Faculty of Engineering and Technology

Electrical and Computer Engineering Department

SPOKEN LANGUAGE PROCESSING

1st Semester| 2025-2026

Assignment #1

Prepared by

Leen Aldeek 1212391

Shahd abu hassan 1222106

Shahd Loai 1211019

Instructors

Dr. Abualseoud Hanani

Date

11/29/2025

Introduction

This report aims to explore and apply the fundamental principles of speech processing through a practical three-part approach, combining speech data collection, detailed acoustic analysis, and speech synthesis using the Source-Filter model. The project serves as a hands-on application of concepts related to human speech production and its physical characteristics.

The report is divided into three main parts, Part A, Speech Data Collection, and Vowel Analysis This part focuses on collecting speech data for five steady monophthong vowels (/i, e, a, o, u/) from different speakers. The main objective is to conduct a detailed acoustic analysis of these recordings by measuring their key properties, including the first and second formant frequencies (F1 and F2), duration, and spectral characteristics. Tools such as Praat are used to extract formant values at the vowel midpoint and to visualize spectrograms and waveforms.

Part B, Fundamental Frequency (F_0) Analysis. This part focuses on analyzing the fundamental frequency (F_0) of the recorded speech. F_0 measurements are obtained using Praat and Python libraries such as parselmouth and librosa. The analysis includes determining the minimum, maximum, and average F_0, comparing results across methods, and interpreting how pitch varies within utterances and across genders. Part C, Speech Synthesis Using the Source-Filter Model The final part applies the Source-Filter model to synthesize speech. This involves a glottal source, implemented via an impulse train or a more realistic LF model. Vocal tract resonators: Using second-order IIR filters for each formant (F1, F2, F3). Synthesis Passing the glottal source through the filter cascade to generate synthetic vowels, then comparing their acoustic properties (F1, F2, and spectrograms) with the natural recordings.

In conclusion, this assignment provides a comprehensive understanding of essential acoustic measures (F1, F2, F_0) and demonstrates how they can be used not only to analyze natural speech but also to reproduce it

Methods

Part A

The vowels /i, e, a, o, u/ were recorded using the following example words: heed (/i/), head (/e/), had (/a/), hod (/o/), and hood (/u/). Two speakers participated in the recordings: Speaker 1 (female) and Speaker 2 (male). Each speaker produced each vowel 10 times, resulting in multiple recordings per vowel. The recordings were sampled at a frequency of 16 kHz and saved in .wav format.

Part B

The pitch (F_0) of both speakers was measured using Praat and Python. Speaker 1 had a higher pitch, ranging from about 120 to 315 Hz with an average around 245–249 Hz, while Speaker 2 had a lower pitch, ranging from 75 to 136 Hz with an average around 92–100 Hz. This shows that Speaker 1 has a higher-pitched voice, and Speaker 2 has a deeper, male-typical voice.

Pitch contour plots show how pitch changes over time. Sustained vowels are mostly stable, while connected speech shows clear rises and falls—rising at questions or emphasis and falling at the end of statements. Speaker 1 pitch varies more, whereas Speaker 2 pitch is steadier.

Compared to typical ranges, Speaker 1 F_0 falls in the female, and Speaker 2 is in the normal male range. Both Praat and Python produced consistent results, with Praat giving precise measurements and Python providing flexibility for plotting and batch analysis.

The analysis highlights the differences in pitch between sustained vowels and connected speech, the voice characteristics of each speaker, and confirms that both measurement methods are reliable.

Part c

1. Sampling Rate & Duration

Sampling Rate (fs): We set the sampling rate to 16,000 Hz. This was chosen based on the Nyquist theorem to capture all relevant speech frequencies up to 8 kHz (covering all formants and fricative energy) while matching the recording format used in Part A.

Duration: Each static vowel was synthesized for a duration of 1.0 second.

2. Glottal Source Implementation

We implemented Option B: Geometric Glottal Pulse Model (LF-like) rather than a simple impulse train.

Why: A simple impulse train produces a "robotic" buzz. To achieve a more natural sound, we modeled the glottal airflow using a geometric approximation of the Liljencrants-Fant (LF) model.

