Homework 1: Speech Analysis

2.Write a python script to extract the following features

For Speaking rate calculation:

Speaking rate was approximated as the number of words per second (#words / duration).

For my own recordings, each utterance contained 12 words. `My neighbor bought a new car last week because his old one stopped working.`

For the MSP samples, the number of words was manually counted as follows:

{19, 13, 17, 31, 16, 26, 9} for {Happy, Angry, Sad, Afraid, Surprised, Disgusted, Neutral}.

The same method (#words/duration) was used for both datasets to ensure consistency.

Duration was obtained using sound.get_total_duration() from Parselmouth.

4. Describe the characteristics of the two sets of emotional speech

Emotion	Your speech	Podcast speech
Happy	The happy speech sounds bright, with a relatively high mean pitch (above 230 Hz) and a wide pitch range, indicating a high-pitched emotion and a cheerful, energetic tone. The intensity is strong and the voice is loud, while the speaking rate is moderate, suggesting active expression and increased vocal tension. The HNR is the highest among all emotions, indicating a clear and resonant voice.	The happy samples exhibit a relatively high mean pitch (~207 Hz) and a wide pitch range. This shows an energetic, high-arousal vocal tone. The mean intensity and large dynamic variation indicate strong, expressive delivery. Speaking rate is the second fastest among all emotions (~4.3 w/s), and HNR is high and Shimmer is the smallest, suggesting clear, resonant, and emotionally positive speech.
Angry	The angry speech shows the highest minimum pitch and a broad pitch range (146–466 Hz), indicating a consistently sharp tone. It also has the highest intensity and the intensity varies most significantly, suggesting loudness and sustained excitation. The speaking rate is relatively fast, reflecting agitation, while jitter and shimmer are low, showing stable vocal fold vibration and a clean voice quality. High HNR (second highest) indicates stable expression of emotions.	Angry speech shows the widest pitch range (93–510 Hz) and high intensity peaks, indicating a forceful and tense vocal style. Pitch variability and large intensity spread further suggest strong emphasis. Although jitter (highest) and shimmer are high, HNR is the lowest among the high-energy emotions (~7.8 dB), showing a rough, strained voice quality consistent with anger.

Sad	The sad speech exhibits the lowest minimum and maximum pitch values, indicating the lowest overall pitch and a deep, subdued tone. Max intensity and min intensity are the weakest, indicating a soft voice with low energy. The speaking rate is slightly slower, and both jitter and shimmer are higher, suggesting a rough or breathy quality. The HNR is moderate, implying a mix of stability and breathiness in the voice.	Sad speech has a narrower pitch range (87–269 Hz) and a moderately low mean pitch. Both minimum and maximum intensity values are low, showing soft and low-energy vocalization. The slowest speaking rate (≈ 2.3 w/s), lowest Jitter, and highest HNR (~15.6 dB) suggest calm but subdued emotional expression with a clear, stable tone.
Afraid	The afraid speech has a low minimum pitch but the highest maximum pitch, with a large pitch standard deviation, indicating great variation and emotional fluctuation. It also shows a large intensity range (low minimum, high maximum) and the highest speaking rate, suggesting agitation and nervous excitement. Jitter and shimmer are both relatively high (second largest), implying vocal instability and tension. A small HNR indicates tension, roughness, and a lot of breathiness.	The afraid samples has smallest min pitch, and display strong pitch variability (SD ≈ 69 Hz compared to avg 48 Hz) and the largest overall pitch range (74–440 Hz). This reflects instability and nervous tension. Although intensity varies widely (1.8–85.7 dB), the mean intensity is moderate. High jitter and the highest shimmer indicate tremor and vocal strain. The low HNR suggests a breathy, unstable voice typical of fear.
Surprised	The surprised speech shows pronounced pitch variation, with the highest mean pitch and the largest pitch standard deviation, indicating a highly excited tone. Intensity also varies substantially, and the speaking rate is the fastest (~3 w/s). Moderate jitter and minimal shimmer indicate clarity, and a high HNR (second highest), indicating excitement and a clear voice quality.	The surprised speech has a high mean pitch and the largest pitch SD. This indicates strong pitch modulation and excitement. Intensity is relatively high (~75 dB compared to avg 72 dB) with notable variability, and the speaking rate is fast. Moderate jitter and shimmer, and with a mid-range HNR (~12.5 dB compared to avg 11.5 dB), show an animated and expressive but slightly tense tone.

