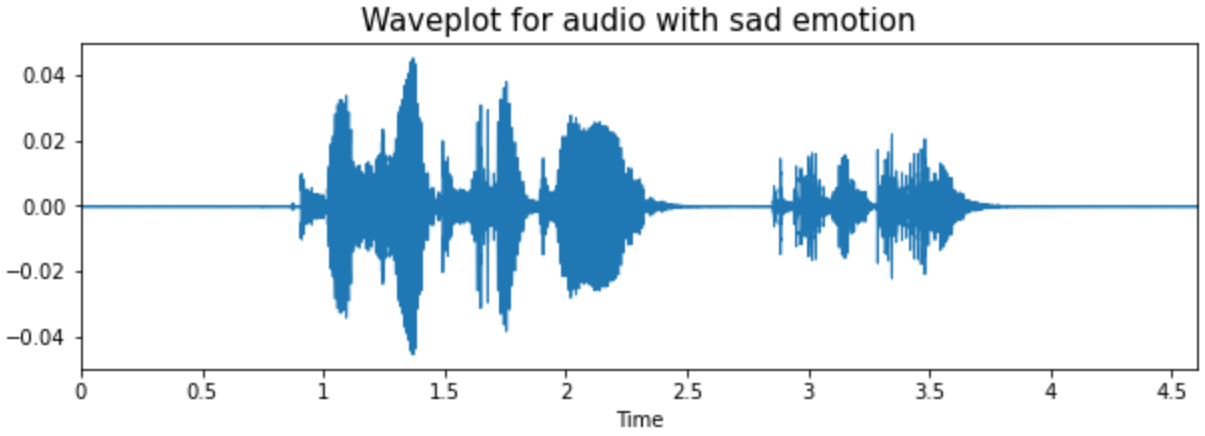
Angry

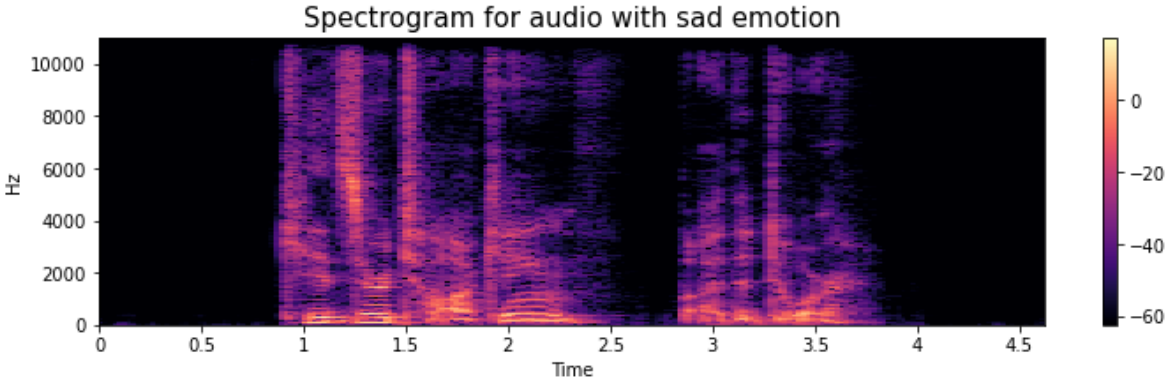




https://prod-files-secure.s3.us-west-2.amazonaws.com/4cc20139-7260-4713-8add-ece438cde12e/d5e15300-3f89-4419-b42c-01ee082c56eb/03-01-05-02-01-01-02.wav

