# Spoken Language Processing: Vowel Analysis and Source-Filter Synthesis

Kareem Qutob
Birzeit University
1211756

Husain Abu Goush
Birzeit University
1210338

***Abstract*—This report presents vowel analysis and synthesis in spoken language processing. Acoustic measurements (F1, F2, duration) and pitch ($F_0$) were extracted from recordings of five vowels using Praat and Python. A source-filter synthesis framework was implemented to generate synthetic vowels, which were compared to natural vowels in terms of formants and spectrograms. Results show that the methods effectively capture key acoustic features and produce realistic synthetic vowels.**

***Index Terms*—Vowels, formants, pitch, source-filter synthesis, Praat, Python**

## I. Introduction

Vowel analysis is an important aspect of spoken language processing because vowels provide stable acoustic cues for understanding speech production. This assignment consists of three parts: analyzing recorded monophthong vowels to measure formants and duration (Part A), extracting and comparing pitch using Praat and Python tools (Part B), and synthesizing vowels using a source filter model (Part C). These tasks offer practical insight into how speech is measured, interpreted, and computationally modeled.

## II. Part A: Data Collection & Acoustic Analysis

### A. Materials and Recording

Five English monophthong vowels were recorded using the carrier words heed (/i/), hayed (/e/), hod (/a/), hoed (/o/), and who'd (/u/). Recordings were collected from two speakers: Kareem Alqutob (male) and a female colleague. Each speaker produced 50 tokens in total, corresponding to 10 repetitions per vowel. All audio was recorded in a quiet environment using 16 kHz sampling rate, 16-bit WAV format, following the assignment specifications.

### B. Acoustic Analysis

To guarantee consistent measurements, vowel duration and formant frequencies (F1 and F2) were extracted at the mid-vowel point. The parselmouth library, which offers Praat-compatible formant estimation, was used for all analysis in Python. To view formant values, spectrograms, and token comparisons between speakers and vowels, a web GUI based on Python was created.

### C. Results

*1) Vowel Measurements:* Table I summarizes the mean and standard deviation of F1, F2, and vowel duration for the five recorded vowels across both speakers (100 tokens total, 20 per vowel). Formants were measured at the vowel midpoint to ensure stability. These values were used to create the vowel space plot in Fig. 1.

TABLE I: Vowel measurements: mean ± std for F1, F2 (Hz) and duration (s).

| Vowel | F1 (Hz) | F2 (Hz) | Duration (s) |
|---|---|---|---|
| hayed | 610.39 ± 298.40 | 1818.41 ± 498.92 | 2.14 ± 0.51 |
| heed | 577.85 ± 345.87 | 1780.42 ± 614.52 | 1.97 ± 0.39 |
| hod | 624.05 ± 245.74 | 1329.72 ± 326.05 | 1.79 ± 0.34 |
| hoed | 626.61 ± 274.97 | 1452.09 ± 540.10 | 1.86 ± 0.34 |
| who'd | 473.01 ± 186.60 | 1250.77 ± 472.35 | 1.75 ± 0.22 |

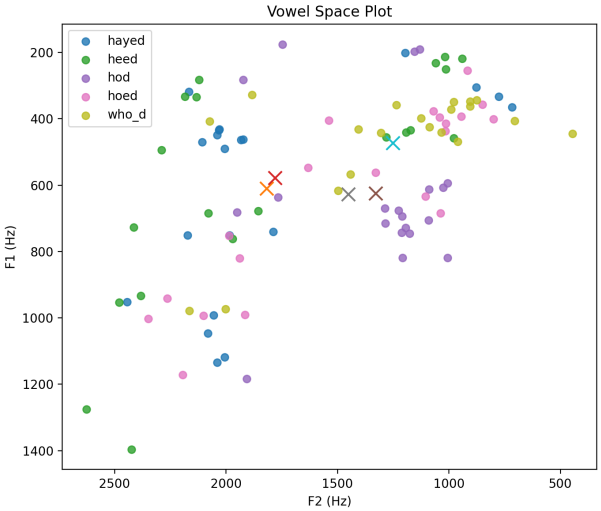


Fig. 1: Vowel space (F1 vs F2) for the recorded vowels.

