CS6300 Speech Technology: Assignment 2 Report

-**Team 7:** Akash Reddy A, EE17B001 and Nikhil Mattapally, EE17B138

1 Convolving Sinusoids

With $\omega_0 = 5$ Hz and $\omega_1 = 10$ Hz, sampled sinusoids $\sin \omega_0 nT, \sin \omega_1 nT$ are generated with 512 samples each.

$T = 1/50$ s has been selected so as to ensure a higher sampling frequency than the Nyquist rate. It is also a sampling frequency that is high enough to allow us to see the sinusoidal nature of the signals when plotted:

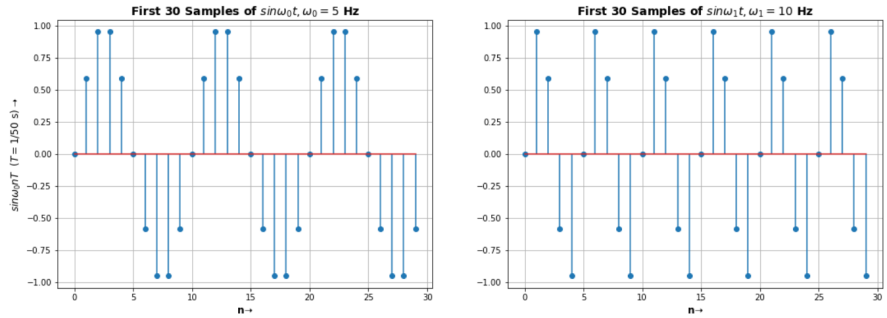

Figure 1: Sinusoids in the time-domain

The convolved signal has $N_1 + N_2 - 1 = 512 + 512 - 1 = 1023$ points. It is displayed below and also appears to be periodic:

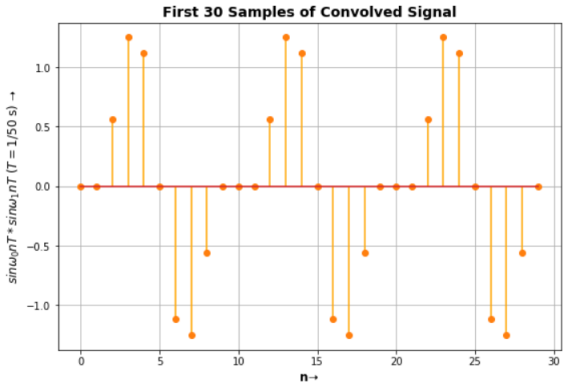

Figure 2: Convolved signal in the time-domain

The DFTs of the sinusoids are displayed below. Each of them have 512 points, since the time-domain signal also has 512 points.

The product of these DFTs is still 512-point. It is compared with the 1023-point DFT of the convolved signal (the difference in number of points can be observed in the x-axis limits). However, the DFT itself is qualitatively similar with similarly spaced pulses of similar heights.

This confirms the theory that convolution in the time-domain is equivalent to multiplication in the frequency domain.

