**EEEM030 Assignment 1**

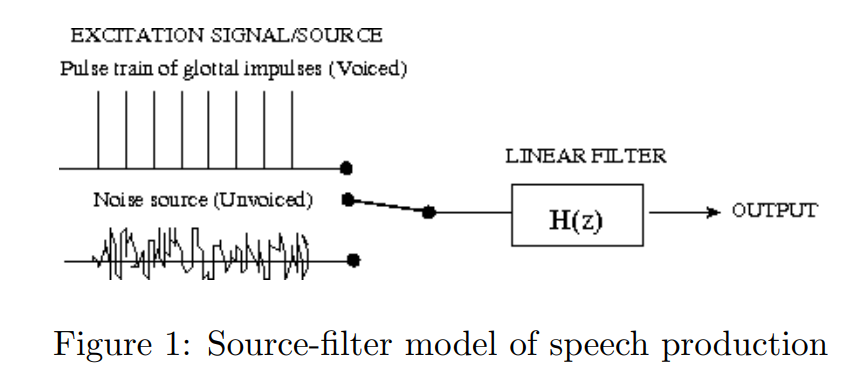
**Linear Predictive Speech Synthesizer**

**TAMAN BACHANI**

**URN: 6846172**

**Abstract**

The goal of this assignment (titled ‘Linear Predictive Speech Synthesizer’), was to leverage the Source-Filter model of Speech Production, to synthesise vowel sounds (monopthongs). For the same, the breakdown into the ‘source’ component, and the ‘filter’ component was studied, understood, and executed.



The aim was to produce vowel sounds that approximate natural human speech. For the ‘source’, as seen in the illustration (Figure 1), a glottal pulse was simulated. And for the ‘filter’, LPC was used to model the characteristics of the vocal tract. The sound is essentially shaped by the LPC coefficients obtained. Various LPC orders were tested (from 6 to 20) to balance formant accuracy and avoid overfitting; ultimately, an order of 20 yielded the best results. The LPC-derived filter was then used to process the impulse train, producing synthesised vowels. Additionally, a low-pass filter was applied to reduce high-frequency noise, enhancing the naturalness of the synthesized vowel. The synthesized vowels were assessed based on several key criteria: the clarity and accuracy of the estimated formant frequencies, the effect of varying LPC order on sound quality, and the perceived naturalness on listening to it after filtering. Though challenging, Python was used to execute and complete the assignment. A deep dive into libraries: librosa, and scipy was performed to achieve the code that shall be explored in the Methodology section.

**Introduction**

As the task at hand was understood, a goal based approach was adopted to achieve the desired results. The following were identified as clear goals for the assignment:

1. To estimate LPC coefficients accurately for short vowel segments.

2. To plot the LPC filter's frequency response and the amplitude spectrum of the original speech samples.

3. To identify the first three formant frequencies of the vowel sound.

4. To estimate the mean fundamental frequency (F0) of the vowel segment.

5. To generate a periodic impulse train as the excitation signal and use it for LPC-based synthesis.

6. To evaluate and compare different LPC model orders and segment lengths to find the best fit for synthesis.

7. To conduct an informal subjective assessment of the synthesised vowel sounds and discuss the quality.

The specific audio files chosen for this assignment were: “*head\_f.wav*”, and “*hood\_m.wav*”. The female vowel sound being “*ea*”, and for the male being “*oo*”. Since LPC is known to clearly represent the stable formant structures, the isolated vowel sounds were analysed and synthesised using the same. A sub-segment of 100ms of each audio file was chosen, so a consistent portion of the vowel sound is considered. Consistency was learned to be a crucial requirement for an accurate LPC analysis.

To the trimmed vowel sounds, the LPC method was applied, which calculated the LPC coefficients, to capture the optimal formant structure. LPC orders of [6, 10, 14, 16, 20] were explored, this allowed the examination of trade-offs between spectral detail and overfitting. High orders capture more frequency (detail) but add noise too. Though the noise can be treated, or tinkered with later. Which was done by adding jitter to the simulated impulse train, and passing the final synthesised waveform through a low-pass filter. To simulate the impulse train, the fundamental frequency (F0) of each segment was calculated and used to set the train’s periodicity, aligning the synthesised sound’s pitch closely with that of the original vowel recordings..

The synthesised vowel sounds, or these frequencies often create an artificial or "robotic" tone. Which was seen across the LPC orders and irrespective of the segment length. Hence, the synthesised vowel sounds were assessed based on the clarity of the formant structure, the alignment of the frequency response with the original sample, and an informal assessment of naturalness, which will be discussed in the Conclusion section of this report. In the following section, specific methods, parameters, and code implementations will be covered. And observations on the synthesis quality and insights on improving the naturalness of LPC-based vowel synthesis will be shared.

**Methodology**

This section discusses the methods adopted to achieve the goal of synthesising vowel sounds using a Python-based approach in place of MATLAB. By utilising the Source-Filter model and various Python libraries (librosa, scipy, matplotlib, and numpy), each step in the synthesis process was carefully tailored to model the characteristics of human vowel production and achieve a realistic approximation of natural sounds.

