Project Milestone Three

The choice of the project is emotional speech recognition, which there was some suspicion to convert the speech to text. However, upon further digging, I discovered that it appears that it is unnecessary for my project. The files of the RAVDESS data have unique file names that consist of seven-part numerical identifiers that define the characteristics of the file: Modality, Vocal Channel, Emotion, Emotional intensity, Statement Type, Repetition, and the Actor. For example, the filename: 03-01-01-01-01-01-11.wav

Here is a breakdown of what the numbers in the name mean:

03 - this file is audio-only

01 - this file has speech, not song

01 - the emotion is neutral

01 - the emotion intensity is normal

01 - the statement is "Kids are talking by the door."

01 - first repetition of the statement

11 - the odd number indicates the actor is male. Even numbers are female.

Thus, no conversion is needed. A visual representation is required in order to further decide the features to extract. Visualizing the waveform in speech is important in feature extraction because looking at the variations of amplitude can show the emotional intensity and stress levels in speech. Higher amplitudes may indicate increased emotional stimulation. The same goes for the energy distribution; it can signal emotional variation; when there are sudden energy bursts, it may represent emotional outbursts (Huzaifah, 2017). Waveforms also reveal temporal aspects of speech, such as the speaking rate (pace of speech fast-slow), pauses, and rhythmic patterns (). This offers insight into the emotional state of the speaker (Schafer, 2007). After reviewing the spectrogram and wave plots, the features to be extracted are the mfcc, chroma, and Mel. Mfcc is the mel frequency cepstral coefficient (founded on the Short-Time Fourier Transform); it represents the short-term power spectrum of a sound. Chroma is the twelve different pitch classes. Mel is the mel spectrogram frequency. The Short-Time Fourier Transform is a fundamental concept in digital speech processing; it describes the frequency components in a signal-averaged and characterizes signals with time-varying frequency components (Carnegie, 2023).

Resources:

Carnegie Mellon University (2023). Short-Time Fourier Transforms. CMU.EDU. <https://course.ece.cmu.edu/~ece491/lectures/L25/STFT_Notes_ADSP.pdf>

Rabiner L., Schafer R. (2007). Introduction to Digital Speech Processing. Foundations and Trends® in Signal Processing: Vol. 1: No. 1–2, pp 1-194

Huzaifah M. (2017). Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks. Arxiv.org. <https://arxiv.org/pdf/1706.07156>