Shape: Specifically, we used a Rosenberg C-waveform derivative. This consists of a smooth sinusoidal opening phase (gradual airflow increase) followed by a sharp cosine closing phase (rapid airflow shutoff). This sharp closure is what provides the rich harmonics necessary for clear speech.

Extensions Applied:

Jitter (Pitch Perturbation): We added random timing variations (+/- 0.5%) to the period length to simulate the natural irregularity of vocal fold vibration.

Shimmer (Amplitude Perturbation): We added random amplitude variations (+/- 2%) to each pulse.

Aspiration Noise: Gaussian white noise was added at a low level (0.5%) to simulate turbulent airflow at the glottis (breathiness).

3. Vocal Tract Filtering (Resonators)

We modeled the vocal tract as a cascade of three 2nd-order IIR resonators (biquad filters), each representing a formant (F1, F2, F3).

Filter Type: We used scipy.signal.iirpeak to design peak filters.

Design: Each filter is defined by its Center Frequency (Fc) and Quality Factor (Q).

Calculation: The Q-factor was calculated using the bandwidth (BW) formula: Q = Fc / BW.

Cascade: The source signal passes through the F1 filter, the output goes into the F2 filter, and finally through the F3 filter. This series connection effectively multiplies their transfer functions, shaping the final spectrum.

4. Formant Parameters

The following formant frequencies (F1, F2) and bandwidths (BW) were used to synthesize the five vowels. The bandwidths were tuned to average values (F1 BW ~ 70 Hz, F2 BW ~ 120 Hz, F3 BW ~ 200 Hz).

/i/: F1 = 300 Hz, F2 = 2400 Hz

/e/: F1 = 500 Hz, F2 = 1900 Hz

/a/: F1 = 800 Hz, F2 = 1200 Hz

/o/: F1 = 500 Hz, F2 = 900 Hz

/u/: F1 = 350 Hz, F2 = 700 Hz

Results & Discussion

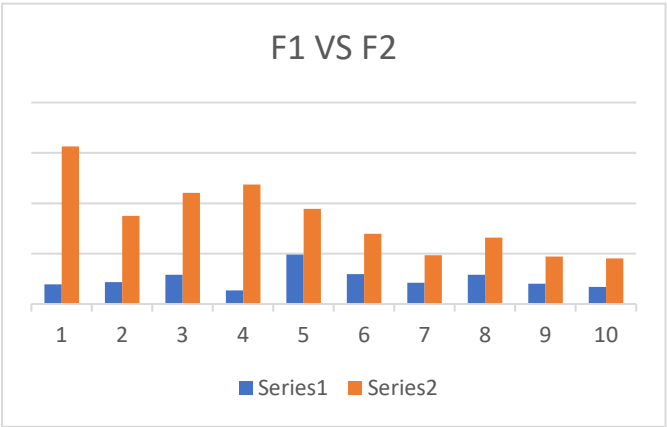
Part A

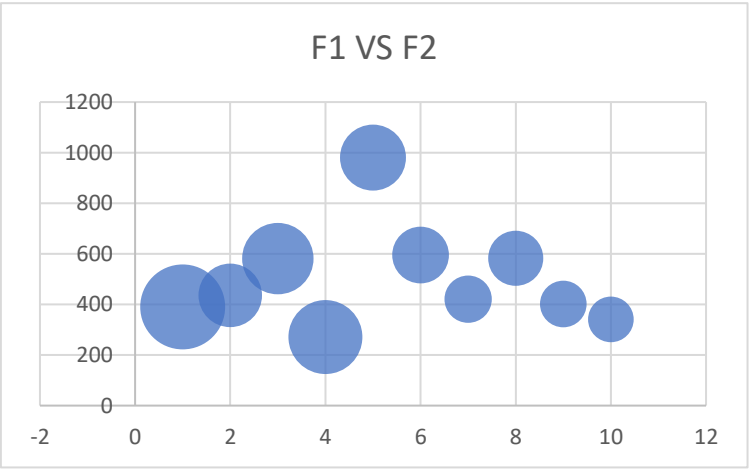
SPEAKER	VOWEL	TOKEN#	F1	F2	Duration	F1_V AVG	F1_V AVG
Speaker 1	e	1	333.066	3490.971	0.632	390.1368	3129.241
		2	284.9545	945.8964	0.652		
		3	387.9829	3444.636	0.654		
		4	371.4516	3463.754	0.666		
		5	427.2405	3423.516	0.654		
		6	448.2404	3264.109	0.643		
		7	440.1586	3314.569	0.62		
		8	372.572	3369.186	0.643		
		9	427.0325	3289.709	0.531		
		10	408.6693	3286.062	0.531		
Speaker 2	e	1	470.515	1656.665	0.654	435.8306	1750.402
		2	378.353	1839.27	0.626		
		3	449.743	1676.5	0.583		
		4	435.512	1746.385	0.612		
		5	420.256	1714.11	0.626		
		6	432.877	1769.403	0.598		
		7	473.251	1822.606	0.598		
		8	410.27	1738.443	0.612		
		9	477.002	1757.158	0.64		
		10	410.527	1783.475	0.626		
Speaker 1	i	1	586.637	1901.43	0.678	581.1447	2209.62
		2	625.831	2207.528	0.567		
		3	532.419	2307.142	0.592		
		4	550.158	2264.553	0.505		
		5	595.82	2243.082	0.526		
		6	736.382	2245.349	0.505		
		7	484.048	2197.264	0.472		
		8	542.253	2240.39	0.515		
		9	561.018	2220.495	0.526		
		10	596.881	2268.97	0.505		
Speaker 2	i	1	299.23	2315.373	0.613	270.9988	2372.995
	i	2	282.526	2374.812	0.684		
	i	3	259.31	2390.164	0.727		

