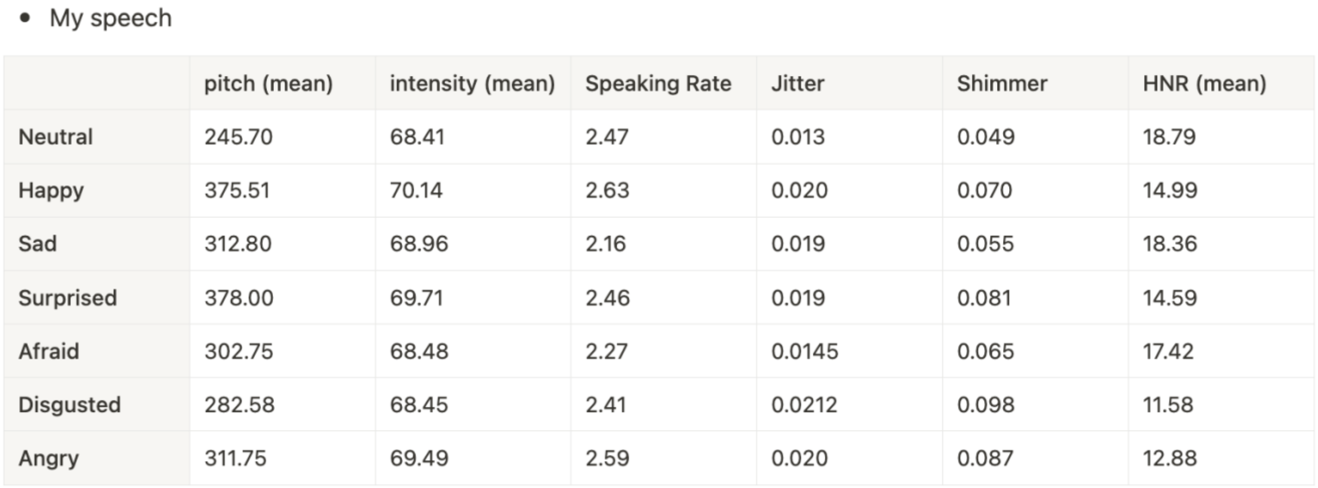
I confirm that I have not used any GPT-generated responses for any part of this assignment.

Recording Transcript: “Oh my gosh, I can’t believe this happened. What am I supposed to do now?”

Speaking Rate Calculation Method: To calculate speaking rates, I first needed to obtain the word count from the transcript. As shown in the *get_transcript()* function, I used the *speech_recognition* module to convert speech to text by applying *recognize_google(audio)*. Since background noise can interfere with accurate transcription, I applied noise reduction using *adjust_for_ambient_noise(source, duration=0.3)* before extracting the transcript. Once the transcript was obtained, I counted the number of words and divided it by the total duration of the speech to approximate the speaking rate.

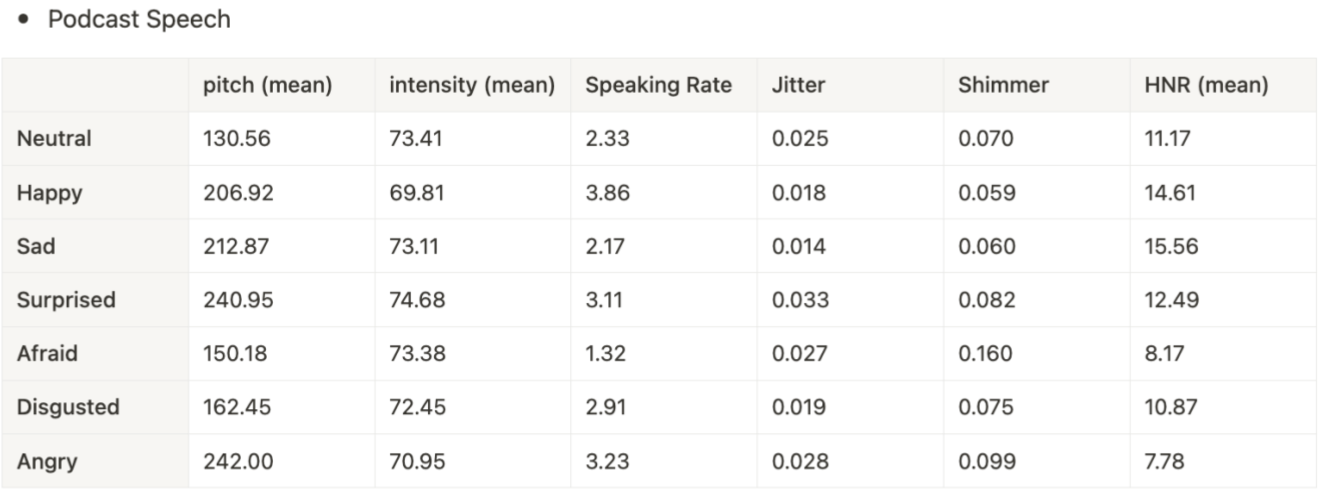
4. Describe the characteristics of the two sets of emotional speech, for each emotion within each dataset. (e.g. In podcast speech, happy v.s. other emotions Do NOT compare across datasets). At least 2-3 sentences in each cell.

