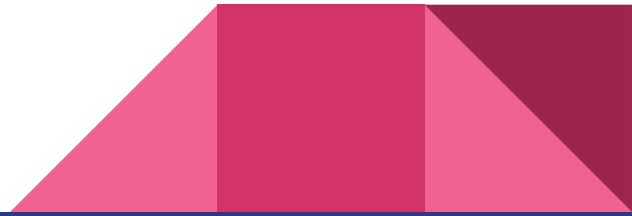
SpeechTeach

McMaster University
Department of Electrical & Computer Engineering

Raphael Capon, Veneta Grigorova, Hira Nadeem, Ben Raubvogel
Group 13

Purpose

“To support speech therapists with an automated stutter recognition program which records patients speech and identifies stutter type and time.”



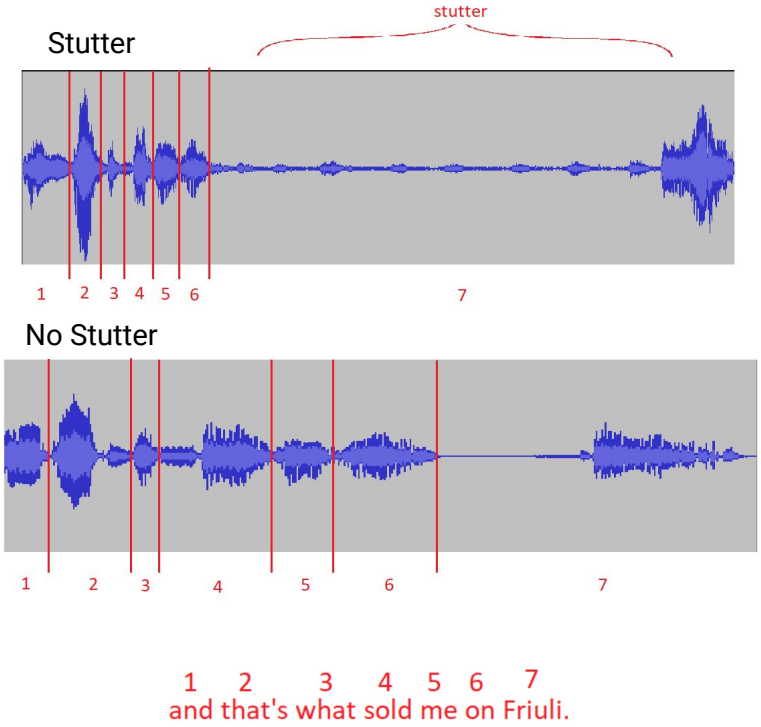
What is a Stutter?

Stuttering: a speech disorder which results in the involuntary disruption of speech by repetitions, elongations, or inability to pronounce sounds.

Three types of stutters:

- Repetition
  - Ex. “Hi my na-na-name is”
- Elongated
  - Ex. “Hi mmm-my name is”
- Locked
  - Ex. “Hi my [pause] name is”

Identifying a Stutter



Stuttering can be easily be identified by the waveform.

Research & Resources

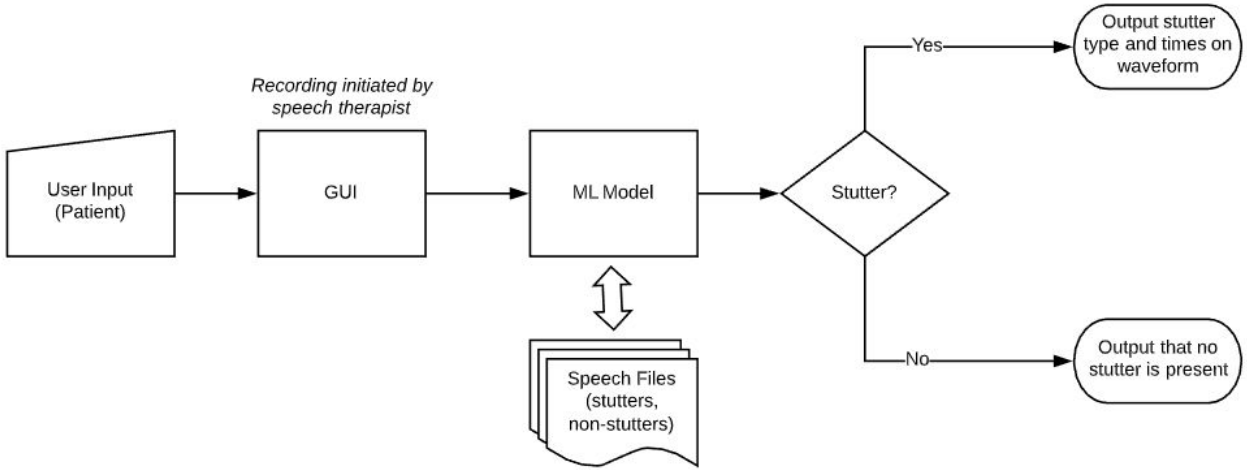
- Speech and Stuttering Institute (Toronto)
- **Anna Tandera**, PhD at Holland Bloorview Kids Rehabilitation Hospital (Toronto)
- The Centre for Advanced Research in Experimental and Applied Linguistics (**AERiAL**) at McMaster
- **Dr. Hassan Ashtiani**, Machine Learning Professor in CAS McMaster