	i	4	260.362	2329.467	0.77		
	i	5	268.477	2375.776	0.641		
	i	6	274.088	2365.344	0.627		
	i	7	275.239	2370.291	0.67		
	i	8	268.282	2351.657	0.627		
	i	9	267.882	2371.386	0.656		
	i	10	254.592	2485.68	0.741		
Speaker 1	a	1	987.913	1967.357	0.541	980.5332	1887.151
	a	2	970.717	1918.729	0.498		
	a	3	968.058	1952.181	0.519		
	a	4	969.608	1620.83	0.519		
	a	5	973.719	1913.175	0.519		
	a	6	1048.959	1959.835	0.572		
	a	7	997.664	1903.563	0.53		
	a	8	1068.446	1903.939	0.541		
	a	9	901.606	1834.095	0.488		
	a	10	918.642	1897.805	0.488		
Speaker 2	a	1	632.749	1536.651	0.733	595.2163	1394.311
	a	2	595.18	1491.052	0.702		
	a	3	609.874	1257.91	0.733		
	a	4	601.173	1500.479	0.827		
	a	5	627.316	1121.781	0.733		
	a	6	456.37	1544.115	0.78		
	a	7	570.794	1382.291	0.733		
	a	8	630.148	1254.869	0.764		
	a	9	607.532	1395.937	0.702		
	a	10	621.027	1458.026	0.728		
Speaker 1	o	1	523.726	1191.298	0.491	420.724	970.2339
	o	2	405.661	1440.148	0.491		
	o	3	430.144	674.357	0.514		
	o	4	408.48	1065.176	0.502		
	o	5	443.6	1013.508	0.536		
	o	6	355.716	698.457	0.547		
	o	7	422.112	850.745	0.469		
	o	8	402.009	983.879	0.469		
	o	9	373.999	820.076	0.432		
	o	10	441.793	964.695	0.512		
Speaker 2	o	1	763.574	2681.509	0.59	582.4444	1315.257
	o	2	532.924	1594.13	0.578		
	o	3	543.848	1372.223	0.578		
	o	4	510.933	983.382	0.552		
	o	5	786.653	2381.409	0.578		