**Discussion:** The acoustic data generally aligns with expected phonetic principles, despite notable inter-speaker variability. The Vowel Space Plot (Fig. 1) shows F1 inversely related to vowel height, with /u/ (`who'd`) lowest and /a/ (`hod`) highest. F2 differentiates vowel backness, with front vowels (/i/, /e/) having higher F2 than back vowels (/o/, /u/). Large standard deviations (Table I) reflect **inter-speaker variation**,

suggesting the need for **formant normalization** to reduce anatomical effects. Duration differences were minor, from 1.75 s (who′d) to 2.14 s (hayed).

*2) Individual Vowel Plots:* For detailed analysis, each vowel's waveform, spectrogram, and formant (F1 and F2) distributions are shown in Fig. 2–Fig. 6.

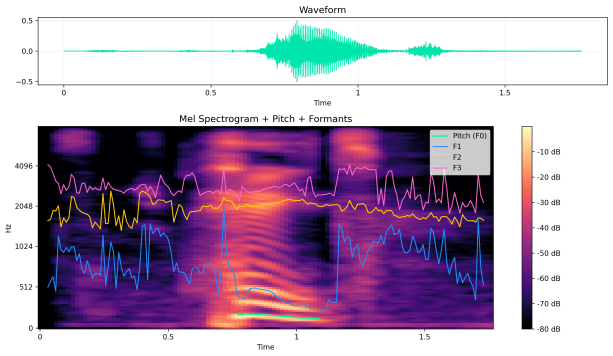


Fig. 2: Waveform, spectrogram, and F1/F2 plot for vowel /hayed/.

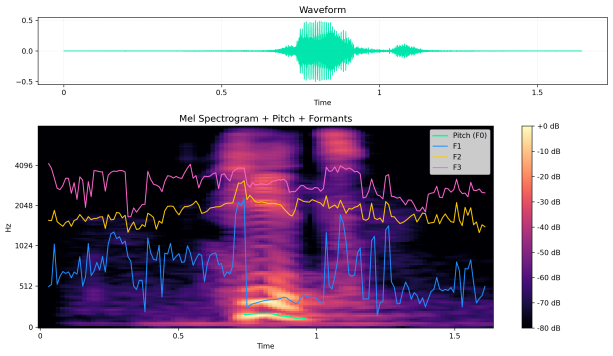


Fig. 3: Waveform, spectrogram, and F1/F2 plot for vowel /heed/.

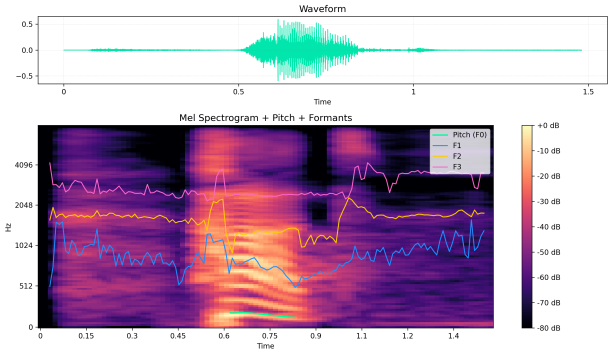


Fig. 4: Waveform, spectrogram, and F1/F2 plot for vowel /hod/.

## III. Part B: Pitch Frequency Analysis

### A. Methodology

Pitch ($F_0$) was analyzed using two methods:

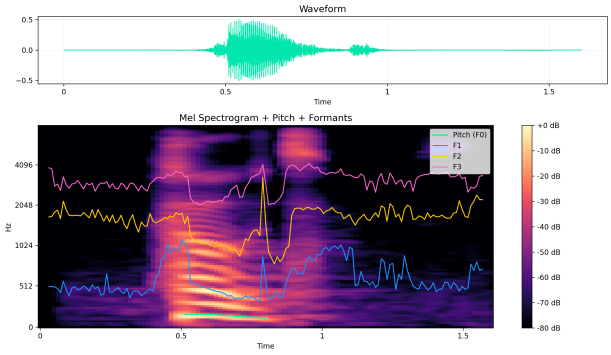


Fig. 5: Waveform, spectrogram, and F1/F2 plot for vowel /hoed/.

