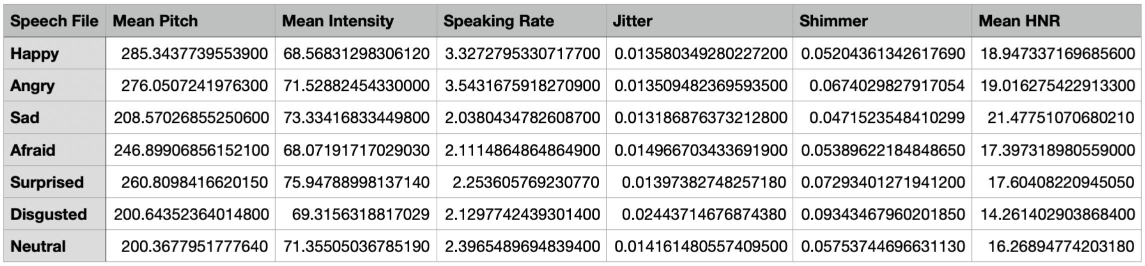
Transcript for my recordings/speech: “My mama lives in Memphis”

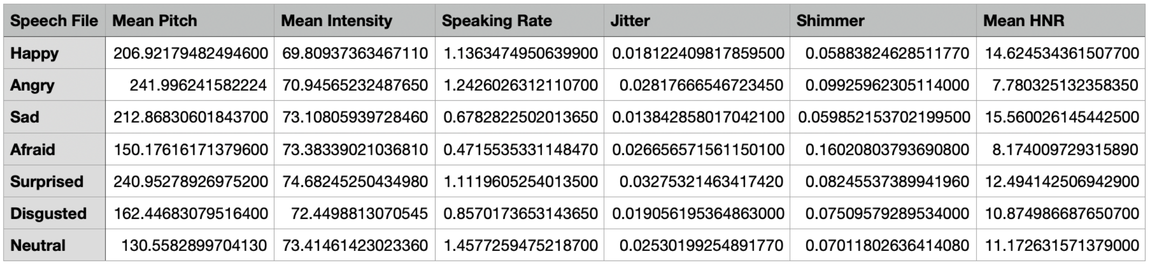
The tables below present the mean values of the extracted speech features from my recordings and from the MSP Podcast samples obtained using the **features_extract.py** script.

Table 1. Extracted Speech Features (My Speech Recordings)



Speech File	Mean Pitch	Mean Intensity	Speaking Rate	Jitter	Shimmer	Mean HNR
Happy	285.3437739553900	68.56831298306120	3.3272795330717700	0.013580349280227200	0.05204361342617690	18.947337169685600
Angry	276.0507241976300	71.52882454330000	3.5431675918270900	0.013509482369593500	0.0674029827917054	19.016275422913300
Sad	208.57026855250600	73.33416833449800	2.0380434782608700	0.013186876373212800	0.0471523548410299	21.47751070680210
Afraid	246.89906856152100	68.07191717029030	2.1114864864864900	0.014966703433691900	0.05389622184848650	17.397318980559000
Surprised	260.8098416620150	75.94788998137140	2.253605769230770	0.01397382748257180	0.07293401271941200	17.60408220945050
Disgusted	200.64352364014800	69.3156318817029	2.1297742439301400	0.02443714676874380	0.09343467960201850	14.261402903868400
Neutral	200.3677951777640	71.35505036785190	2.3965489694839400	0.014161480557409500	0.05753744696631130	16.26894774203180

Table 2. Extracted Speech Features (MSP Podcast Speech Samples)



Speech File	Mean Pitch	Mean Intensity	Speaking Rate	Jitter	Shimmer	Mean HNR
Happy	206.92179482494600	69.80937363467110	1.1363474950639900	0.018122409817859500	0.05883824628511770	14.624534361507700
Angry	241.996241582224	70.94565232487650	1.2426026312110700	0.02817666546723450	0.09925962305114000	7.780325132358350
Sad	212.86830601843700	73.10805939728460	0.6782822502013650	0.013842858017042100	0.059852153702199500	15.560026145442500
Afraid	150.17616171379600	73.38339021036810	0.4715535331148470	0.026656571561150100	0.16020803793690800	8.174009729315890
Surprised	240.95278926975200	74.68245250434980	1.1119605254013500	0.03275321463417420	0.08245537389941960	12.494142506942900
Disgusted	162.44683079516400	72.4498813070545	0.8570173653143650	0.019056195364863000	0.07509579289534000	10.874986687650700
Neutral	130.5582899704130	73.41461423023360	1.4577259475218700	0.02530199254891770	0.07011802636414080	11.172631571379000