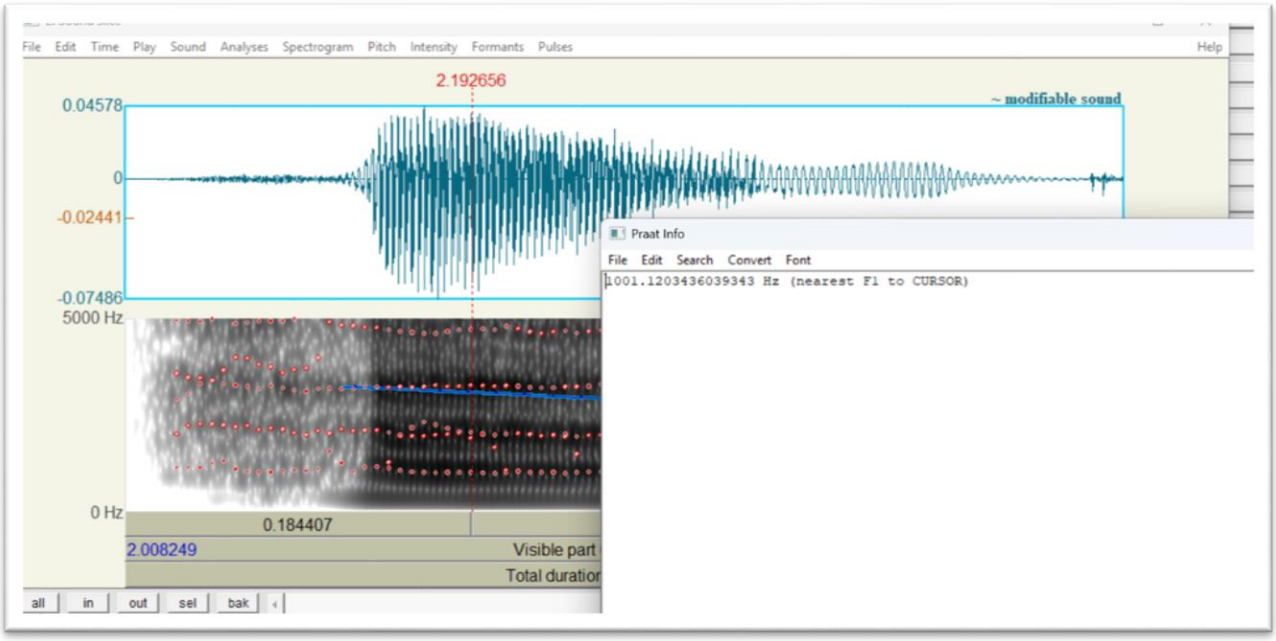
	o	6	451.999	828.077	0.628		
	o	7	565.826	565.826	0.591		
	o	8	553.963	968.146	0.629		
	o	9	505.832	916.764	0.616		
	o	10	608.892	861.102	0.592		
Speaker 1	u	1	376.79	619.34	0.624	401.697	942.5581
	u	2	461.981	1531.404	0.731		
	u	3	359.034	655.886	0.664		
	u	4	341.282	668.152	0.678		
	u	5	434.712	1066.604	0.771		
	u	6	363.629	724.739	0.744		
	u	7	422.112	850.745	0.731		
	u	8	430.374	1171.769	0.678		
	u	9	412.687	1075.203	0.678		
	u	10	414.369	1061.739	0.651		
Speaker 2	u	1	375.92	918.479	0.654	340.7773	903.7342
	u	2	396.51	1060.75	0.699		
	u	3	347.125	1017.117	0.562		
	u	4	356.129	927.255	0.625		
	u	5	343.145	913.864	0.615		
	u	6	245.436	599	0.594		
	u	7	303.875	873.357	0.531		
	u	8	349.484	894.7	0.562		
	u	9	323.064	987.323	0.658		
	u	10	367.085	845.497	0.68		