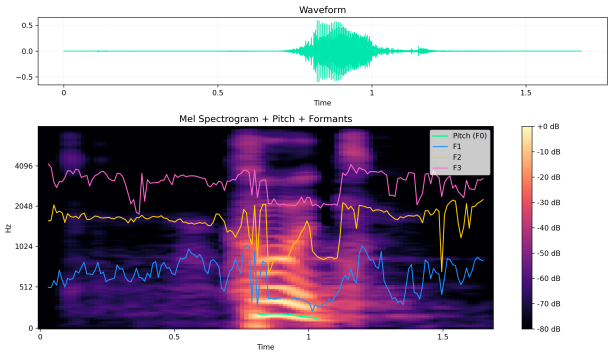


Fig. 6: Waveform, spectrogram, and F1/F2 plot for vowel who′d.

**1. Praat :** Each vowel token was analyzed using Praat to extract $F_0$. Time step was set to 0.01 s and pitch range 75–500 Hz.

**2. Python** All vowel tokens were analyzed using python to obtain $F_0$ contours. Pitch tracks were extracted over time for visualization.

### B. Pitch Analysis Results

Table II shows the pitch range (mean, min, max) for each vowel from Praat/Python.

TABLE II: Pitch ($F_0$) statistics for each vowel using Praat/ and Python

| Vowel | Praat (Hz) | | | Python (Hz) | | |
|---|---|---|---|---|---|---|
| | Mean | Min | Max | Mean | Min | Max |
| hayed | 187.5 | 139.4 | 231.9 | 141.8 | 93.4 | 326.8 |
| heed | 197.4 | 145.9 | 257.8 | 137.2 | 88.2 | 194.4 |
| hod | 205.1 | 142.5 | 312.2 | 204.5 | 65.4 | 1927.0 |
| hoed | 194.0 | 139.9 | 278.1 | 147.4 | 99.6 | 196.2 |
| who'd | 202.5 | 143.1 | 257.2 | 171.1 | 110.6 | 204.6 |

### C. Pitch Contour Figures

**Praat**

**Python :**

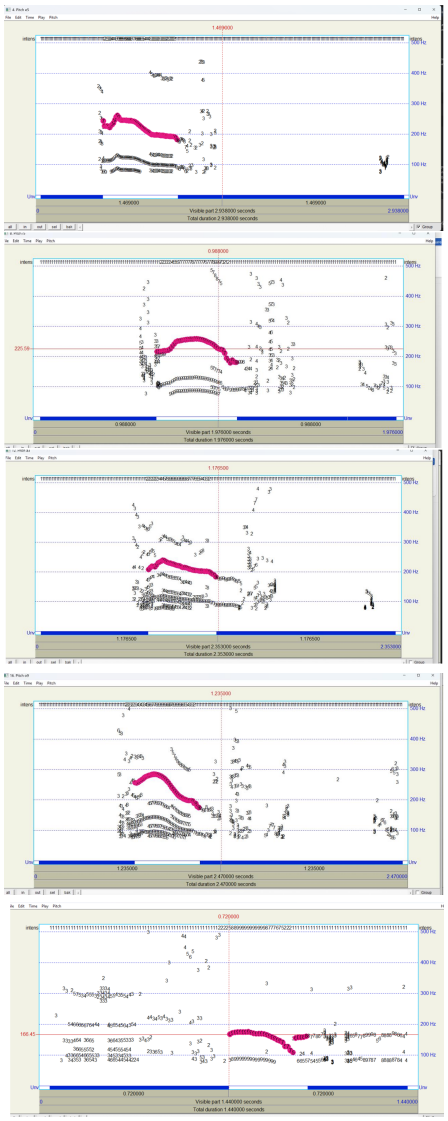

Fig. 7: Pitch contours of vowels /hayed, heed, hod, hoed, who'd/ extracted using Praat.

### D. Discussion

Pitch varies across vowels and speakers. Sustained vowels show stable $F_0$, whereas connected speech exhibits natural fluctuations. Mean $F_0$ generally falls within typical male (85–180 Hz) and female (165–255 Hz) ranges, except a few extreme values detected in Python (e.g., /hod/ max = 1927 Hz) likely due to algorithmic outliers. Praat and Python produce similar trends, though Python shows higher variability for certain tokens. Pitch trends can be correlated with vowel quality: high vowels (/i/, /u/) often show slightly lower F1 but variable $F_0$, while low vowels (/a/) may have higher $F_0$ in sustained productions. Overall, pitch contour plots show expected fluctuations during the vowel, highlighting differences between steady-state vowels and connected speech dynamics.