Challenges Identified

- Recording audio of patients
- Identifying stutters in audio
- Timestamping when the stutter occurs

Approach

- Create a detailed and easy to use Graphical User Interface (GUI)
- Train a machine learning model to identify stutters from recorded speech
- Train a machine learning model to identify when a stutter occurs from inputted speech



A Look at Machine Learning

Two goals

1. Identify if a stuttered occurred, and what type.
2. Identify where the stutter occured.

Selected Tools

- Scikitlearn
- Keras
- Tensorflow
- Praat-Parselmouth

ADD LOGOS

Understanding the Features of Speech

General Audio Analysis Techniques (time series, frequency domain)

- Time series analysis
- Frequency domain analysis

Specialized Speech/Audio Tools

- Mel Frequency Cepstral Coefficients (MFCC)
- Praat-Parselmouth API in Python
    - Speech features examples: Jitter, Shimmer, Voice Breaks

Image processing approach

- Convolutional Neural Network (CNN)
- Spectrograms

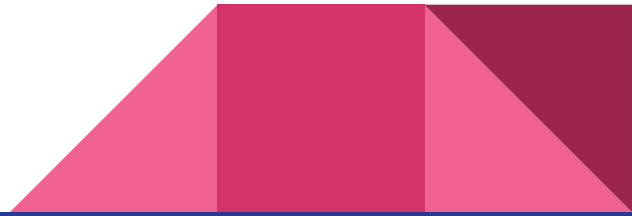
Data Set

A machine learning model is only as strong as its dataset

- Large data set
- Clean, normalized and sorted data
- Equal parts stuttered and non-stuttered data

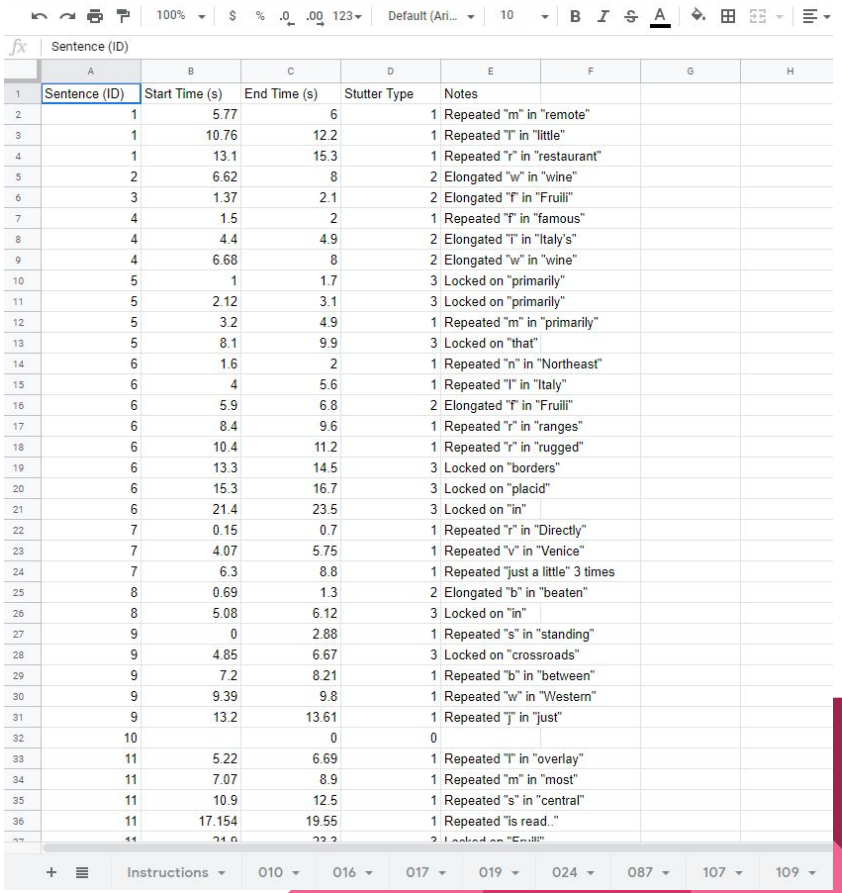
Data collected from:

- FluencyTalkBank Adults with Stutter database
- Anna Tendara, PhD
- Google Form for non-stuttered data



Organizing the Data

- A data set of approximately 25 different patients, each speaking 14 sentences from a transcript
- Each sentence was notated based on type and time (s) of stutter instances
- Type
  - 1 = repeated
  - 2 = elongated
  - 3 = locked



Sentence (ID)

Sentence (ID)	Start Time (s)	End Time (s)	Stutter Type	Notes
1	5.77	6	1	Repeated "m" in "remote"
1	10.76	12.2	1	Repeated "l" in "little"
1	13.1	15.3	1	Repeated "r" in "restaurant"
2	6.62	8	2	Elongated "w" in "wine"
3	1.37	2.1	2	Elongated "f" in "Fruili"
4	1.5	2	1	Repeated "f" in "famous"
4	4.4	4.9	2	Elongated "i" in "Italy's"
4	6.68	8	2	Elongated "w" in "wine"
5	1	1.7	3	Locked on "primarily"
5	2.12	3.1	3	Locked on "primarily"
5	3.2	4.9	1	Repeated "m" in "primarily"
5	8.1	9.9	3	Locked on "that"
6	1.6	2	1	Repeated "n" in "Northeast"
6	4	5.6	1	Repeated "l" in "Italy"
6	5.9	6.8	2	Elongated "f" in "Fruili"
6	8.4	9.6	1	Repeated "r" in "ranges"
6	10.4	11.2	1	Repeated "r" in "rugged"
6	13.3	14.5	3	Locked on "borders"
6	15.3	16.7	3	Locked on "placid"
6	21.4	23.5	3	Locked on "in"
7	0.15	0.7	1	Repeated "r" in "Directly"
7	4.07	5.75	1	Repeated "v" in "Venice"
7	6.3	8.8	1	Repeated "just a little" 3 times
8	0.69	1.3	2	Elongated "b" in "beaten"
8	5.08	6.12	3	Locked on "in"
9	0	2.88	1	Repeated "s" in "standing"
9	4.85	6.67	3	Locked on "crossroads"
9	7.2	8.21	1	Repeated "b" in "between"
9	9.39	9.8	1	Repeated "w" in "Western"
9	13.2	13.61	1	Repeated "j" in "just"
10		0	0	
11	5.22	6.69	1	Repeated "l" in "overlay"
11	7.07	8.9	1	Repeated "m" in "most"
11	10.9	12.5	1	Repeated "s" in "central"
11	17.154	19.55	1	Repeated "is read.."