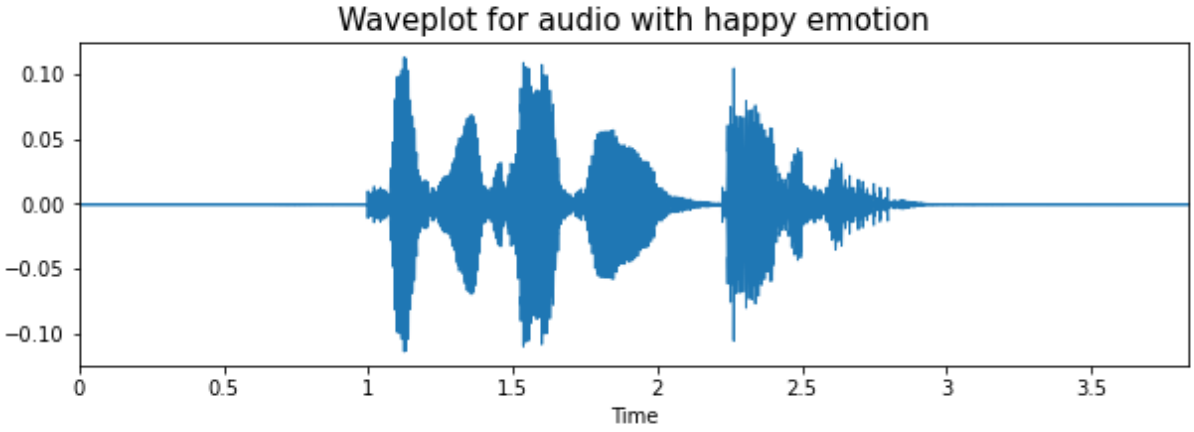
Sad

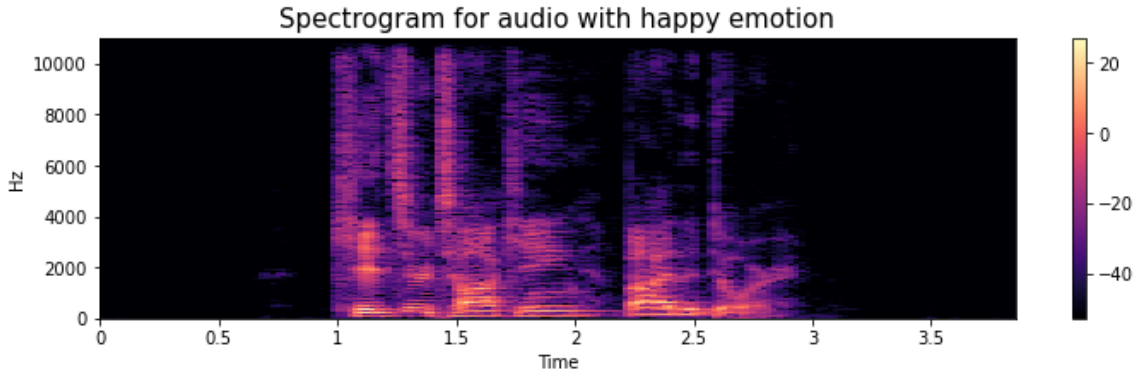




https://prod-files-secure.s3.us-west-2.amazonaws.com/4cc20139-7260-4713-8add-ece438cde12e/221bc263-dc5b-4052-93b3-44229f11551c/03-01-03-01-01-02-02.wav

Happy





https://prod-files-secure.s3.us-west-2.amazonaws.com/4cc20139-7260-4713-8add-ece438cde12e/a2a90efe-e191-4191-b14c-6824e15e1600/03-01-04-02-02-02-02.wav

My Approach

Feature Extraction

https://medium.com/heuristics/audio-signal-feature-extraction-and-clustering-935319d2225

Librosa library was used for feature extraction

https://librosa.org/doc/main/feature.html#spectral-features

https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d

Here we have utilized hand crafted features :

1. Zero Crossing Rate : The rate of sign-changes of the signal during the duration of a particular frame.
2. Chroma_stft : chromagram from a waveform or power spectrogram
3. MFCCs Mel-frequency cepstral coefficients (MFCCs)
4. RMS(root mean square) value: root-mean-square (RMS) value for each frame, either from the audio samples
5. MelSpectogram: mel-scaled spectrogram