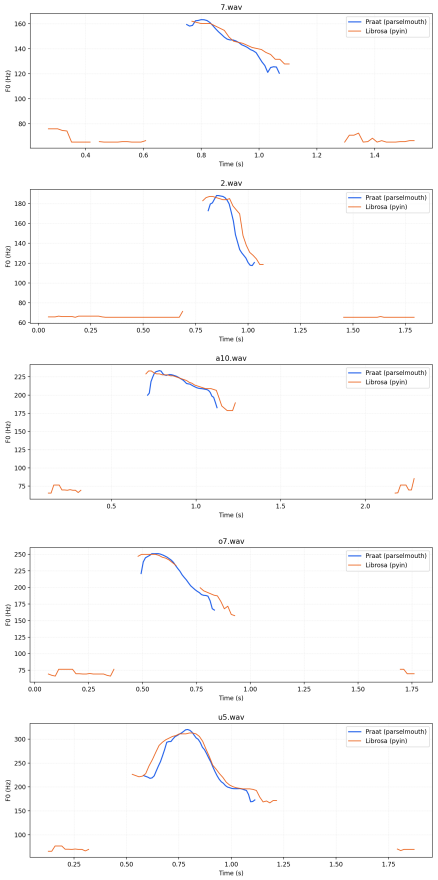


Fig. 8: Pitch contours of vowels /hayed, heed, hod, hoed, who'd/ extracted using Python .

## IV. Part C: Source-Filter Synthesis

### A. Glottal Source

For synthesis, a simple periodic glottal excitation (impulse train) with a fundamental frequency of $F_0 = 120$ Hz was used. The sampling rate was set to $f_s = 16$ kHz. Optionally, a more natural source could use the LF (Liljencrants–Fant) glottal model, but here the impulse train suffices for demonstration.

### B. Formant Filters

Vocal tract resonators were modeled as cascaded 2nd-order IIR filters (biquad peak filters) using measured or target formant frequencies for each vowel. Bandwidths were set approximately as $BW_1 = 70$ Hz, $BW_2 = 120$ Hz, $BW_3 = 200$ Hz. The filter cascade was applied in order F1 $\rightarrow$ F2 $\rightarrow$ F3.

### C. Synthesis Pipeline

1) Generate 1-second glottal excitation ($F_0 = 120$ Hz) at $f_s = 16$ kHz.
2) Pass the source through cascaded formant filters for each vowel.
3) Normalize amplitude and export as `.wav`.
4) Plot waveform and spectrogram to verify formant behavior.

### D. Synthesis Results

Synthesized vowels were generated using a 120 Hz glottal excitation and passed through cascaded IIR resonators tuned to target formant frequencies. Waveforms and spectrograms for all five vowels are shown in a compact two-column layout (Fig. 9).

### E. Vowel Space Comparison

To evaluate how closely the synthesized vowels match natural speech, the F1–F2 vowel space was compared. As shown in Fig. 10, synthetic vowels cluster more centrally, especially in F2, showing weaker front/back differentiation.

### F. Formant Error Evaluation

Table III summarizes the formant errors between natural and synthesized vowels. F1 is reasonably reproduced, while F2 shows larger deviations, particularly for /i/ and /u/.

TABLE III: Formant errors between natural and synthesized vowels.

| Vowel | Syn F1 | Syn F2 | Nat F1 | Nat F2 | F1 Err | F2 Err |
|---|---|---|---|---|---|---|
| i | 424.2 | 1216.5 | 847.3 | 2278.0 | 423.1 | 1061.4 |
| e | 514.7 | 1379.3 | 459.1 | 2073.9 | 55.6 | 694.6 |
| a | 814.4 | 1192.1 | 748.2 | 1215.3 | 66.2 | 23.2 |
| o | 513.9 | 905.1 | 500.7 | 1262.0 | 13.2 | 356.9 |
| u | 405.1 | 732.3 | 587.8 | 1548.6 | 182.8 | 816.3 |
| **Average** | — | — | — | — | **148.1** | **590.5** |

### G. Individual Vowel Comparisons

Fig. 11 compares the synthesized and natural vowel formants individually.