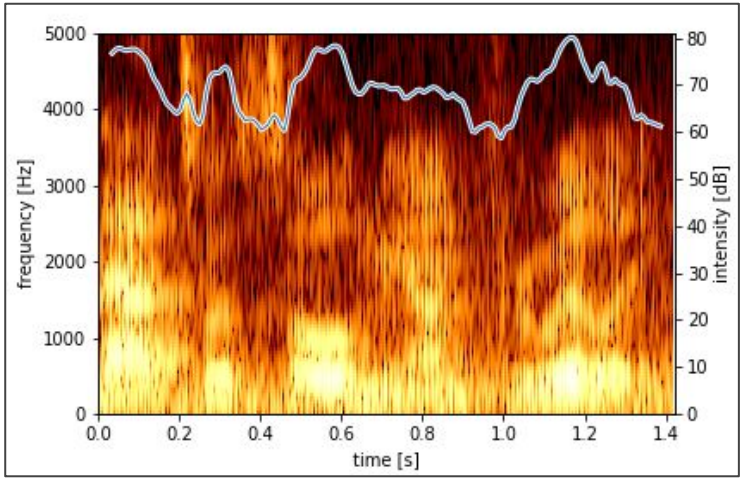
Instructions | 010 | 016 | 017 | 019 | 024 | 087 | 107 | 109

Spectrograms

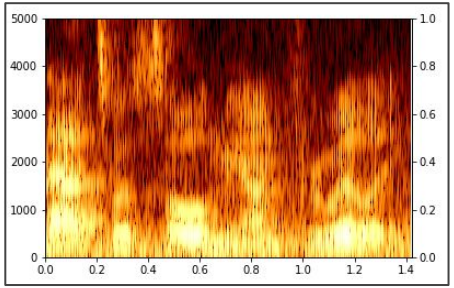
Spectrograms are a very useful tool for visualizing speech

- Gives a 3D perspective of **time**, **frequency** and **intensity**
- Plots generated using parselmouth

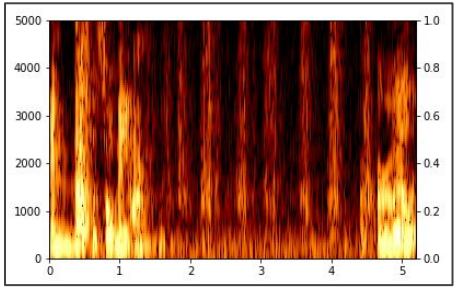
Sample Spectrogram



Spectrogram Visualization of Stuttering



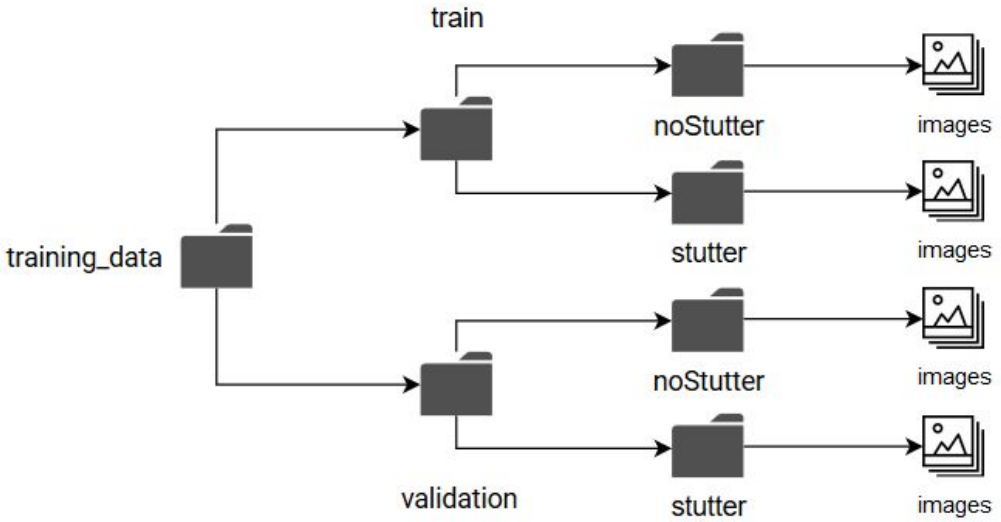
Non-Stutter: Even delivery



Stutter: Clear repetition of a syllable

File Structure

The model receives the data in a file structure with sorted images.



Model Specifications

Hidden Layers

- 4 **Convolution** and **MaxPooling** layers.
- Several **Activation**, **Dense**, **Flatten** and **Dropout** layers.