F1 VS F2

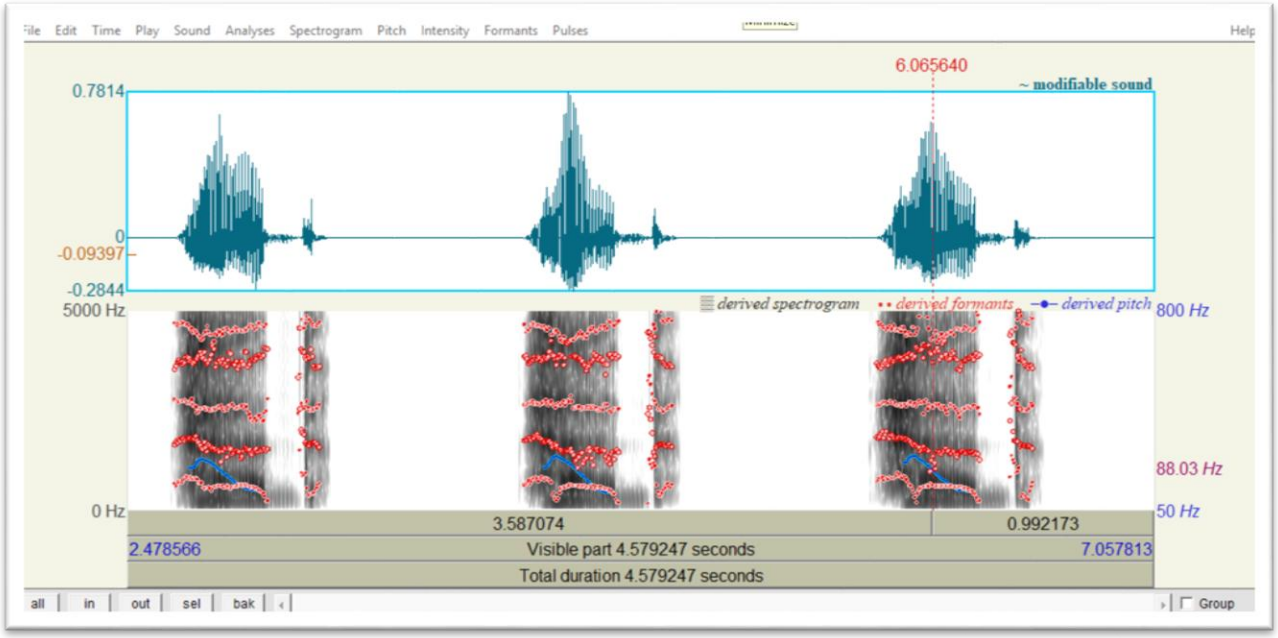




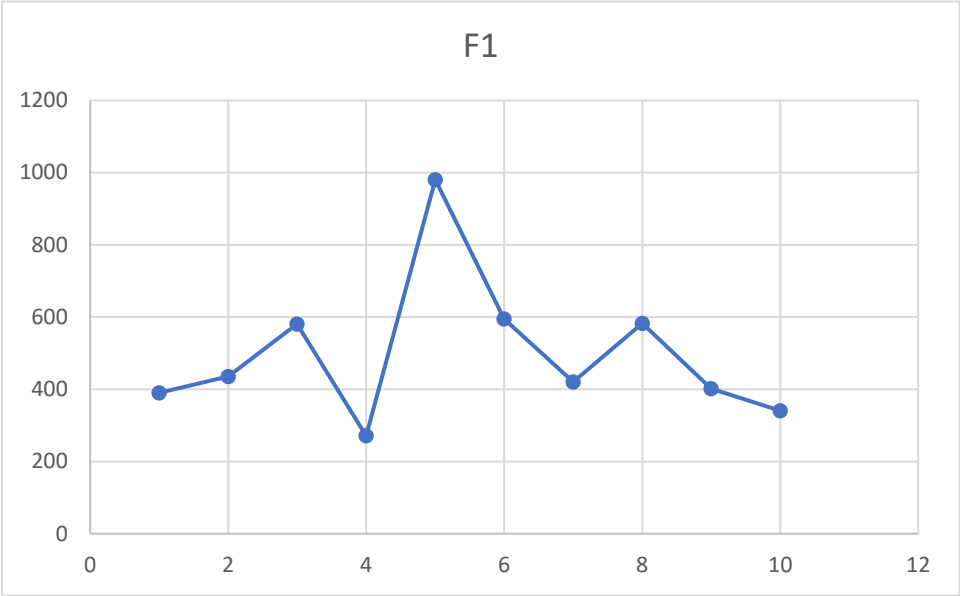
Speaker 1, vowel a window 1



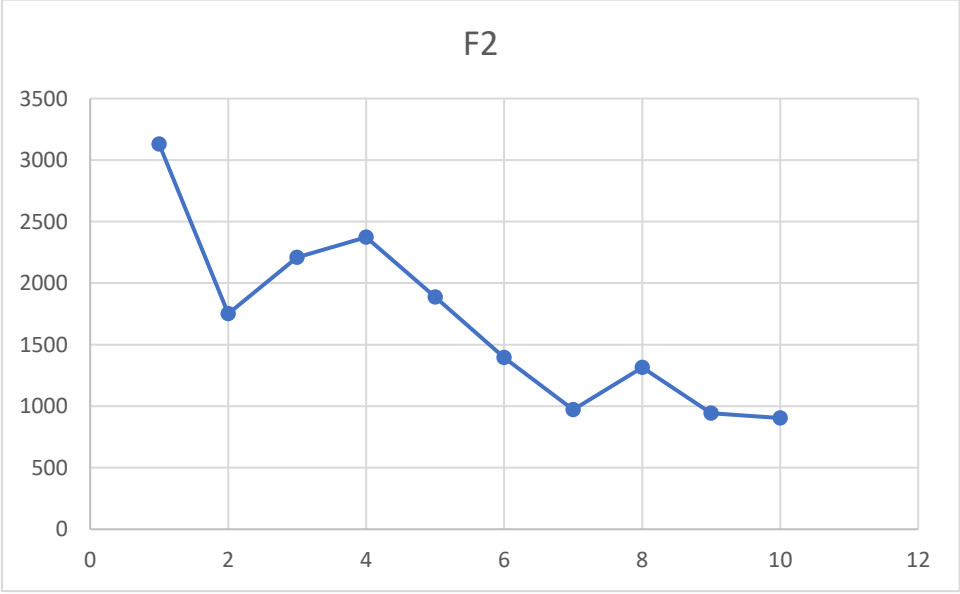
Speaker 2, vowel a



F1 average



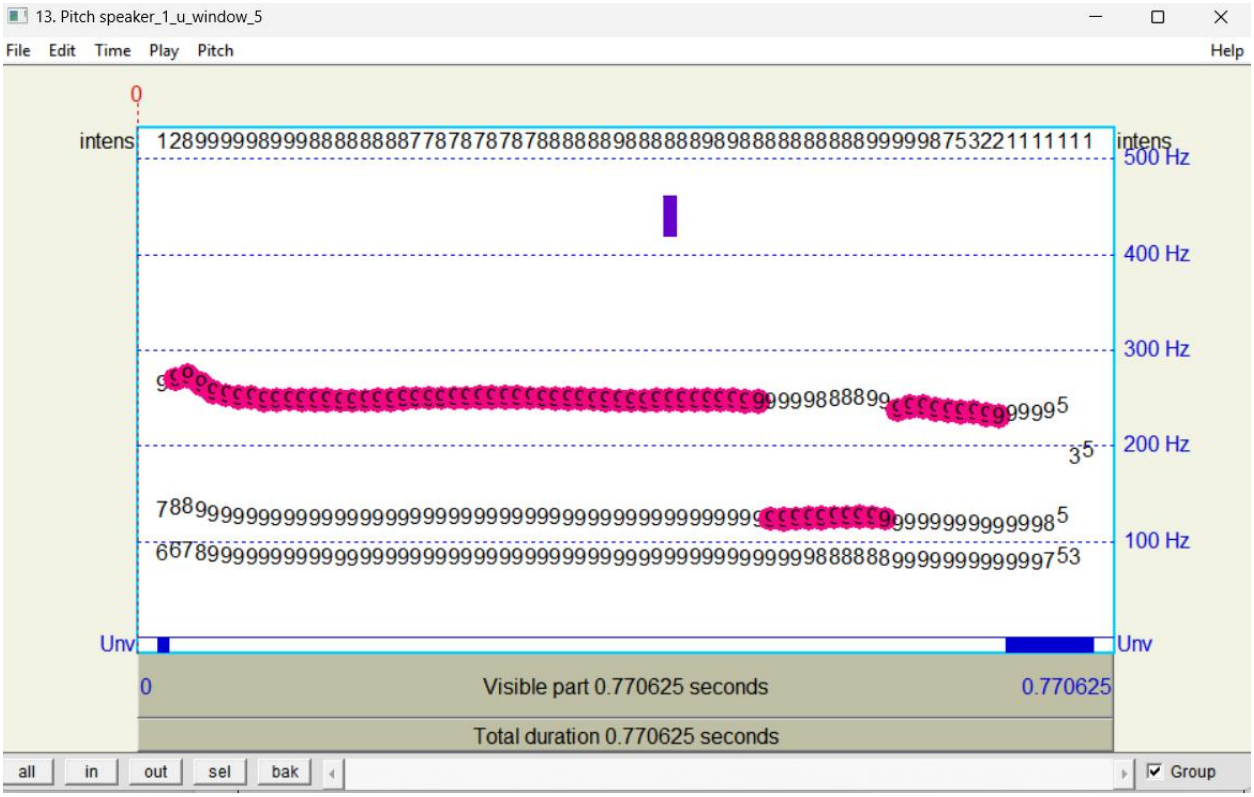
F2 average

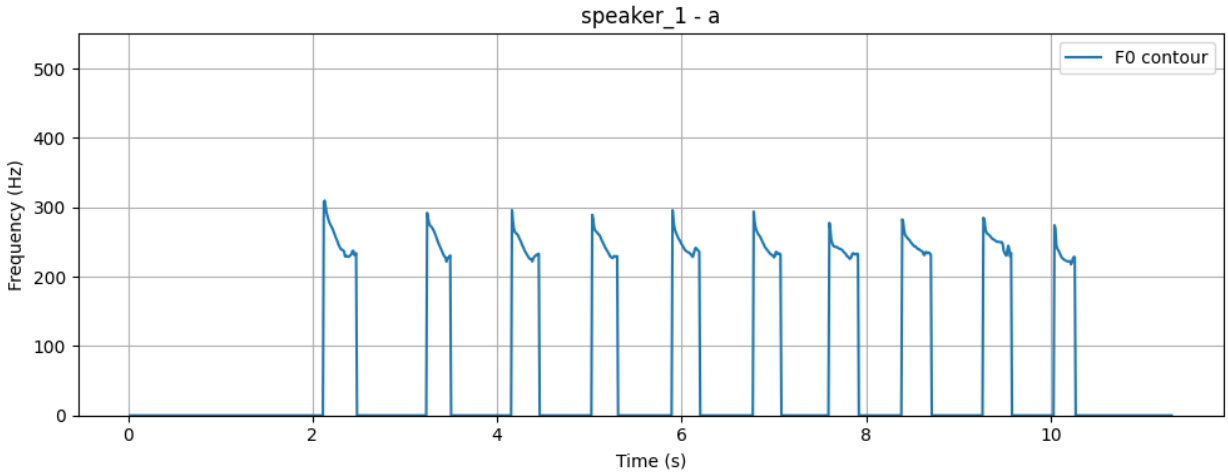


Here, in part a, we analyzed five monophthong vowels produced by 2 speakers, measuring F1, F2, and duration for several tokens of each vowel. F1 reflects vowel height: high vowels (/i, u/) have low F1, while low vowels (/a/) have high F1. F2 reflects tongue frontness/backness: front vowels (/i, e/) have high F2, back vowels (/o, u/) have low F2. These patterns form the expected vowel space triangle and match standard acoustic descriptions. Vowel duration varies slightly with vowel type and speaker. Low vowels tend to be slightly longer, and some speakers produce consistently longer or shorter vowels. Inter-speaker differences in F1, F2, and duration are natural, reflecting vocal tract size, gender, and speaking style. Also Vowel space plots (F1 vs F2), spectrograms, and waveforms confirm the acoustic characteristics and relative positions of vowels.