**Python-based Approach**

Python was chosen for this assignment due to its extensive libraries for signal processing and audio analysis, which serve as effective replacements for MATLAB’s functions. Key libraries include `*librosa*` for audio loading, analysis, and LPC (Linear Predictive Coding) coefficient estimation, `*scipy*` for filtering and Fourier transformations, *‘matplotlib’* for plotting crucial graphs, and `*numpy*` for generating synthetic signals and performing mathematical operations. These libraries collectively enabled the implementation of the Source-Filter model with LPC for vowel synthesis, replicating MATLAB’s functionality. The installation commands, code, and dependencies are detailed in the appendix (Code Block 1).

**Processing the Source Files**

The source files selected for synthesis were “*hood\_m.wav*” and “*head\_f.wav*”, as mentioned above. These files were uploaded into the Colab runtime directly and subsequently loaded into the environment using `librosa`. Each file was trimmed to a 100 ms segment, isolating stable vowel sounds, with the male and female segments stored as `wave\_male\_segment` and `wave\_female\_segment` (Code Block 2). Fundamental frequency (*F0*) estimation was then performed over these segments to establish the base frequency for the periodic impulse train that shall be simulated later. The mean F0 values calculated were ***90.38 Hz*** for the male segment and ***220.11 Hz*** for the female segment. These values, as observed in the *F0* graph, are typical for male and female voices, with the male segment reflecting a lower pitch consistent with typical male vocal frequencies and the female segment a higher pitch. The *F0* estimation plot (Figure 2, Figure 3), verified the stability of *F0* over time (Code Block 3), which was essential to ensure consistent pitch in the synthesis, and was set up as a debugging mechanism when diagnosing and curing the problem of artificial or robotic sound.

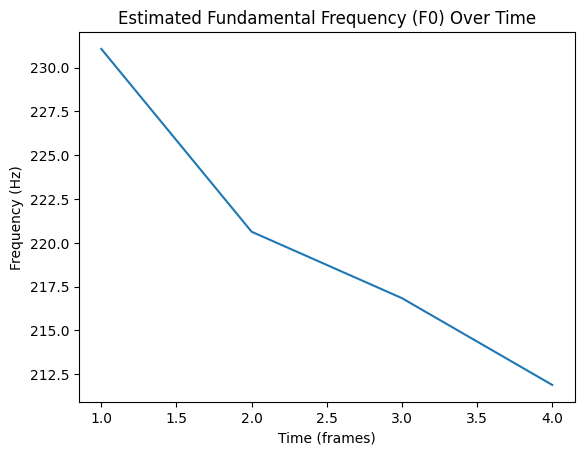
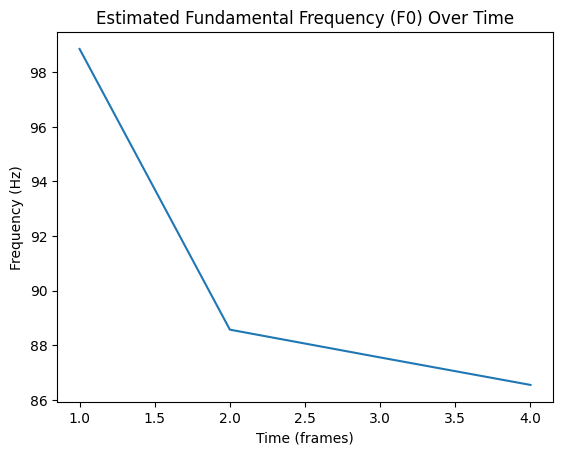


Figure 2: For Male Figure 3: For Female

**The LPC Analysis**

The LPC analysis required experimentation with various LPC orders, as this parameter directly affects the clarity and quality of the synthesised vowel sound. Orders of ***6, 10, 14, 16, and 20*** were tested. Higher orders were observed to improve the distinctiveness of the vowel sound, transitioning from an initial robotic buzz to a more recognisable vowel quality. The final choice, an LPC order of **20**, provided a well-defined formant structure, while still maintaining stability in the output, this was achieved by some neat tricks and some minor post-processing. The `***librosa.lpc***` function was used to generate LPC coefficients for each segment (Code Block 4), which modelled the vocal tract’s resonance characteristics. The resulting coefficients (Figure 4) captured the formant frequencies necessary to shape the impulse train into a recognisable vowel sound (‘ea’ and ‘oo’ in our case).