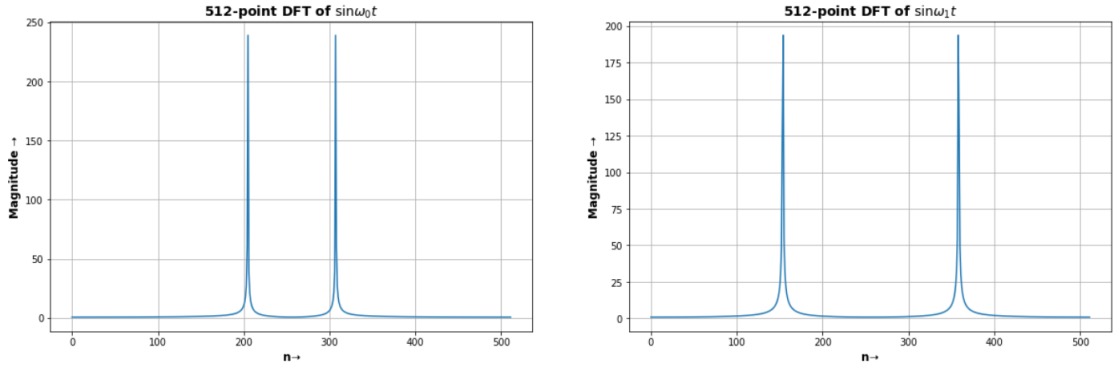

Figure 3: DFT of Sinusoids

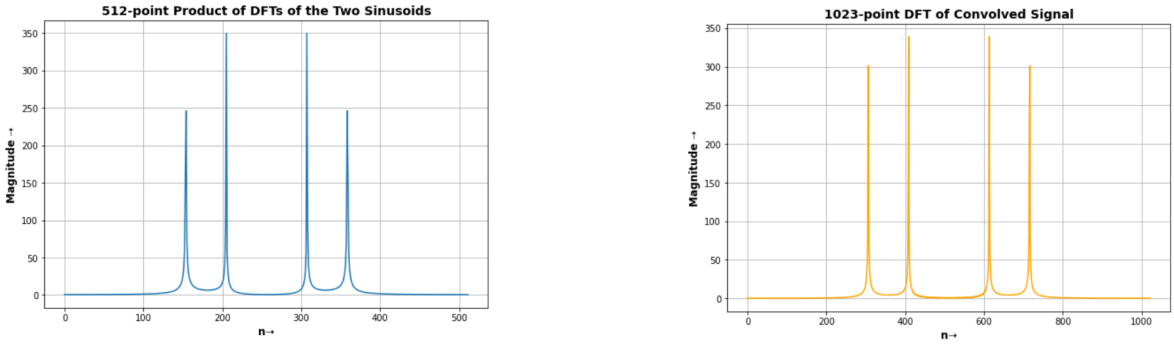

Figure 4: Product DFT of Sinusoids vs DFT of Convolved Signal

2 Isolated Vowel DFT and Formants

The recorded .wav files of the isolated vowels are read in Python using `scipy.io.wavfile.read`, which fills the time-series data of the audio into a Numpy array. Then, the DFTs of the vowels are generated. They are of different lengths because the audio files themselves have varying number of samples. We can observe that the frequency peak regions (formants) are in different regions for different vowels.

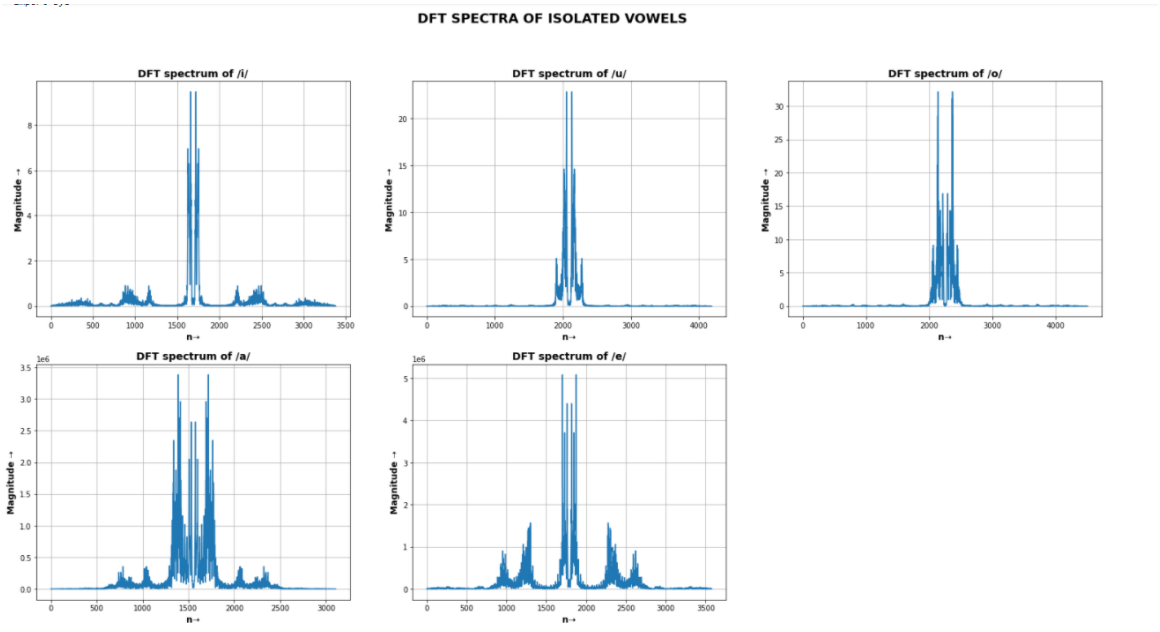

Figure 5: DFTs of Isolated Vowels

We can use Wavesurfer's formant plots to dive deeper into the analysis of these formants.