Description of the characteristics of each emotion within each dataset. For each dataset separately, the features of a given emotion are discussed in relation to the other emotions in that dataset (for example, in the MSP podcast speech, happy is compared against the other emotions). Cross-dataset comparisons are not included.

Emotion	Your Speech	MSP Podcast Speech
Happy	“Happy” has the highest mean pitch (285 Hz) among all emotions. It also shows the second fastest speaking rate (3.33 Hz) just behind “Angry” (3.54 Hz). Jitter and shimmer values are lower than those observed in emotions such as “Afraid,” “Surprised,” “Disgusted,” and “Neutral,” which contributes to a clearer, more positive tone.	“Happy” also exhibits a moderately high mean pitch relative to the overall mean. Its speaking rate is faster than the speaker’s average with moderate jitter and shimmer, reinforcing the impression of a positive emotional state.
Angry	“Angry” has the second highest mean pitch (276 Hz) and the fastest speaking rate (3.54 Hz). It also shows higher shimmer than the	“Angry” displays the highest mean pitch (242 Hz) and a very wide pitch range (up to 510 Hz). Jitter and shimmer values are higher than in most

	calmer emotions resulting in a tenser and harsher vocal quality.	other emotions, again producing a tense, strained sound.
Sad	“Sad” has the lowest speaking rate (2.03). Although the maximum pitch is very high (541 Hz), the mean pitch is moderate (209 Hz). It also has the highest HNR (21.49), resulting in a clear and resonant vocal quality.	“Sad” has the second slowest speaking rate (0.68) just above “Afraid.” The mean pitch is relatively high (213 Hz) and it shows the highest HNR (15.6), which contributes to a clearer tone compared to other emotions.
Afraid	“Afraid” shows a relatively high mean pitch (247 Hz), though still lower than “Happy” and “Angry.” Its speaking rate (2.11) is slower than most of the other emotions giving it a more hesitant quality.	“Afraid” has a mean pitch of 150 Hz, the second lowest across all emotions. Its speaking rate (0.47) is the slowest of all reflecting a subdued and restrained delivery.
Surprised	“Surprised” is marked by a wide pitch range with the second highest maximum pitch (544 Hz). The mean intensity (76 dB) is the highest among all emotions, giving the speech a lively and energetic character.	“Surprised” also has a broad pitch range with a mean pitch of 241 Hz, the second highest overall. Its intensity (75 dB) is the strongest of all emotions and the speaking rate (1.11) is relatively fast further reinforcing its energetic tone.
Disgusted	“Disgusted” has the second lowest mean pitch (200.6 Hz) and a moderate speaking rate (2.13). It exhibits the highest shimmer (0.09) and the lowest HNR (14.26), which combine to produce a rougher, less smooth vocal quality.	“Disgusted” has the third lowest mean pitch (162 Hz) and a slower speaking rate (0.857) compared to “Angry,” “Surprised,” and “Neutral.” Its HNR is the lowest (10.87) creating a harsh and rougher sound.
Neutral	“Neutral” has the second-lowest mean pitch (200.6 Hz) and the third-fastest speaking rate (2.4). Jitter and shimmer values are moderate producing a balanced and steady tone.	“Neutral” has the lowest mean pitch across all emotions (131 Hz) but the fastest speaking rate (1.46). Both jitter and shimmer are moderate, resulting in a stable and even vocal quality.

What are some similarities and differences between the features from the two datasets?
Both datasets show relatively high pitch for "Happy" compared to their overall average pitch across emotions. However, in my recording "Happy" has an extremely high mean pitch (285Hz) and is relatively fast (3.33) whereas MSP's "Happy" is more moderate in mean pitch and speed.