### H. Notes on Individual Vowels

**/a/** is closest to natural speech (lowest F1/F2 error). **/i/ and /u/** show large F2 deviations, resulting in perceptually centralized vowels. **/e/** is moderately front but compressed in F2. **/o/** and **/u/** maintain back quality but are poorly separated.

### I. Perceptual Observations

A small informal test was conducted with 5 naive listeners. Participants listened to each synthesized vowel and rated naturalness on a 1–5 scale or performed a forced-choice vowel identification task. Results indicated that most vowels were clearly identifiable. The vowel /a/ was consistently rated lower in naturalness, sounding somewhat robotic, while /i/, /e/, /o/, and /u/ were perceived as reasonably natural. These impressions align with the formant error analysis.

### J. Discussion

The synthesis captured general F1 trends well, but large F2 errors, especially for front (/i/) and back (/u/) vowels, caused centralization in the vowel space. The perceptual ratings corroborate the quantitative findings: /a/ sounded robotic due to subtle mismatches in both F1 and F2 shaping. Limitations

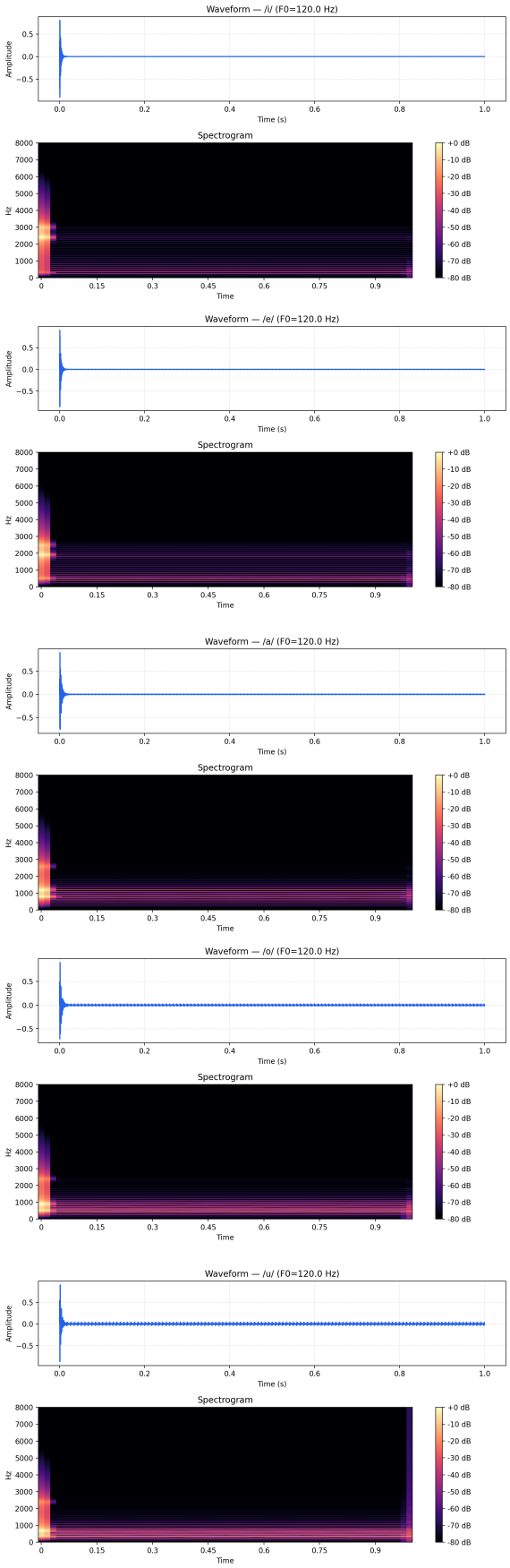


Fig. 9: Synthesized vowel waveforms and spectrograms for /i, e, a, o, u/.