Upon extracting the formant data and calculating average values of first and second formants (F1 and F2), we plot F1 vs F2. The vowels should form different groups at the vertices of a triangle as discussed in class.

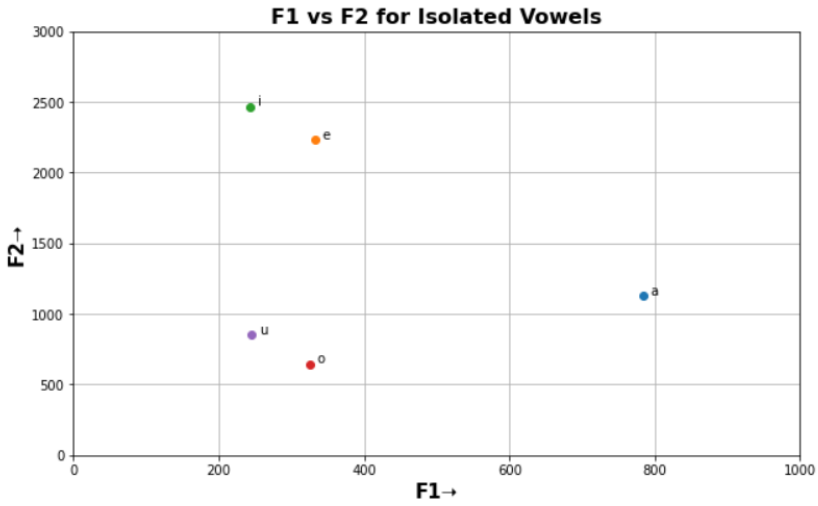

Figure 6: F1 vs F2 for Isolated Vowels

- /a/ has a high F1 and medium F2, hence it lies in the right vertex of the triangle that is formed.
- /e/, /i/ have a low F1 and high F2. They are constricted vowels that are also high-pitched in sound. Hence, they form the upper vertex of this triangle.
- Finally, /o/ and /u/ have both low formants. They are the deeper sounding vowels, so they naturally form the lower vertex of the triangle.

3 Consonant Formants (UV-UA, UV-A, V-UA, V-A)

In all types of consonants, the consonant sound at the beginning produces different frequency peaks right at the beginning of the sound. This leads to changes in the average formant frequencies, as we shall see in the F1 vs F2 plots.

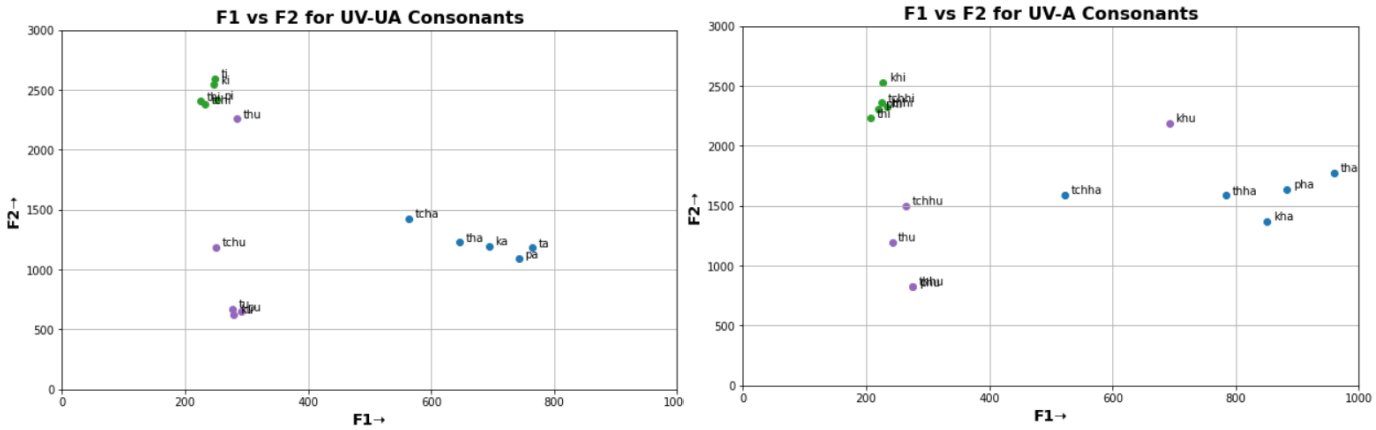

Figure 7: F1 vs F2 for Unvoiced Consonants (Unaspirated and Aspirated)

