Capstone Final Proposal

Speech Recognition and Classification

Harish Ram, Zeqiu Zhang, Sisi Zhang, Jiachen Sands

Project Motivation:
Speech detection is a critical research topic, with applications ranging from analyzing speech patterns to disease classifications for schizophrenia or Alzheimer patients. Our group will concentrate on how speech data may be utilized to analyze sentiment and understand emotions, gender, accents, and age groupings, as well as the disease classification in this research. We will employ data from a variety of sources, as well as voices with various accents and voice lines. By doing this, we can assure diversity in our data and increase model accuracy when we use neural network-based algorithms to model our data. This research will be used for further analysis for a degree of mental health degradation by voice as a pre-trained model as a final result.

Project Plan:
The group will collect data from a variety of sources and create models for each category. The deep speech model will be utilized as a benchmark, and several transformers will be pre-trained from scratch and applied to increase model accuracy. As a final product, the team will create a graphical user interface (GUI) with two main components. Users can upload an audio file to the application as an mp3, WAV, WMA, etc and several models will analyze it to classify the speaker's emotions, accent, gender, and age. Another feature of the GUI determines whether the speaker has Parkinson disease and whether or not it is noticeable in their voice. Additionally, the audio file will be transcribed and the text will be color-coded for different types of emotions. Along with the project's development, more related data visualizations will be added to interpret the final output.

Category	**Data Source**	**Plan To Do**
Emotions (Harish Ram)	1. https://github.com/siddiquelatif/urdu-dataset (Urdu language, 27 male and 11 female, no age, Angry, Happy, Neutral, Sad) 2. https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=doi%3A10.5683%2FSP2%2FE8H2MF (English language, 2 women, young and old, anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral) 3. http://m3c.web.auth.gr/research/aesdd-speech-emotion-recognition/ (Greek	● Perform emotion classification ● Combine datasets to increase the number of sound files for each emotion to provide more diversity of the speaker (age, gender, language) ● First use pre-trained deep speech models to determine a baseline

	language, multiple men and women, multiple ages (unknown), anger, disgust, fear, happiness, and sadness) 4. https://www.kaggle.com/a13x10/basic-arabic-vocal-emotions-dataset (Arabic, 6 men and women, multiple ages (unknown), 3 levels of emotions) 5. https://github.com/CheyneyComputerScience/CREMA-D (91 actors, 48 male and 42 female actors, 12 sentences, 6 emotions with 4 emotion levels)	● Tune and pre-train numerous transformer models (Wav2Vec, etc) to exceed baseline score ● Understand how the architecture works ● Choose the most representative model that does emotional classification the best
Accent (Jiachen Sands)	1. https://www.kaggle.com/rtatman/speech-accent-archive?select=recordings (2138 Mp3 files, speakers from 176 countries, 50% speakers for each gender) 2. https://datashare.ed.ac.uk/handle/10283/3443 (10GB of FLAC files in total, will only use 1GB data from this source) 3. The first data source has at least 400 audio files from native English speakers. So we will use non-English speakers data from the second source only)	● Give a score of the pronunciation on the input audio file ● Detect speaker's origin (Mandarin, Korean, Irish, etc.) ● Combine two data sources together. May combine the audio files for each speaker in the second data source. ● Use Deep Speech model baseline ● Try and pre-train multiple transformers and decide a best one

Gender (Sisi Zhang)	1. https://github.com/NeuroLexDiagnostics/voice_modeling_starter (52 females wav files + 52 males wav files) 2. https://github.com/NeuroLexDiagnostics/voice_gender_detection (females 2312 m4a files +3683 males m4a files total size 390mb) 3. https://github.com/CheyneyComputerScience/CREMA-D (7443 wav files with csv file contains age, sex, race, ethnicity information. 91 actors, 48 male and 42 female actors, 12 sentences, 6 emotions with 4 emotion levels)	● Data preprocessing and EDA to know the data ● Perform gender classifications with Deep Speech model to get the baseline ● Data Augmentation and compare the result ● Use different transformer model and determine the best one ● Train on the best model to further improve the score
Age (Everyone)	1. https://github.com/CheyneyComputerScience/CREMA-D (7443 wav files with csv file contains age, sex, race, ethnicity information. 91 actors, 48 male and 42 female actors, 12 sentences, 6 emotions with 4 emotion levels) 2. https://www.kaggle.com/rtatman/speech-accent-archive?select=recordings (2138 Mp3 files, speakers from 176 countries, 50% speakers for each gender)	
Disease Classification (Zeqiu Zhang)	1. https://zenodo.org/record/2867216#.XeTbN59R2BZ 2. https://github.com/Mak-Sim/Troparion/tree/master/SPA2019	
Transcription (Everyone)	1. https://keithito.com/LJ-Speech-Dataset/ (3.5GB, the metadata.csv provides the transcription, 278 samples)	●
Making GUI (Everyone)	N/A	● Use Dash ● Design a webpage that people can upload audio files and it will output the

		classification result with a transcription

Timeline:

- (3 Weeks) Speech Data Wrangling:
    - Collecting data and engineering datasets
    - Define data loader to convert the data for later modeling
- (2 Weeks) Speech Data Processing:
    - EDA, experiment LibROSA, PyAudio for feature extraction
    - Explore text features, audio features, mixed features and meta features
    - Dimensionality reduction and feature selection
    - Possible data augmentation (Time Shift & Time and Frequency Masking)
- (1 Week) Get model baseline:
    - Run deep speech model to get the baseline
- (1 Week) Modeling with deep learning models:
    - CNN, LSTM, ensemble models
- (4 Weeks) Modeling with different transformers:
    - SEW, UniSpeech, UniSpeechSat, WavLM, Wav2Vec, Wav2Vec2Phoneme
- (1 Week) Classifications, transcriber heads
    - Test on different classification tasks
    - Test on speech transcriptions
- (1 Week) Analysis
    - Model analysis
    - Challenges and future work suggestions
- (1 Week) Writing Up a paper and submission
    - Writing a paper and style it properly
- (1 Week) Final Presentation
    - Make slides and practice for presentation

Feedback/Concerns:

- How to use the cloud to download massive sets of audio data such as Common Voice and Gigaspeech?
- What type/size of data do we need to pre-train our models on?
- Credits for GCP? How to store data on the cloud?
- Schedule a session with the professor?
- 1 - 2 GB data size for pre-training or training?
- How do we combine datasets? How to combine models from different categories?
- Accent: Should I combine audio files for each speaker in the second data source to make it consistent to the first one?

- Accent: Should I delete duplicate audio files (same person saying same thing, but time is different)
- Accent: Is that possible to do speech analysis (strength, weakness, areas for improvement)