Note: The MSP podcast samples are from a variety of speakers under different recording conditions. Because of this, direct comparisons of exact feature value are not useful; instead consider comparisons of relative feature value (e.g. compared to each speaker’s mean).



- My speech

	pitch (mean)	intensity (mean)	Speaking Rate	Jitter	Shimmer	HNR (mean)
Neutral	245.70	68.41	2.47	0.013	0.049	18.79
Happy	375.51	70.14	2.63	0.020	0.070	14.99
Sad	312.80	68.96	2.16	0.019	0.055	18.36
Surprised	378.00	69.71	2.46	0.019	0.081	14.59
Afraid	302.75	68.48	2.27	0.0145	0.065	17.42
Disgusted	282.58	68.45	2.41	0.0212	0.098	11.58
Angry	311.75	69.49	2.59	0.020	0.087	12.88



- Podcast Speech

	pitch (mean)	intensity (mean)	Speaking Rate	Jitter	Shimmer	HNR (mean)
Neutral	130.56	73.41	2.33	0.025	0.070	11.17
Happy	206.92	69.81	3.86	0.018	0.059	14.61
Sad	212.87	73.11	2.17	0.014	0.060	15.56
Surprised	240.95	74.68	3.11	0.033	0.082	12.49
Afraid	150.18	73.38	1.32	0.027	0.160	8.17
Disgusted	162.45	72.45	2.91	0.019	0.075	10.87
Angry	242.00	70.95	3.23	0.028	0.099	7.78

The above table presents the mean values of the extracted speech features from my *feature_extraction.py* script (included here as a reference for easier comparison). In accordance with the provided instruction, I have structured my analysis by comparing **mean values** within each dataset separately rather than directly across datasets.

Emotion	**Your Speech**	**Podcast Speech**
Happy	Happy speech is rapid, pitched, and lively, and thus the most expressive of all the emotions. Specifically, it has a high mean pitch of about 375.51 Hz, slightly lower than Surprised (387.00 Hz) but greater than any other emotion. Its intensity, at 70.14 dB, is the highest, being louder than Surprised (69.71 dB) and Angry (69.49 dB). Furthermore, it has the fastest speaking rate at 2.63 words/sec, the most vivid tone. Jitter (0.020) and shimmer (0.070) are moderate, higher than Neutral (0.013) and Afraid (0.015) but lower than Disgusted (0.021), indicating some variation but not so roughness. The HNR (14.99 dB) is relatively lower than Neutral (18.79 dB) and Sad (18.36 dB), which means it is a bit more noisy.	Happy speech is energetic, fast-paced, and lively, contributing to its expressive nature. Compared to the speaker's typical speech patterns, Happy speech generally features a higher pitch and faster speaking rate. For example, the mean pitch, which is 206.92 Hz, is relatively high, though still lower than Surprised (240.95 Hz), Angry (242.00 Hz), and Sad (212.87 Hz). Its intensity of 69.81 dB is one of the lower values, softer than Neutral (73.41 dB), Surprised (74.68 dB), and even Sad (73,11 Hz). The speaking rate is. He fastest among all emotions, suggesting a dynamic and engaging delivery. In terms of stability, jitter (0.018) and shimmer (0.059) are relatively low, ensuring a clear and resonant voice. The HNR, which is 14.61 dB, is relatively high, higher than Neutral (11.17 dB) but slightly lower than Sad (15.56dB), indicating a fairly clear signal with some background noise.
Angry	Angry speech is loud, high-pitched, and fast, making it intense and forceful. Specifically, it has a high mean pitch, which is 311.75 Hz, slightly lower than Sad (312.80 Hz) and significantly lower than Happy (375.51 Hz) and Surprised (378.00 Hz). Its intensity, which is 69.49 dB mean, is also high, just slightly lower than Happy (70.14 dB) and Surprised (69.71 dB). It has the second-fastest speaking rate at 2.59 words/sec, after Happy (2.63 words/sec). In terms of stability, jitter (0.020) and shimmer (0.087) are relatively high, suggesting a more unstable and aggressive voice. The HNR, which is 12.88 dB, is low, indicating a noisier and rougher tone.	Angry speech is characterized by a forceful and intense delivery, reflecting aggression and urgency. It has the highest mean pitch (242.00 Hz) among all emotions, making it sharper. The intensity, which is 70.95 dB, is moderate to low, lower than Neutral, Surprised, Sad, and Afraid, but slightly higher than Happy (69.81 dB). The speaking rate, which is 3.23 words/sec, is also fast, which supports an assertive and urgent delivery. However, jitter (0.028) and shimmer (0.099) are relatively high, indicating a more unstable vocal quality. The HNR (7.78 dB) is the lowest among all emotions, meaning this speech has the most noise and roughest vocal texture.
Sad	Sad speech is slow, steady, and low in intensity, making it sound calm and steady compared to more expressive emotions. Specifically, it has a mean pitch of 312.80 Hz, which is slightly higher than Angry (311.75 Hz) but lower than Happy (375.51 Hz) and Surprised (378.00 Hz). Its intensity, which is 68.96 dB, is moderate, lower than	Sad speech is slow, steady, and lower in intensity compared to other emotions. It has a moderate mean pitch of 212.87 Hz, higher than Neutral (130.56 Hz), Afraid (150.18 Hz), and Disgusted (162.45 Hz) but lower than Surprised (240.95 Hz) and Angry (242.00 Hz). Its intensity, which is 73.11 dB, is relatively strong, very close to Neutral (73.41 dB), and

