Project Milestone: Emotional Vocalization Classification

Team Members: Nithya Srinivasan (nsrinivasan@berkeley.edu), Kris Mehra (kris.mehra@berkeley.edu), Bjorn Melin (bjorn_melin@berkeley.edu), Pratheek Sankeshi (psankesh9@berkeley.edu)

Course: DS207 - F25S8 **Date:** 10/12/2025

Motivation:

Emotions are a critical part of human communication, and machines have not been able to successfully detect them automatically. Accurate emotion detection has the potential to enhance human-computer interaction, which will be beneficial in several applications such as online therapy, customer service systems, and accessibility tools. Our team is especially motivated by the connection between emotion detection and text-to-speech (TTS) research. By tackling emotion detection in both speech and song, we aim to build foundational skills and insights that directly support future research on emotionally aware TTS systems.

Data Description:

We use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. RAVDESS contains 24 speech and song recordings in eight different emotions (neutral, calm, happy, sad, angry, fearful, surprise, and disgust). Each clip is labeled with emotion, intensity, and modality. This dataset is widely used in emotion recognition research and provides a balanced and well-curated benchmark for evaluating speech-based emotion detection models.

Data Preprocessing:

The initial step involved programmatically parsing the filenames of the 1,440 speech files to extract metadata. Each filename, such as 03-01-08-02-02-01-12.wav, was broken down into its seven numeric identifiers. This process allowed for the automated labeling of each audio sample with its corresponding Emotion, Emotional Intensity, Statement, and Actor ID. These extracted labels are the target variables for the model.

To ensure consistency across all files, a standard audio processing pipeline was applied.

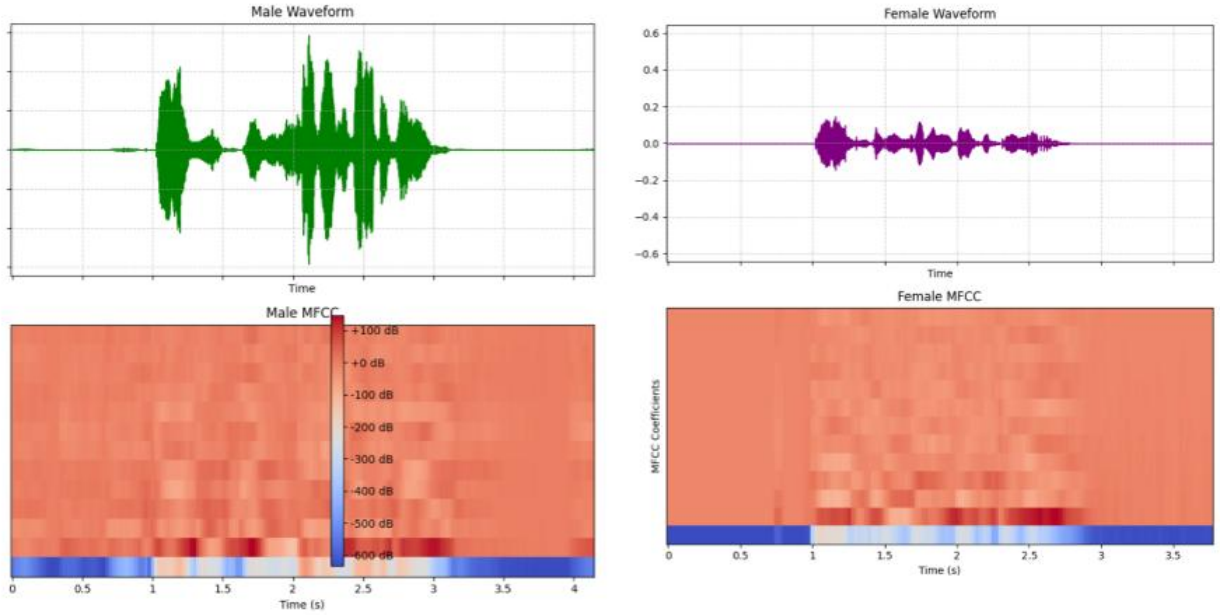
- Resampling: All audio files were down-sampled from 48kHz sampling rate to 22kHz.
- Silence Trimming: Leading and trailing silences were removed from each clip.
- Normalization: Audio amplitude for all files was normalized to a peak value of 1.0.

Feature Extraction:

The processed waveforms were then converted into numerical features for machine learning. From each audio file, the following primary acoustic features were extracted:

- MFCCs: 40 Mel-Frequency Cepstral Coefficients were extracted to capture the timbral and spectral characteristics of the voice.
- Mel Spectrogram: A log-mel spectrogram was generated to provide a 2D representation of how the spectral content of the speech evolves over time.
- Chroma Features: 12 chroma features were extracted to represent the tonal content of the speech.