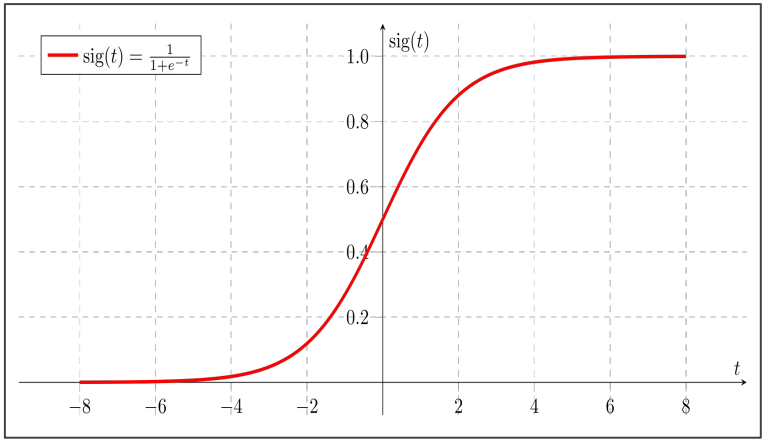
Probability Function: **Sigmoid**

Output: [noStutter stutter]

- Example: A stutter would be classified as [0 1]

Number of **Epochs**: 30

Sigmoid Function



Retrieved from https://towardsdatascience.com

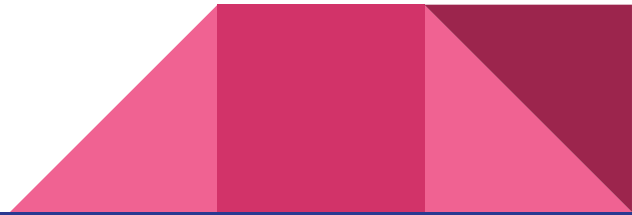
Training Results

Performance

- 3 Models were trained (one for each sentence prompt).
- Consistently reported over 80% accuracy during training.
- Model struggled with user-generated recordings.
    - Potentially due to lack of consistency in recording equipment, signal processing or other factors.

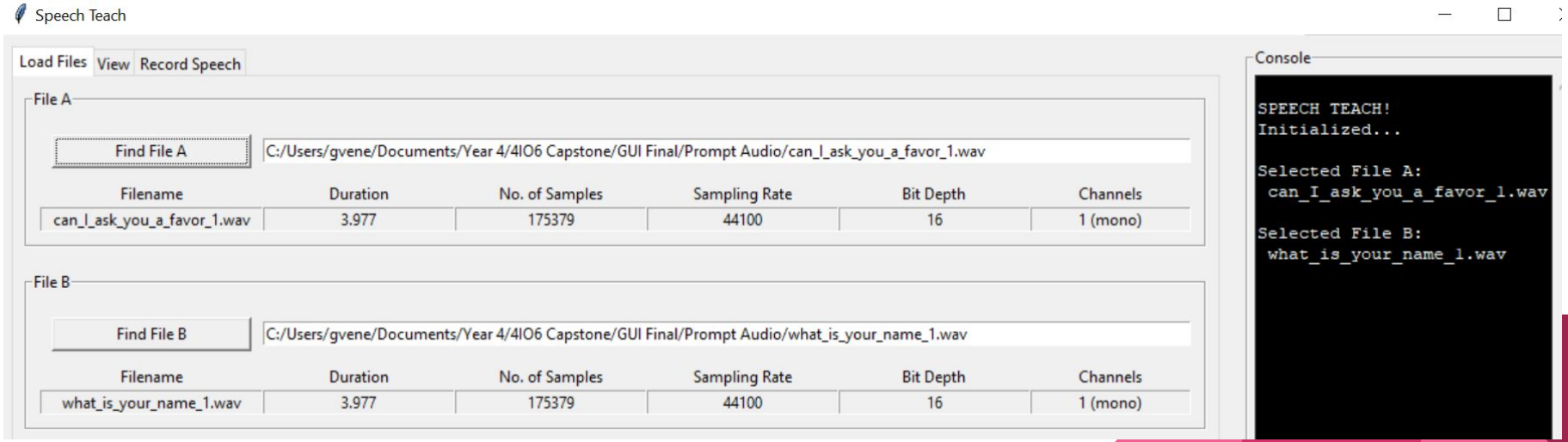
Limitations from Data Set Size

- Reliable classification by stutter type was not possible given size of data set.