We can make a few observations from the plots that are different from the isolated vowels F1 vs F2 plot:

- In general, various consonants prefixed to the vowels cause them to get scattered in various directions the F1 vs F2 plot.

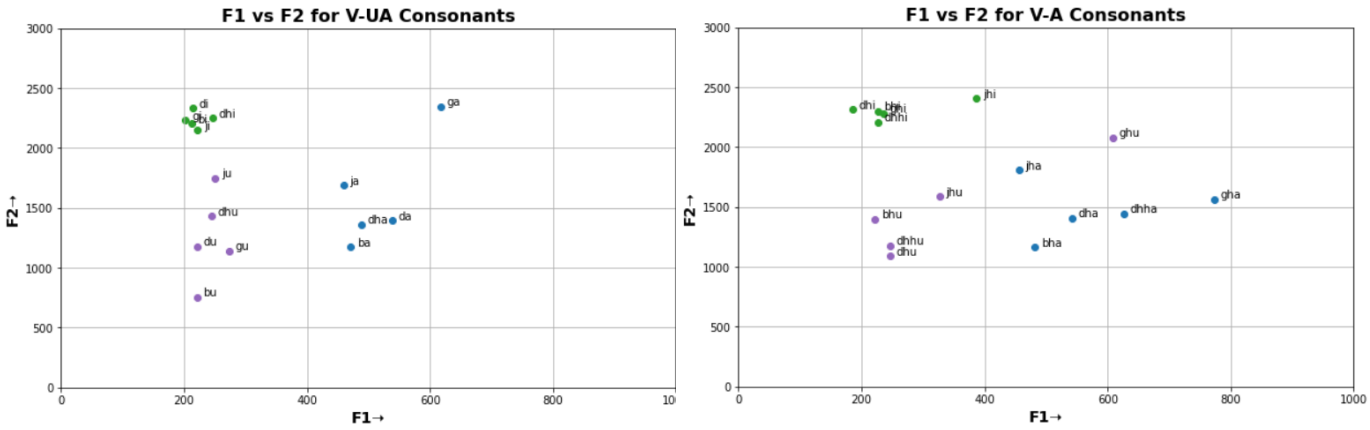

Figure 8: F1 vs F2 for Voiced Consonants (Unaspirated and Aspirated)

- The scattering for the vowel /i/ is lesser, and the scattering for the vowel /a/ seems to be the most.
- The voiced consonants tend to have a lower F1 than the unvoiced ones. Upon examining the waveforms closely, we observed that the voiced consonants are quite similar to the unvoiced equivalents, except there is a small duration where the voice has to start up before it can pronounce the consonant. In this region, the voice is low amplitude and low frequency. (Plus, when we played the audio after skipping this voice region, the recording sounded like the unvoiced equivalent! Therefore, this initiation of voice is instrumental to voiced consonants.)
  This is probably the factor that brings down the F1 frequency in voiced consonants.

4 Diphthongs - Spectrograms

To begin with, the spectrograms of the individual vowels are shown below, followed by those of the /ai/ diphthong (with window size = 1 pitch period and ≈2.8 pitch periods) are shown below.

