COMS 6998 Advanced Topics in Spoken Language Processing -- Spring 2025
Homework 1: Speech Analysis (20%)

In this homework you will practice extracting features from speech. You will analyze your own emotional speech as well as the speech in the provided examples.

GPT-use affirmation: Please confirm your compliance by typing the following statement at the top of your report: **'I confirm that I have not used any GPT-generated responses for any part of this assignment.'**

Submission (10 pts)

Submit a zipped folder <uni.zip> containing the following:

- 2 spreadsheets with feature values, based on the provided templates: my_features.csv; msp_features.csv.
    - Note: your spreadsheets should ***exactly match*** the templates provided on naming and row/column headers.
- Your code (.py) for feature extraction
- Text file (.pdf or .docx) with your responses
- Your recordings (Happy.wav, Angry.wav, Sad.wav...)
    - Note: your recording names should ***exactly match*** the names in the MSP samples folder.
- Read.me file explaining how to run your script and references.

Please make sure the file names are correct.

DO NOT include the MSP samples in your submission.

Feature extraction notes

- Pass praat script commands into the parselmouth.praat.call() function for feature extraction. Do NOT use numpy functions or built-in methods of the form `to_feature()`.
- For pitch extraction, set pitch floor to 75Hz, and pitch ceiling to 600Hz.
    - Note: avoid using autocorrelation.
- For intensity extraction, set the pitch floor to 100Hz. Use 'energy' averaging method to get mean intensity.
- For jitter, extract local jitter only, and set period floor to 0.0001s, period ceiling to 0.02s, and maximum period factor to 1.3.
- For shimmer, extract local shimmer only, and set period floor to 0.0001s, period ceiling to 0.02s, maximum period factor to 1.3, and maximum amplitude factor to 1.6.
- To calculate HNR (harmonics-to-noise ratio), extract harmonicity (cc) first. Set time step to 0.01, minimum pitch to 75Hz, silence threshold to 0.1, and number of periods per window to 1.0.
- Speaking rate can be approximated with #words/duration. Please indicate which method you use in your submission and how you obtain the results, e.g. transcripts, helper scripts.

Resources (see resources link from syllabus)

Praat
Parselmouth: a Python library for Praat software

1. (14pts) Record yourself producing emotional speech (one sentence each). For best results, do this in a quiet room and wear a headset with a microphone. Trim leading and trailing silence.

a) Happy
b) Angry
c) Sad
d) Afraid
e) Surprised
f) Disgusted
g) Neutral

2. (30 pts) Write a python script to extract the following features using parselmouth. Print them to a csv.

- Pitch (min, max, mean, sd)
- Intensity (min, max, mean, sd)
- Speaking rate
- Jitter
- Shimmer
- HNR

3. (11 pts) Extract the same features from the files provided from the MSP-Podcast corpus. Print them to a csv. Please make sure the column and row names match the given template.
MSP-Podcast corpus description
MSP-Podcast samples (Coursework->Files->HW1->msp-samples.zip)

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 values are not useful; instead consider comparisons of relative feature values (e.g. compared to each speaker's mean).

Emotion	Your speech	Podcast speech
Happy		
Angry		
Sad		
Afraid		
Surprised		
Disgusted		
Neutral		

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?
b. Which of the datasets would be more useful for emotion recognition applications? Why?
c. Which of these datasets would be easier for an emotion recognition system to classify? Why?
d. What other features would be useful for emotion recognition? Why?

Bonus problem (2 pts): Using Praat, manipulate the pitch contour of your neutral utterance to convert it to sound happy. Please submit both the manipulation text file as ”bonus.Manipulation” and the wav file as “bonus.wav”. Note the expected file endings as underlined.