A Look at the GUI | Loading Files

- Load .wav files for plotting in the View tab.
- Read basic file properties (sample rate, channels, sample width etc.).
- Two files can be loaded for comparison.

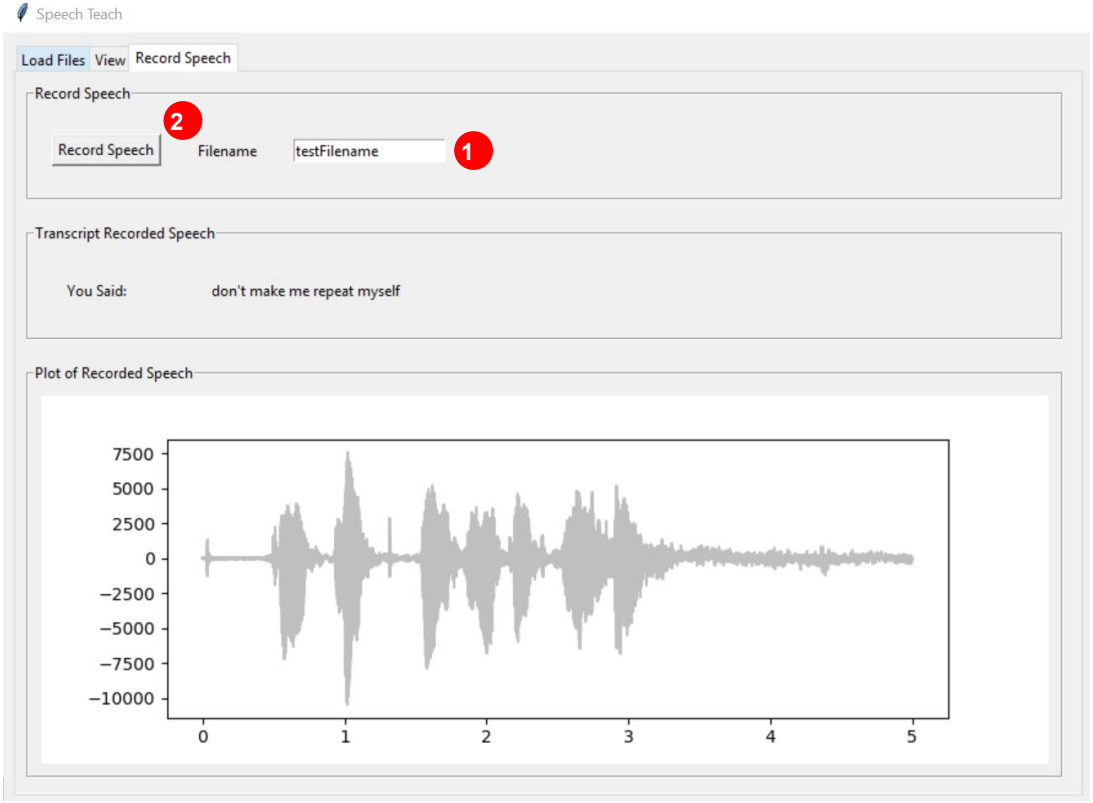


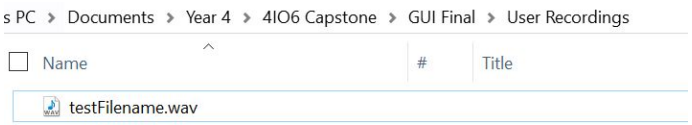
A Look at the GUI | The Record Tab

The Record tab records audio and prepares it for analysis.

1. The user enters the desired name of the file they want to save to
2. The user clicks “Record Speech” button
3. The program records for 5 seconds
4. After recording stops, the speech-to-text transcript and the plot are displayed

A Look at the GUI | The Record Tab





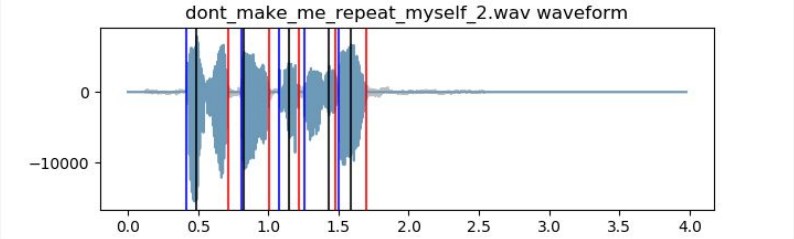
A Look at the GUI | The View Tab

1. User selects type of information to plot.
    a. “Chunks” of spoken text
    b. Silences between spoken text
    c. Peaks of the spoken text
2. “Play” and “Stop” buttons are provided for playback
3. “Analyze” function detects a stutter and displays the result
4. Secondary plot provided for comparing with another file