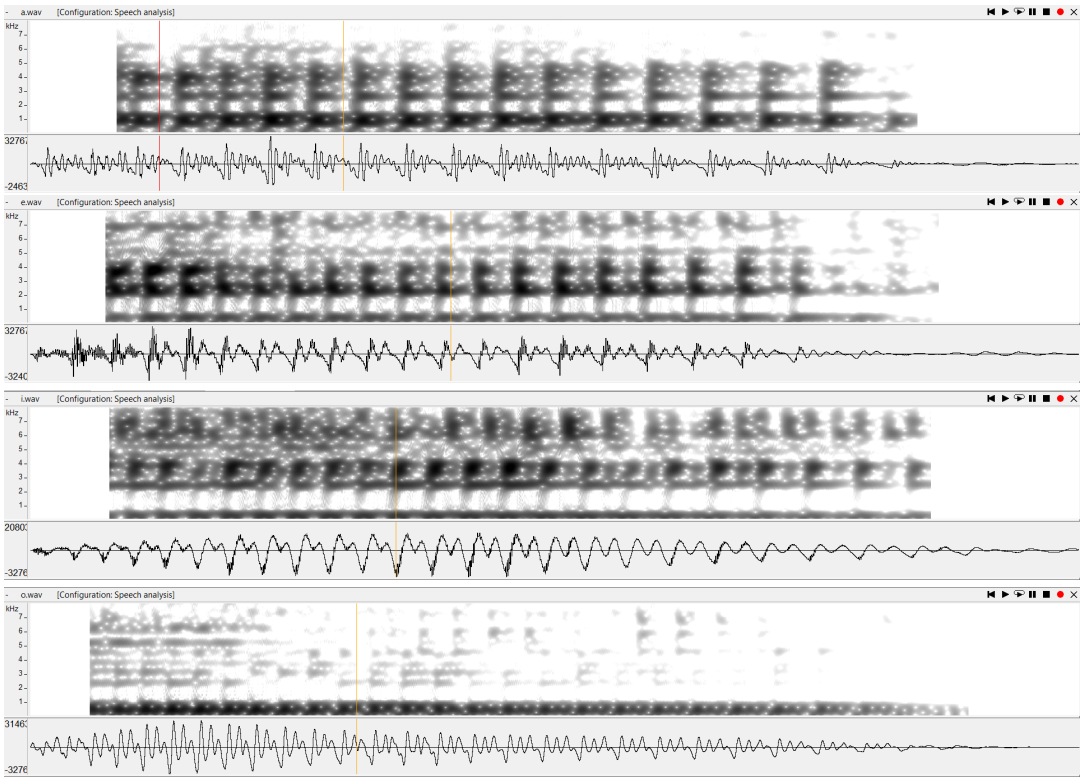

Figure 9: Spectrograms of /a/,/e/,/i/,/o/ respectively

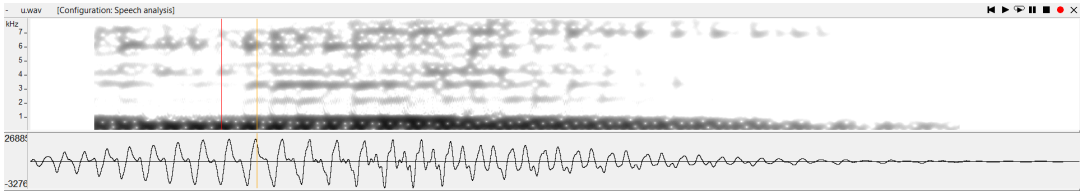

Figure 10: Spectrogram of /u/

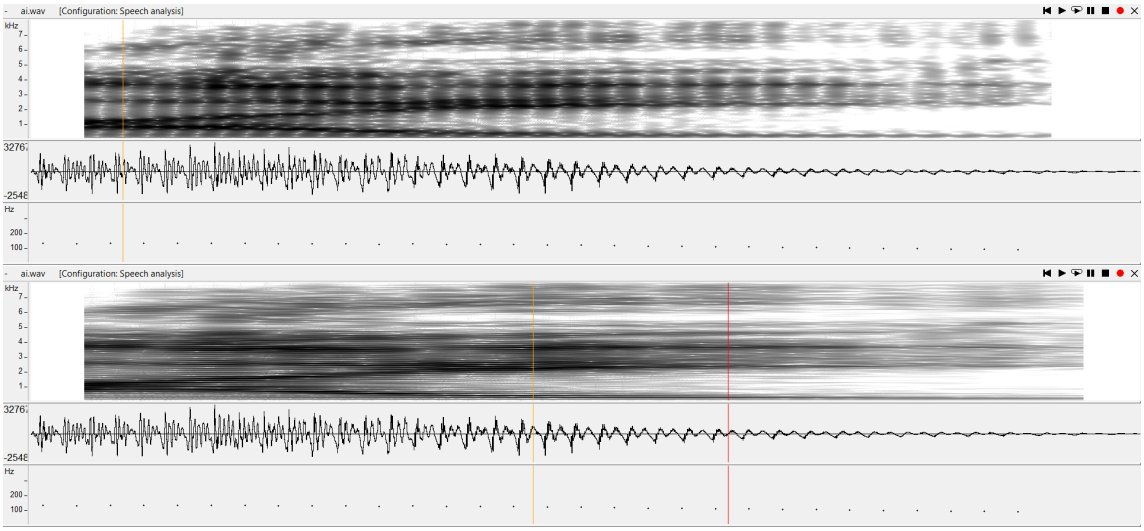

Figure 11: Spectrograms of /ai/ with the different window sizes

- The spectrogram of /ai/ is firstly different from the vowels in that the formants change drastically through their duration, whereas those of the vowels remain nearly constant throughout the duration of the vowel.
- More specifically, we can see that the /ai/ spectrogram starts off like the spectrogram of /a/. F1 and F2 then move farther apart from each other (giving lower F1 and higher F2). Towards the