For "Angry," both datasets show the highest mean pitch ("Angry" has the second highest mean pitch after "Happy" in my dataset. "Angry" has the highest mean pitch in the MSP dataset). The speaking rate for "Angry" is the fastest in my dataset, whereas the speaking rate is on the higher end in the MSP dataset.

For both datasets, "Surprised" has a wide pitch range. "Surprised" has the second highest max pitch in my dataset while "Surprised" has the second highest mean pitch in the MSP dataset.

In both datasets, "Sad" has the highest HNR, which suggests less noisy and cleaner sounds for "Sad" recordings. Also, "Sad" has the slowest speaking rates for both datasets. However, while "Sad" has a very high max pitch in my dataset (541Hz), in the MSP dataset "Sad" has a more contained pitch range.

In both datasets, "Afraid" has relatively slow speaking rates. However, in my dataset "Afraid" has a relatively higher pitch compared to other emotions while in the MSP dataset "Afraid" has the second lowest mean pitch.

In both datasets, "Disgusted" has the lowest HNR out of all emotions and higher shimmer. For "Neutral," both datasets have the lowest mean pitch (in my dataset, "Neutral" has the second lowest mean pitch. In the MSP dataset "Neutral" has the lowest mean pitch). Both datasets have faster speaking rates for "Neutral."

Therefore, both datasets show similar patterns in that strong emotions such as "Surprised," "Angry," and "Happy" exhibit higher pitch, faster speaking rates and higher intensity compared to other emotions. "Angry" and "Surprised" show peaks in max pitches in both datasets compared to more subtle emotions. "Sad" and "Afraid" show slower rates and lower intensity. In both datasets, more tense emotions such as "Angry" and "Disgusted" show high shimmer and lower HNR.

My dataset has generally higher mean pitches across all emotions compared to the MSP dataset. Furthermore, the speaking rate is higher than that of the MSP dataset.

Which of the datasets would be more useful for emotion recognition applications? Why?
I believe the MSP Podcast dataset would be more valuable for emotion recognition applications due to its diverse speaker representation and varied recording conditions. For practical deployment in real world scenarios, a robust dataset must encompass a broad spectrum of speaking styles, vocal characteristics, and recording environments to ensure effective generalization across different use cases and user populations. In contrast, my single speaker dataset presents a limitation as models trained exclusively on one voice would likely overfit to my specific speaking patterns, intonation, and vocal characteristics. This overfitting would compromise the model's ability to accurately recognize emotions in voices that differ from mine, ultimately rendering it ineffective for widespread real world applications where users exhibit diverse vocal traits and speaking styles.

Which of these datasets would be easier for an emotion recognition system to classify? Why?
I believe my dataset would be easier to classify because it contains the same voice and consistent recording conditions, which allows emotion recognition systems to more easily identify the differences and patterns between each emotion, leading to improved classification accuracy. With the MSP dataset, it might be more challenging for emotion recognition systems to focus solely on

the differences and patterns between emotions due to the variability introduced by different speakers and diverse recording conditions.

What other features would be useful for emotion recognition? Why?
Temporal prosodic features such as pause frequency and duration, phrase length, and speech rhythm would be valuable for emotion recognition as they capture the timing and fluency patterns of speech which vary significantly across different emotional states. For instance, emotions like "Afraid" and "Sad" typically exhibit more frequent and longer pauses, reflecting hesitation or contemplation, while emotions such as "Angry," "Happy," and "Surprised" tend to demonstrate more fluent speech with fewer interruptions due to heightened arousal and urgency.

Additionally, formant frequencies (F1, F2, F3) would provide information about vocal tract configuration changes across different emotions. These acoustic features reflect how emotional states physically alter speech production mechanisms. For example, "Happy" emotions often result in higher formant frequencies due to increased jaw opening and elevated larynx position creating a brighter vocal quality. Conversely, "Sad" emotions typically produce lower formant frequencies as a result of larynx lowering and reduced articulatory precision, contributing to a more muffled or dampened vocal quality. These physiological changes in vocal tract shape directly correlate with the acoustic signatures that distinguish emotional expressions in speech.