	Happy (70.14 dB), Surprised (69.71 dB), and Angry (69.49 dB) but slightly higher than Neutral (68.41 dB) and Afraid (68.48 dB). The speaking rate at 2.16 words/sec is the slowest among all emotions. In terms of stability, jitter (0.019) and shimmer (0.055) are relatively low, indicating a steady voice. The HNR, which is 18.36 dB, is high, suggesting a clear and stable vocal quality.	higher than Disgusted (72.45 dB) and Happy (69.81 dB). The speaking rate, which is 2.17 words/sec, is one of the slowest, reflecting a melancholic tone. Jitter (0.014) and shimmer (0.060) are relatively low, keeping the voice stable and controlled. The HNR, which is 15.56 dB, is the highest, suggesting a clean and less noisy vocal tone.
Afraid	Afraid speech has moderate pitch and slow speed, sounding hesitant and unstable. Specifically, it has a moderate pitch (302.75 Hz), lower than Happy (375.51 Hz), Surprised (378.00 Hz), and Angry (311.75 Hz) but higher than Neutral (245.70 Hz) and Disgusted (282.58 Hz). It has one of the lowest intensities (68.48 dB), similar to Neutral (68.41 dB) and Disgusted (68.45 dB). The speaking rate, which is 2.27 words/sec, is moderate, slower than Happy and Angry but faster than Sad, indicating hesitant and unstable speech. In terms of stability, shimmer (0.065) is moderate, indicating mild voice perturbation. The HNR, which is 17.42 dB, is relatively high, meaning it has a clear signal.	Afraid speech conveys hesitation and unease, characterized by a moderate pitch and a slow pace. Its mean pitch (150.18 Hz) is lower than Surprised (240.95 Hz), Happy (206.92 Hz), and Sad (212.87 Hz) but higher than Neutral (130.56 Hz). The intensity, which is 73.38 dB, is relatively high, close to Neutral (73.41 dB) and sad (73.11dB). The speaking rate, which is 1.32 words/sec, is the slowest, emphasizing uncertainty and hesitation in speech delivery. The shimmer (0.160) is the highest among all emotions, suggesting a more unstable vocal quality which may reflect physiological tension. The HNR of 8.17 dB is very low but higher than Angry (7.78 dB), meaning this speech has more noise and less clarity than most other emotions.
Surprised	Surprised speech is characterized by a high pitch and is energetic but slightly unstable. Specifically, it has the highest mean pitch, which is 378.00 Hz, higher than Happy (375.51 Hz), making it the most exaggerated in tone. Its intensity, which is 69.71 dB, is relatively loud, close to Happy (70.14 dB). The speaking rate, which is 2.46 words/sec, is moderate, similar to Neutral (2.47 words/sec) and slower than Happy (2.63 words/sec) and Angry (2.59 words/sec). In terms of stability, shimmer (0.081) is high, suggesting a more fluctuating vocal quality. The HNR (14.59 dB) is lower than Neutral (18.79 dB) and Sad (18.36 dB), indicating a slightly noisier tone.	Surprised speech is high-pitched, loud, and dramatic, creating an exaggerated and expressive tone. It has a high mean pitch of 240.95 Hz, which is high but lower than Angry (242.00 Hz), contributing to its excited nature. The intensity, which is 74.68 dB, is the strongest among all emotions, making it the loudest speech among all emotions. The speaking rate, which is 3.11 words/sec, is relatively fast, but it is slower than both Happy (3.86 words/sec) and Angry (3.23 words/sec), suggesting a speech style that is energetic but not rushed. Jitter (0.033) is the highest among all emotions, indicating significant pitch variations, while shimmer (0.082) is also high, meaning the amplitude fluctuates more. The HNR (12.49 dB) is low, meaning the voice is noisier compared to Happy (14.61 dB) and Sad (15.56 dB) but still clearer than Angry (7.78 dB).
Disgusted	Disgusted speech is low in pitch, intensity, and clarity, making it sound rougher and less articulated. Specifically, its mean pitch (282.58 Hz) is the second-lowest, higher only than Neutral (245.70 Hz). The intensity,	Disgusted speech is rough, low-pitched, and unclear, making it sound harsh and less articulate than other emotions. It has a mean pitch of 162.45 Hz, which is higher than Neutral (130.56 Hz) and Afraid (150.18 Hz)