end, the spectrogram looks like that of /e/ or /i/. This is expected because /ai/ is a concatenation of /a/ and /i/.
- When we increase the window size to more than two pitch periods, we observe that the spectrogram loses resolution and gets very blurry.

Similar observations can be made for the /ao/ diphthong, which is a combination of /a/ and /o/. It starts off similar to /a/ once again and looks like /o/ towards the end.

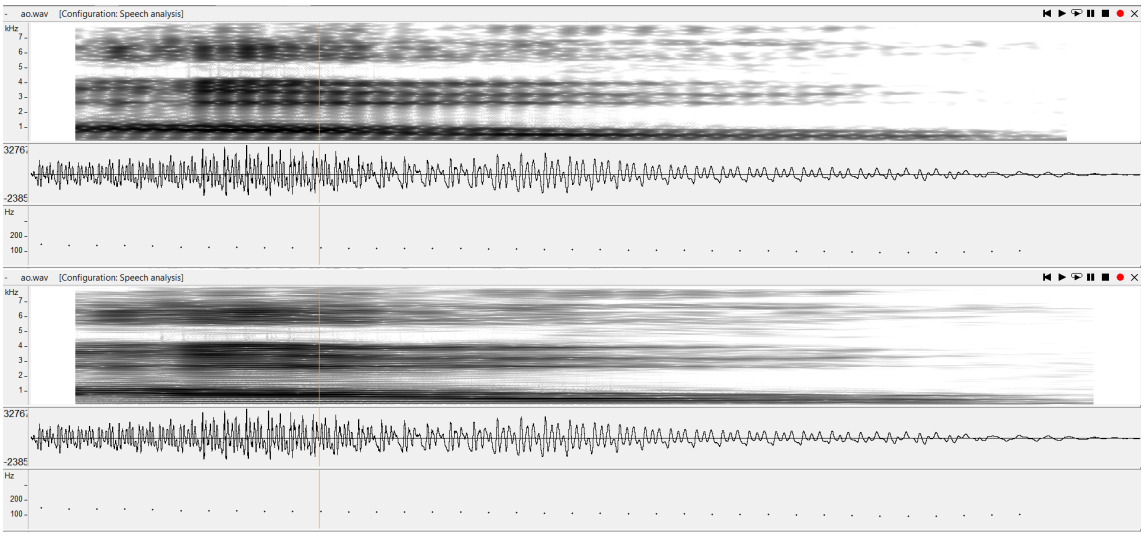

Figure 12: Spectrograms of /ao/ with the different window sizes

5 Fricatives - Spectrograms

The spectrograms of /sa/ and /ha/ are shown below. They have been obtained using the default window size of 64 samples.