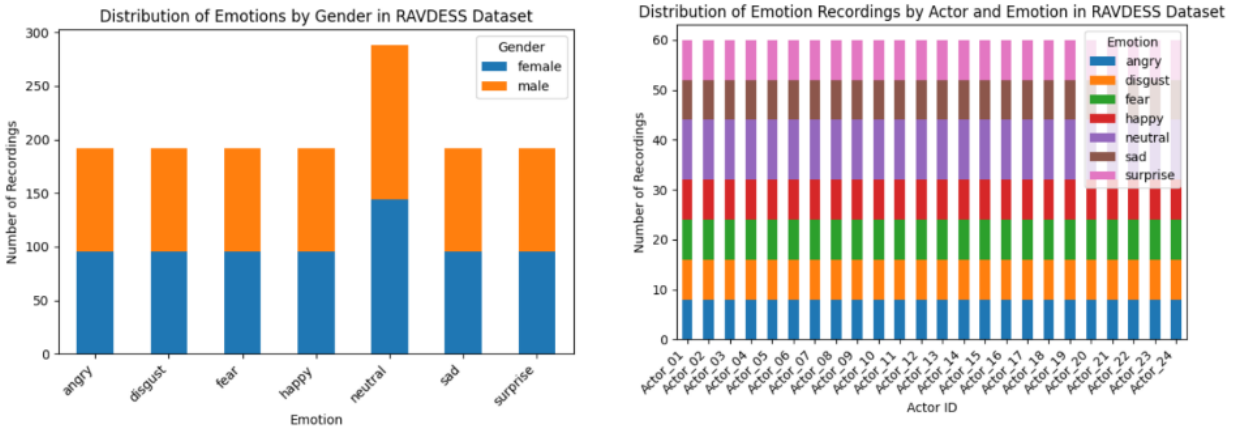
Below is a comparison of a male versus female recording and MFCC plots.



EDA:

Dataset Balance and Distribution

An initial analysis of the dataset's structure showed its balanced design containing an equal number of samples from the **12 male and 12 female actors**. Also, the distribution of emotional categories is perfectly balanced, with **192 samples for each of the seven emotions** with the exception of 'neutral' which has 96 samples. This balance ensures that a machine learning model will not be biased toward any particular emotion during training.



Acoustic Feature Analysis Across Emotions:

Exploring the extracted acoustic features revealed clear and consistent patterns across different emotional categories. Visualizations confirmed that these acoustic differences provide a strong quantitative basis for distinguishing between the emotional classes.

Impact of Intensity and Gender:

Recordings labeled as **'strong' intensity** consistently exhibited more extreme feature values than their **'normal' intensity** counterparts for the same emotion. Additionally, gender-based differences were observed as expected. While predictable, confirming this pattern is crucial for building a model that can generalize across different speakers

Data Challenges:
While we did not experience substantive challenges in obtaining, loading, processing and analyzing data in this phase of the project, we anticipate that the RAVDESS dataset will present several challenges that can limit model performance and generalization in next steps of our project. Given that the dataset is relatively small, with only 24 actors and varying numbers of samples per emotion, which can lead to overfitting and bias toward more common emotions like *neutral* or *happy*. The dataset also contains acted rather than spontaneous emotions, meaning the emotional expressions are often exaggerated and may not reflect real-world affective speech. Moreover, because all recordings come from a small group of speakers in clean, studio conditions, models risk learning speaker-specific or recording-specific features rather than true emotional cues, leading to poor performance on unseen voices or noisy data.

Methods and Experiments:
The next phase of the project will focus on model development and experimentation. The plan is to implement and compare the following models to identify the most effective approach:

- **Support Vector Machine (SVM)** will be developed first to serve as a baseline.
- **XGBoost** model will be built to assess its power as an ensemble method.
- **Convolutional Neural Network (CNN)** will be created with 2D Mel Spectrograms that will allow the model to learn features automatically from the time-frequency data, providing a deep learning alternative to the other methods.

To ensure a fair comparison, all models will be trained and evaluated using a standardized framework. The pre-defined speaker-independent splits will be used, with the 80% training set for model fitting and the 10% validation set for hyperparameter tuning and remining 10% as test set. While overall **accuracy** will be the primary metric, class-specific metrics will also be analyzed and a confusion matrix will be generated to visually inspect various emotions.

Data augmentation:
The best-performing model will be retrained on a dataset expanded with artificial noise and pitch variations to measure whether this improves its robustness and generalization capabilities. We also plan to include Songs datasets as well as additional relevant datasets to test the robustness of our final model selection.

Team Contributions:
We all collaborated on all aspects of this milestone and we will continue to share responsibilities based on team members availability, expertise and results-based team spirit. For this phase, Nithya worked on overall approach and data prep, Kris contributed to EDA and drafting the document, Bjorn helped with model selections and next steps and Pratheek assisted with data challenges and final review of the document.

Project GitHub Repository:

https://github.com/psankesh9/EDFS