Speech Teach

Load Files | View | Record

File A



File A Options

Refresh A

- [ ] show where silent segments start
- [ ] show where silent segments end
- [x] show where sound chunks start
- [x] show where sound chunks end
- [x] show chunk peaks

Play Audio

Stop

File A Analysis

Perform Stutter Analysis

No Stutter Detected

Each model was trained using a single prompt:

Model 3
And that's what sold me on Friuli.

Model 10
As a result, things look different here.

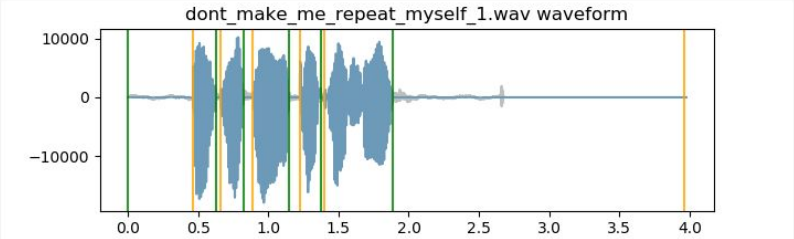
Model 13
Here you'll find grey stone castles rather than sundrenched villas.

Select a model:

10

File A Playback Progress

File B



File A Options

Refresh B

- [x] show where silent segments start
- [x] show where silent segments end
- [ ] show where sound chunks start
- [ ] show where sound chunks end
- [ ] show chunk peaks

Play Audio

Stop

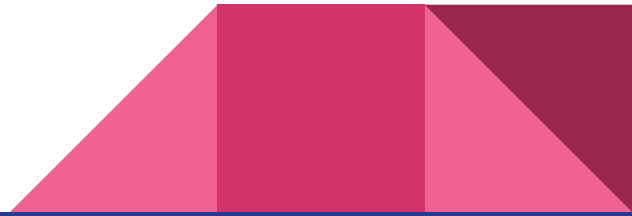
File B Playback Progress

Console

```
Playback Stopped

read new file conte

 (working) Getting
Theshold= 55.0
min. silence sample
min. chunk samples=

(working) Getting C
silences...

Getting Peak sample

potting waveform...

adding event lines

adding event lines

read new file conte

 (working) Getting
Theshold= 55.0
min. silence sample
min. chunk samples=

(working) Getting C
silences...

Getting Peak sample

potting waveform...

adding event lines

adding event lines

adding event lines

read new file conte

 (working) Getting
Theshold= 55.0
min. silence sample
min. chunk samples=
```

Ideal Functionality

- Improved accuracy (more training data).
- Stutter type classification
- Stutter event timestamping (using sequence to sequence prediction).
- Improved flexibility in the GUI.
- Improved computation time (approaching real-time)



How COVID-19 has Affected our Capstone

- Delay in receiving data to train the machine learning
  - Due to closures, our contact was unable to send us more audio samples in time for us to parse them and train the AI.
- Resistance in collecting own data
- Loss of access to in-person resources
- Inability to meet in person hindered collaboration
- Difficulty acclimating to new workflow
- Negative emotional health outcomes caused by stress

Test Plan / Demonstration

Pre-recorded Demo (Stutter):

1. Upload sample of stuttered speech (193_3, 218_10, 217_13) (do all three models)
2. Play file to demonstrate stuttered speech
3. Analyse speech to output classification
4. Show user spectrogram

Pre-recorded Demo (No Stutter):

1. Upload sample of non-stuttered speech(010_3, 017_10, 111_13) (do for all three models)
2. Play file to demonstrate non-stuttered speech
3. Analyse speech to output classification
4. Show user spectrogram

User Input Demo:

1. Record user input for speech (for all three sentences to test all three models)
2. Play file
3. Analyse speech to output classification
4. Show user spectrogram