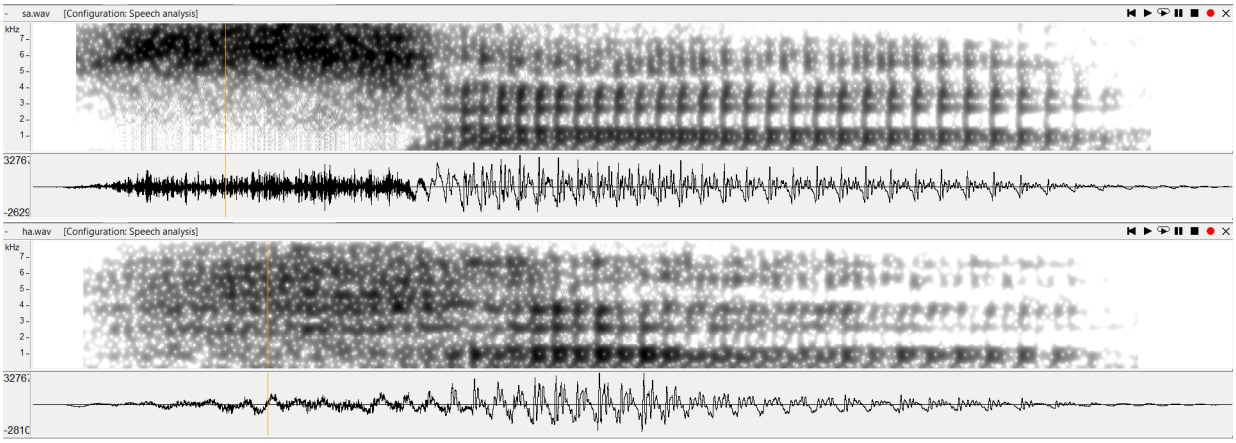

Figure 13: Spectrograms of /sa/ and /ha/

- Both spectrograms end similar to /a/ as expected.
- However, the initiations of the two sounds are different. /sa/ has much higher formants (frequency content) than /ha/ in the beginning. This is natural, as the /s/ sound is a much higher frequency consonant than /h/.

6 Discernible Tones through Masking Noise

The masking noise has been generated by adding up waveforms of {350, 351, 352, ..., 399, 400} Hz, each of fixed amplitude 0.0635 (or $\approx$ -24 dB) where $dB = 20 \log_{10} amp$. Setting this -24 dB as the new reference level, the new dB amplitude relative to the noise, required for a tone to be discerned, is plotted as a function of frequency below.

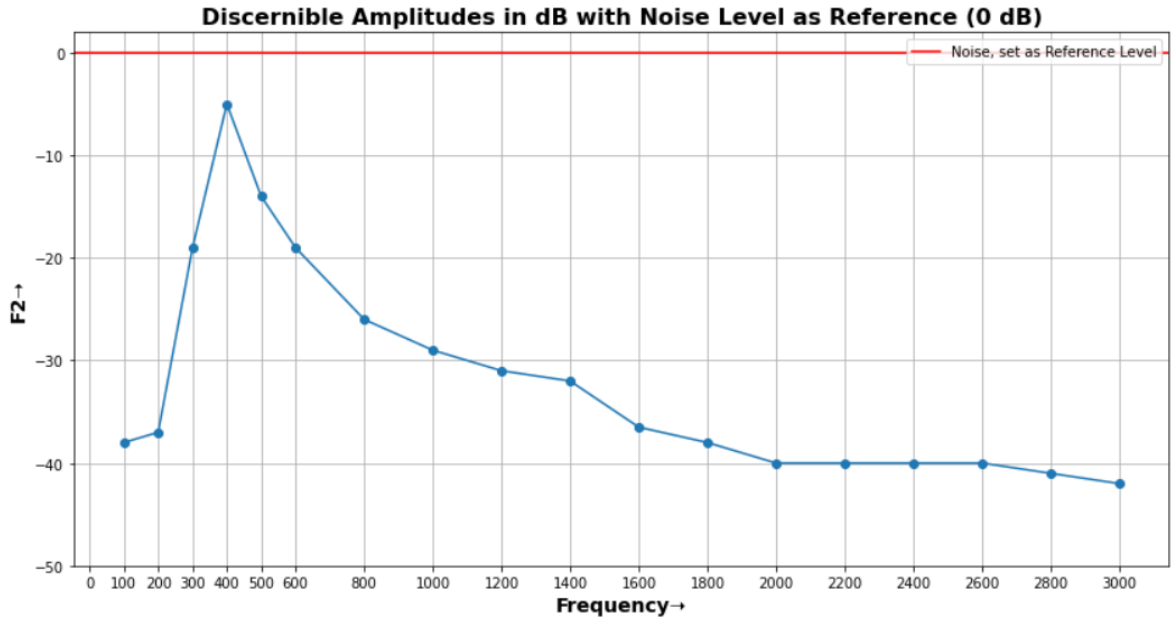

Figure 14: Discernible Amplitude (dB) vs Frequency of Tone

- As the tone frequency approaches the range of the masking noise (350-400 Hz), we need a significantly higher amplitude tone to be able to distinguish it from the masking noise.