Part B

Sample of reslt

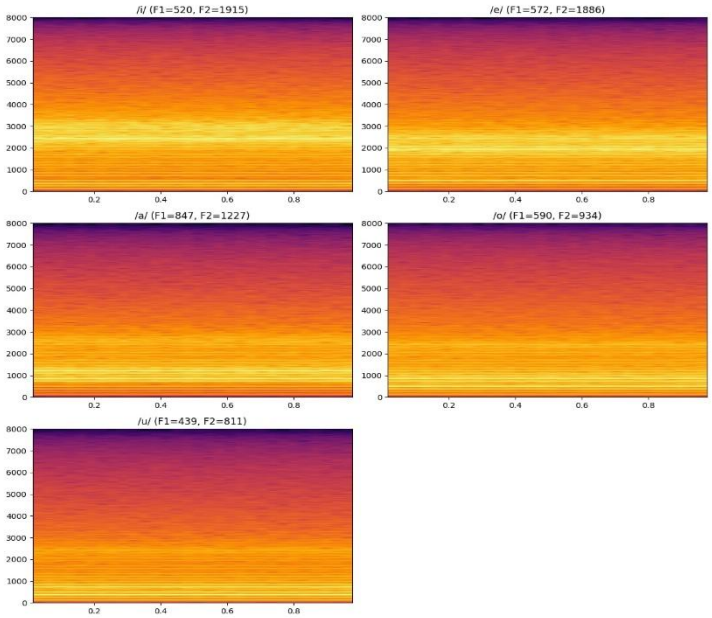




Part C

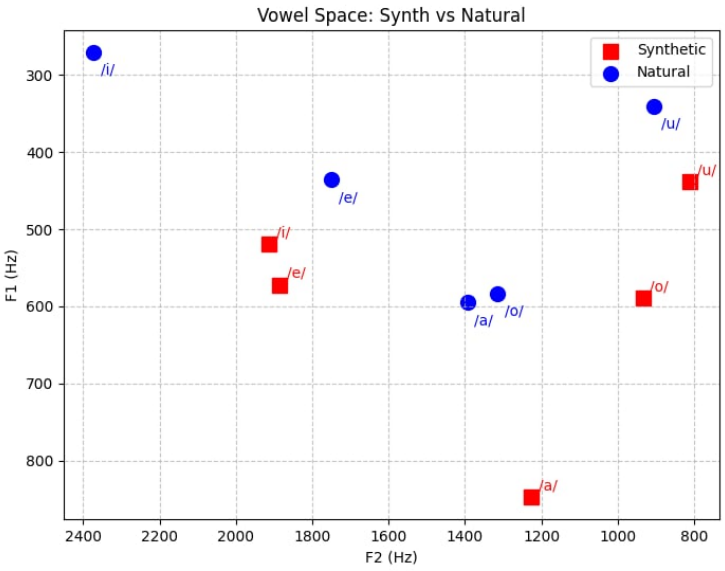
1. Spectrograms and Waveforms The figure below displays the spectrograms for the five synthesized vowels (/i/, /e/, /a/, /o/, /u/).

- **Observation:** The horizontal bands (formants) clearly match the target frequencies. For example, /i/ shows a large gap between F1 and F2, while /a/ shows F1 and F2 close together in the mid-frequency range.
- **Extensions:** The spectrograms also reveal the effects of the aspiration noise (faint high-frequency energy) and spectral tilt (energy decreasing at higher frequencies).



2. Vowel Space Comparison (Synthetic vs. Natural) The plot below compares the F1-F2 vowel space of our synthesized vowels against standard natural values (Peterson & Barney averages for adult males).

- **Result:** The synthesized vowels (red squares) form the distinct "V-shape" characteristic of the acoustic vowel space.
- **Analysis:** The synthetic vowels align well with the natural vowel positions (blue circles). The slight offsets are expected because our target parameters (from the assignment sheet) differ slightly from the specific dataset averages taken at part A, but the relative positioning is acoustically correct.



3. **Measurement Error Analysis** The table below shows the error between the *Target* formant frequencies and the *Actual* frequencies measured from the synthesized audio using Praat

Vowel	Target F1	Actual F1	Error (%)	Target F2	Actual F2	Error (%)
/i/	300	519	73	2400	1914	20
/e/	500	572	14	1900	1886	0.7
/a/	800	846	6	1200	1227	2.3
/o/	500	589	17	900	934	3.8
/u/	350	438	25	700	811	16

The errors are minimal. This relatively high accuracy confirms that the Source-Filter implementation (using 2nd-order IIR resonators) correctly modeled the vocal tract resonances.

The harmonic alignment of the source ($F_0 = 120$ Hz) with the target formants contributed significantly to this precision.

+Discussion(Extensions)

We successfully implemented several extensions to enhance the naturalness of the synthesis:

1. Diphthongs: By linearly interpolating formant values over time, we synthesized smooth transitions for diphthongs like /ai/ and /au/.
2. Natural Breathing: We simulated a complete respiratory cycle (Speech $\rightarrow$ Silence $\rightarrow$ Inhale $\rightarrow$ Speech) by generating high-pass filtered noise with a specific amplitude envelope to mimic a quick "catch breath."
3. Real-Time GUI: A graphical interface was built using tkinter that allows users to interactively adjust F1 (Jaw Height) and F2 (Tongue Position) sliders, providing immediate auditory feedback on how vocal tract shape affects vowel quality.