	which is 68.45 dB, is one of the lowest, similar to Afraid (68.48 dB). The speaking rate, which is 2.41 words/sec, is moderate, similar to Neutral (2.47 words/sec). In terms of stability, shimmer (0.098) is the highest, indicating significant vocal roughness. The HNR, which is 11.58 dB, is the lowest among all emotions, meaning it contains the highest level of noise, which aligns with the perception of this emotion as rough and breathy.	but lower than most other emotions. The intensity, which is 72.45 dB, is moderate, higher than Happy (69.81 dB) and Angry (70.95 dB), but slightly lower than Neutral (73.41 dB) and Afraid (73.38 dB), indicating a steady but not overly strong vocal presence. The speaking rate, which is 2.91 words/sec, is moderate, showing no extreme speed variation. Shimmer (0.075) is relatively high, suggesting an unstable vocal tone. The HNR, which is 10.87, is low, making it one of the noisier emotions with reduced vocal clarity.
Neutral	Neutral speech is balanced in all features, making it steady and clear but not overly expressive. It has the lowest mean pitch, which is 245.70 Hz, significantly lower than all other emotions. Its intensity, which is 68.41 dB, is moderate, lower than Happy (70.14 dB), Surprised (69.71 dB), and Angry (69.49 dB) but slightly higher than Afraid (68.48 dB) and Disgusted (68.45 dB). The speaking rate, which is 2.47 words/sec, is similar to Surprised (2.46 words/sec) and Disgusted (2.41 words/sec). In terms of stability, jitter (0.013) and shimmer (0.049) are the lowest, indicating a stable voice. The HNR, which is 18.79 dB, is the highest among all emotions, meaning it has the clearest and least noisy signal.	Neutral speech is relatively steady and less expressive compared to other emotions. Specifically, it has the lowest mean pitch among all emotions, which is 130.56 Hz, indicating a calm and stable vocal tone. Its intensity (73.41 dB) is relatively strong, slightly lower than Surprised (74.68 dB) but stronger than Disgusted (72.45 dB) and Happy (69.81 dB), making it a clear tone but not overly energetic. The speaking rate at 2.33 words/sec is within a mid-range speed, neither particularly slow like Afraid (1.31 words/sec) nor fast like Happy (3.86 words/sec) or Angry (3.23 words/sec). Jitter (0.025) and shimmer (0.070) are moderate, indicating the voice is relatively stable. The HNR (11.17 dB) is relatively low, suggesting more background noise, but it is cleaner than Angry (7.78 dB) or Afraid (8.17 dB).