- After 400 Hz, the minimum amplitude required drops off in an asymptotic manner as the tones become more distinguishable from the noise once again and the logarithmic distance between tones decreases. (Reason for asymptotic behaviour is that, we hear music and tones on a logarithmic scale, so equally spaced tones start sounding more and more similar as we go up in the linear scale.)

7 F1 vs F2 for Lab 1 Data (Non-Isolated) Vowels

We have taken 7 instances of each vowel from the Lab 1 sentences and calculated the average F1 and F2 values. They are plotted below.

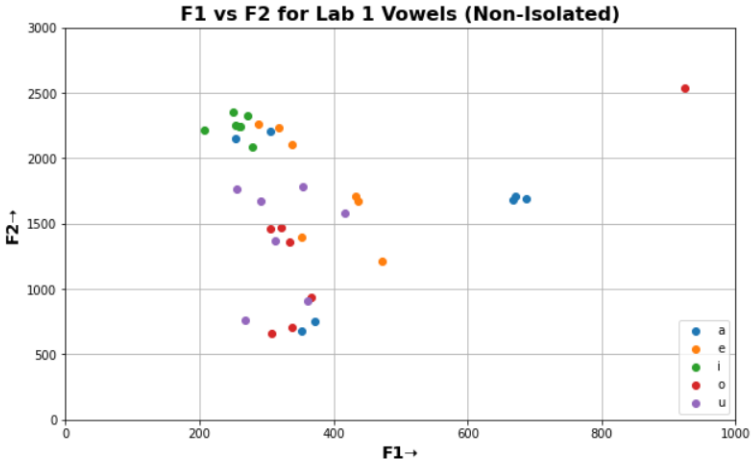

Figure 15: F1 vs F2 on Lab 1 Data vowels

- The fact that the vowels are non-isolated creates a lot of noisy frequency formants. The residues of the prefixed or suffixed consonants contribute to this noise. Besides, the pronunciation of the vowel itself varies slightly based on the word and the context. For example, the vowel /a/ is pronounced audibly differently in the words "pursue" and "father". But if we were recording just isolated vowels, all instances of /a/ would sound more similar.
- Despite the noise from surrounding consonants, we can still observe that the same triangular trend as in Question 2 is followed to an extent. Most /a/ are at the right vertex, most /e/ and /i/ are at the top-left vertex, and most /o/ and /u/ are at the bottom-left vertex.

Obstacles and Learnings

- I had first recorded taken the time to record all my audio files in Audacity. I exported them and tried to get my formants in Wavesurfer, but the formants were not showing up (F1 was always a default constant value of 500, F2 always 1500, and so on). Upon reaching out to Saish, he suggested that we use someone else's audio files and mention their name. However, it seemed like most students hadn't recorded all consonants.
  Just as I was about to do this anyway, I opened an Audacity waveform in Wavesurfer and simply normalised it to 0 dB, and formants suddenly appeared. The waveform suddenly turned "Wavesurfer-compatible", somehow. I am not sure why this happened, but I had to redo all the .frm files.
- When I used the label tracks to automatically extract the vowel pieces from Lab 1 Data, I found that they were not accurate whatever I did. After a lot of exploration, I realised that all the sliced audio was just sliced a few milliseconds before the expected area. Upon further digging, I found that Audacity has a Default Latency Compensation of -130 ms. So every recorded audio started 130 ms earlier. However I had obtained the label files with 0 ms as the reference, therefore there was a slight offset which I had to correct and re-obtain all the label files.
- Both were unforeseen obstacles that caused hours of delay, however we now understand better how Audacity and Wavesurfer work together and how to efficiently proceed with future assignments.