> STFT ⇒ Short-time Fourier Transform

Model

Keras & tensorflow framework was utilized here

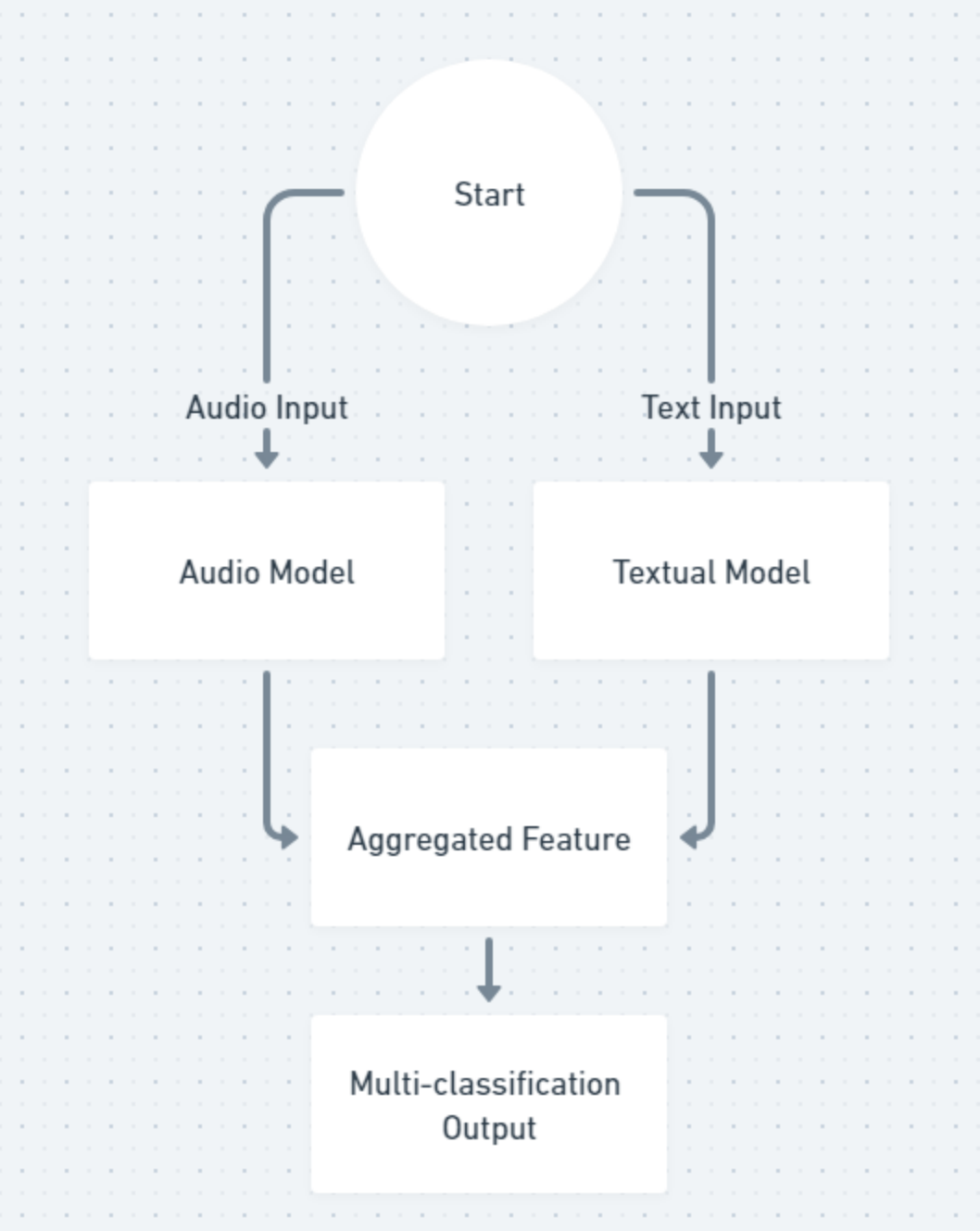
- LSTM Model
  - 5 Layer LSTM with 128 units each followed by a dropout layer of 0.2 dropout rate
- CNN Model
  - VGG16
  - Resnet101
  - EfficientNetB3
  - MobileNetV2

Over-Engineering (Overkill 🙂)

We only 2 textual prompts here, so lets over-engineer to build a multi-modal model

- We only have 2 text inputs
  - "Kids are talking by the door",
  - "Dogs are sitting by the door"
- but to experiment with multi-modal architecture here

  - [ ] Identify & experiment on Some other dataset



Pytorch & HuggingFace Library was used here

Pretrained model was utilized to generate the required features

Audio Model

- Whisper

    - WhisperProcessor
- Wav2Vec2
    - Wav2Vec2Processor
- Wav2Vec2-BERT
    - Wav2Vec2BertProcessor

Text Model

- BERT Base Uncased
- Google T5-small

Feature Aggregation & Classifier

Here i have used 2 sets of approachs

- ML
    - XGBoost
    - CatBoost
- DL
    - Feed Forward Neural Networks
    - Attention Mechanisms
        - Self Attention (Scaled Dot-Product Attention)
        - Cross Modal Attention (pending)
    - Simple Transformer