5. (20 pts) 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: Firstly, pitch trends across emotions are similar between the two datasets. In both datasets, Surprised and Angry tend to be relatively high in the mean pitch, while Neutral and Disgusted are relatively low. For instance, Surprised speech has the highest pitch in both datasets (My speech: 378.00 Hz, Podcast: 240.95 Hz), while Neutral has the lowest (My speech: 245.70 Hz, Podcast: 130.56 Hz). Additionally, patterns of speaking rate between datasets are also similar. Happy and Angry are the fastest, and Sad is relatively slow. Specifically, in my speech, Happy (2.63 words/sec) and Angry (2.59 words/sec) are the fastest, while Sad (2.16 words/sec) is the slowest. Similarly, in podcast speech, Happy (3.86 words/sec) and Angry (3.23 words/sec) are the fastest, and Sad (2.17 words/sec) is relatively slow.

Differences: Firstly, my speech has higher overall pitch values compared to the podcast speech dataset. My pitch ranges from 245.70 Hz to 378.00 Hz, while the Podcast speech dataset has a lower pitch range of 130.56 Hz to 240.95 Hz. For example, for Happy speech, my pitch is 375.51 Hz, but the Podcast dataset has 206.92 Hz. Similarly, in the case of Neutral speech, my pitch is 245.70 Hz, and for the Podcase dataset is 130.56 Hz. This suggests that my recordings may have involved a more extreme emotional expression style than the podcast dataset. In addition, while my speaking rates in the different emotions are quite consistent, the podcast dataset is more variable. My speaking rate also falls within a narrower range from 2.16 words/sec to 2.63 words/sec, while the Podcast dataset is from 1.32 words/sec to 3.86 words/sec. This may be because Podcast speech is highly conversational and expressive. Lastly, HNR values are greater in my speech, which suggests a cleaner signal, while the Podcast dataset has lower HNR values, which indicates more background noise or naturally rougher vocal textures. My HNR values range from 11.58 dB to 18.79 dB, while the Podcast dataset has a lower range, from 7.78 dB to 15.56 dB. This suggests that my recordings would contain fewer environmental noise variables, and the Podcast dataset would contain more background noise or naturally rougher vocal textures.

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

I think that the Podcast speech dataset would likely be more useful for emotion recognition applications due to its greater variation in pitch and speaking rate, which are key factors in distinguishing emotions. Since the goal of emotion recognition is to detect patterns that differentiate emotions clearly, the Podcast speech dataset, with its wider range of vocal expressions and variations, provides richer information for training a robust model. Even though I tried to immerse myself in each emotion while recording, my speech can still have individual speaking habits and environmental influences that introduce inconsistencies and may make it less ideal for training emotion recognition models. Podcast speakers, who are likely more experienced in expressive speech, may also be able to convey emotions more effectively than I, a non-native English speaker. Their ability to control vocal nuances, pitch, and rhythm might result in a more natural and dynamic dataset for emotion recognition. Moreover, the Podcast dataset includes multiple speakers, while my dataset consists of a single speaker. In machine learning, training on data from diverse speakers helps a model generalize better to different voices, making the Podcast dataset more practical for real-world emotion recognition applications.

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

In this case, my speech dataset would likely be easier for an emotion recognition system to classify because it exhibits more consistency in pitch and speaking rate across emotions. For machine learning models, a consistent distribution of features makes classification more straightforward, reducing the need for extensive normalization. The speaking rate remains relatively stable across emotions, reducing ambiguity in tempo-based classification. Additionally, in terms of the recording environment, since I intentionally recorded each emotion with clear articulation, the dataset is cleaner and has less background variation, making it easier for a model to learn from. On the other hand, the Podcast speech contains multiple speakers and varying recording conditions, leading to greater vocal diversity and inconsistencies, which can make classification more challenging.

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

Based on my insights from the referenced paper (in README), Mel-Frequency Cepstral Coefficients (MFCCs), a representation of the spectral characteristics of an audio signal, would be useful for emotion recognition. Since MFCCs are specifically designed to simulate the way the human ear perceives different frequencies, they effectively capture subtle variations in tone and timbre, which are essential for distinguishing between emotions. Additionally, formant frequencies (F1, F2, F3), which are resonant frequencies of the vocal tract, can help capture how speech articulation differences using different emotions. These features rely on physiological changes in speech production and are, therefore, beneficial for emotional state recognition.

Bonus Problem) Manipulation

Based on the analysis of my speech features, intensity showed only minor variations, whereas pitch increased by approximately 1.5 times, and the speaking rate increased by about 1.1 times. Therefore, to transform the neutral speech into happy speech, I manipulated only the pitch and duration, increasing them by factors of 1.5 and 1.1, respectively.