Disgusted	The disgusted speech has a moderately low pitch with a low minimum, giving a heavy or subdued impression. Its intensity is relatively low, with the smallest intensity standard deviation, which means little loudness variation and a flat prosodic contour. HNR is low, accompanied by the highest jitter and shimmer values, revealing a rough and unstable voice quality that reflects aversion or repulsion.	Disgusted speech shows a mid-low pitch (~162 Hz compared to avg 192 Hz) , smaller pitch range, and lowest intensity variation. This suggests a restrained delivery and reduced prosodic fluctuation. The Speaking rate is the fastest. And the medium high shimmer and low HNR imply a coarse, unstable voice quality that conveys aversion or repulsion.
Neutral	The neutral speech is acoustically balanced, with the lowest mean pitch (~154 Hz) and moderate intensity. The speaking rate is slow, with low jitter, medium shimmer, and relatively high HNR. This indicates stable, smooth vocal production without marked emotional variation.	Neutral speech displays the lowest max and mean pitch and the smallest pitch SD, showing monotonic, steady vocalization. Intensity levels are moderate with little variation (SD ≈ 10.6 dB compared to avg 12 dB). The Speaking rate is a little slower. Moderate jitter and shimmer with mid-range HNR (~11.2 dB) indicate a stable, balanced, and emotionally unmarked tone.

5. Answer the following questions in several sentences each. Remember to briefly justify each of your answers.

a. What are some similarities and differences between the features from the two datasets?
Similarities:
Both datasets show consistent emotional patterns across acoustic features. For example, happy and surprised speech in both sets shows higher mean pitch, larger pitch variability, and faster speaking rates, reflecting high-arousal, energetic emotions. Conversely, sad and neutral emotions in both datasets show lower pitch, smaller intensity variation, slower speaking rate, and higher HNR, indicating calmer speech. These parallel trends suggest that the pitch, loudness, and rate convey emotion reliably across speakers and recording conditions.

Differences:
Due to speaker variability and environmental noise, the absolute feature values vary significantly between datasets. MSP samples collected from different speakers exhibit a wider range of pitch and intensity, as well as more variation in speaking rate, indicating greater acoustic diversity. In contrast, my recordings are more consistent and stable, with a nearly

uniform speaking rate, because all utterances were produced by a single speaker in a single setting. Therefore, while the relative emotional patterns are similar, differences between speakers and recording conditions lead to significant differences.

b. Which of the datasets would be more useful for emotion recognition applications? Why?

For emotion recognition applications, I think the MSP dataset would be more useful overall. It contains a large variety of speakers, recording conditions, and emotional expressions, which provides a richer and more diverse feature for training and evaluating models. Such diversity helps a model learn speaker-independent emotional patterns rather than overfitting to one speaker's voice characteristics.

c. Which of these datasets would be easier for an emotion recognition system to classify? Why?

I think my own recordings would be easier for an emotion recognition system to classify. All samples were produced by me in a quiet, controlled environment, leading to consistent pitch, intensity, and voice quality across recordings. The emotional distinctions are therefore clear and systematic, and the model is easilier to detect feature differences such as higher pitch and faster rate in *happy* speech versus lower pitch and slower rate in *sad* speech. In contrast, the MSP Podcast dataset includes many speakers, microphones, and recording conditions, making the model hard to separate emotional characteristics from speaker or environment specific effects.

d. What other features would be useful for emotion recognition? Why?

First, I think Spectral features like formant frequencies or other contour patterns would be useful. These features reflect differences in articulation, which often vary across emotions — for example, tense or bright vowels in *happy* speech versus more relaxed articulation in *sad* speech.

Second, prosodic and temporal features such as pause duration, hesitation, or changes in speaking rate could also help capture emotional states. Such rhythm-related patterns can reveal how speakers express excitement, uncertainty, or calmness beyond static acoustic measures.