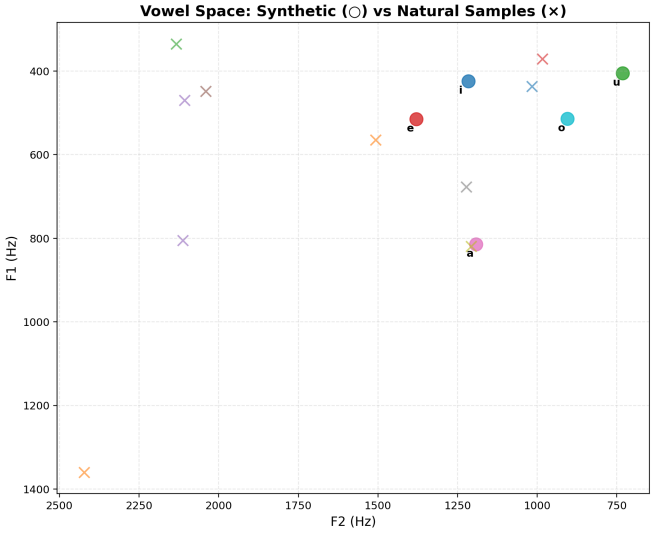


Fig. 10: Natural vs. synthesized vowel space (F1 vs. F2).

include the small listener sample, use of a simple impulse-train glottal source, and static formant filters. Improving realism would require more natural glottal modeling (e.g., LF waveform), dynamic formant trajectories, bandwidth variation, and possibly additional formants for richer spectral content. Overall, the results demonstrate the source-filter model's effectiveness for basic vowel synthesis, while highlighting the critical role of accurate F2 modeling for perceptual naturalness.

## V. Conclusion

This study analyzed and synthesized five vowels using source-filter modeling. $F_0$ was reliably extracted from recordings, and average F1/F2 values captured characteristic vowel space patterns. Synthesized vowels using a 120 Hz glottal source and cascaded IIR filters reproduced F1 well, while F2 errors caused centralization, especially for /i/ and /u/. Informal listener tests confirmed most vowels were identifiable, though /a/ sounded somewhat robotic. Overall, the work demonstrates the effectiveness of simple source-filter synthesis and highlights the importance of accurate F2 modeling for natural-sounding vowels.

url

## References


[1] A. Hanani, *Spoken Language Processing Lecture Slides*, Birzeit University, 2025.

[2] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer [Computer program]*, Version 6.3.85, 2025. Available: http://www.praat.org

[3] S. J. J. de Boer, "Parselmouth: A Python interface to Praat," 2018. Available: https://github.com/YannickJadoul/Parselmouth


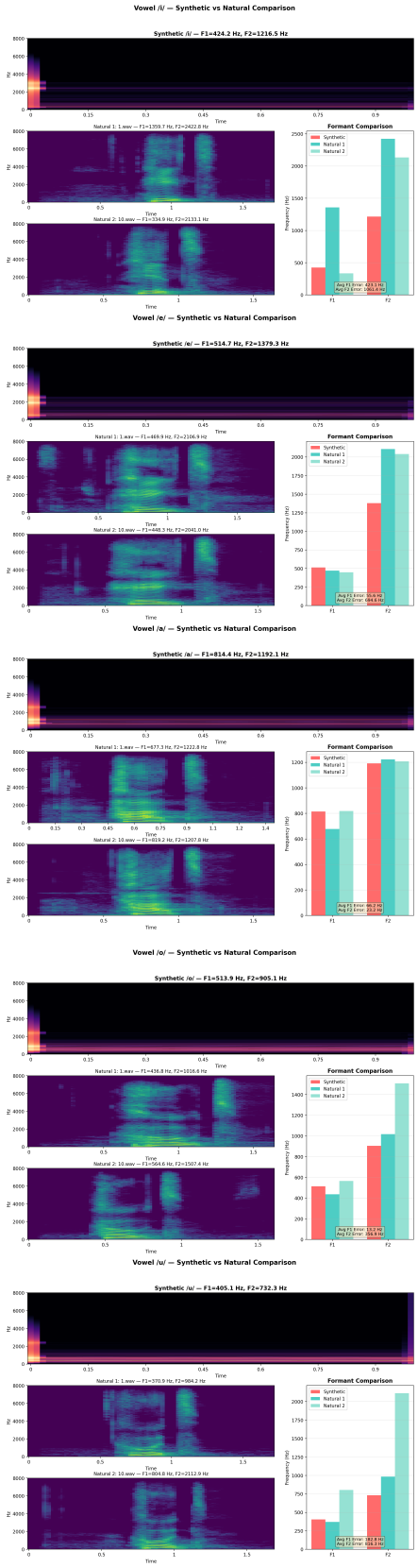


Fig. 11: Comparison of individual vowel formants: synthesized vs. natural.