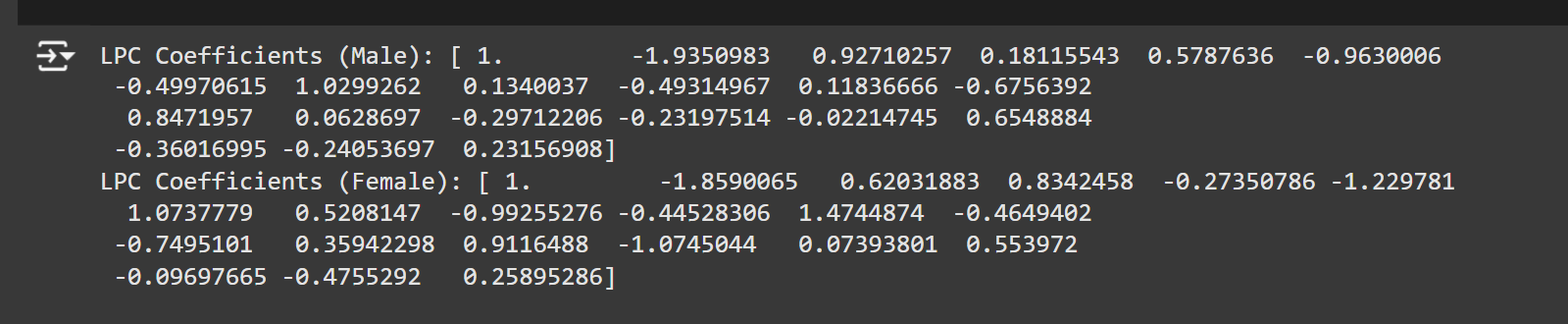


Figure 4: LPC Coefficients for Male and Female

**Estimation of the Amplitude Spectrum, the Frequency Response, and the Formant Frequencies**

The amplitude spectrum and frequency response of the LPC filter were critical in analysing the formant structure and confirming the accuracy of the synthesised sound. To obtain the amplitude spectrum, each vowel segment was transformed using the Fourier transform, with `*scipy*` performing the spectral analysis. The results were plotted using `*matplotlib*` to visualise the distribution of frequencies (Figures 5, 6, 7, & 8). The frequency response of the LPC filter, plotted using `***scipy.signal.freqz***`, shows how the LPC filter emphasises specific frequencies corresponding to formants, aligning closely with the natural vowel structure. These plots confirmed that the synthesised signal captured essential formant peaks, making the sound more realistic. (Code Block 5)

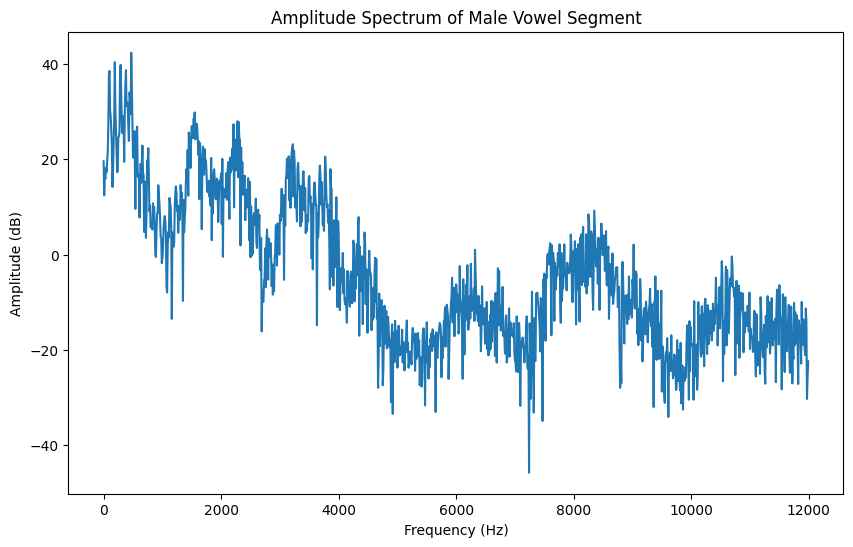


Figure 5: Amplitude Spectrum of Male Vowel Segment

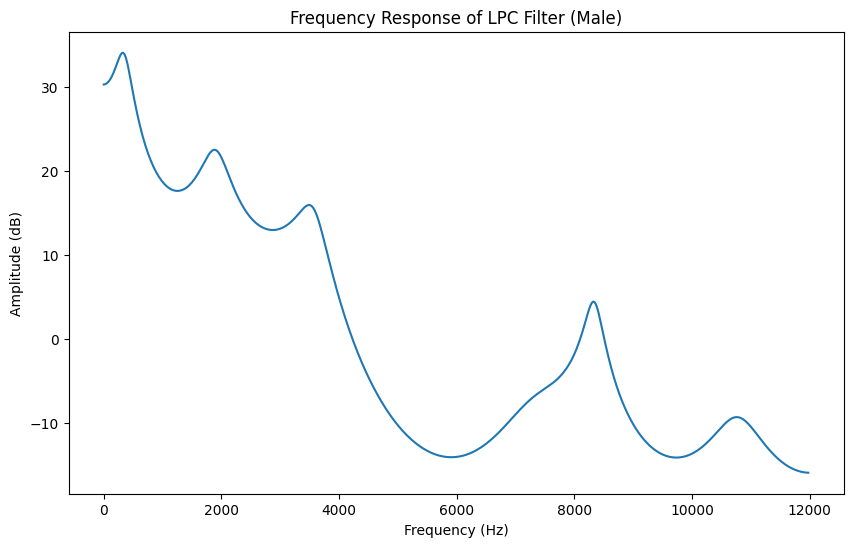


Figure 6: Frequency Response of Male Vowel Segment

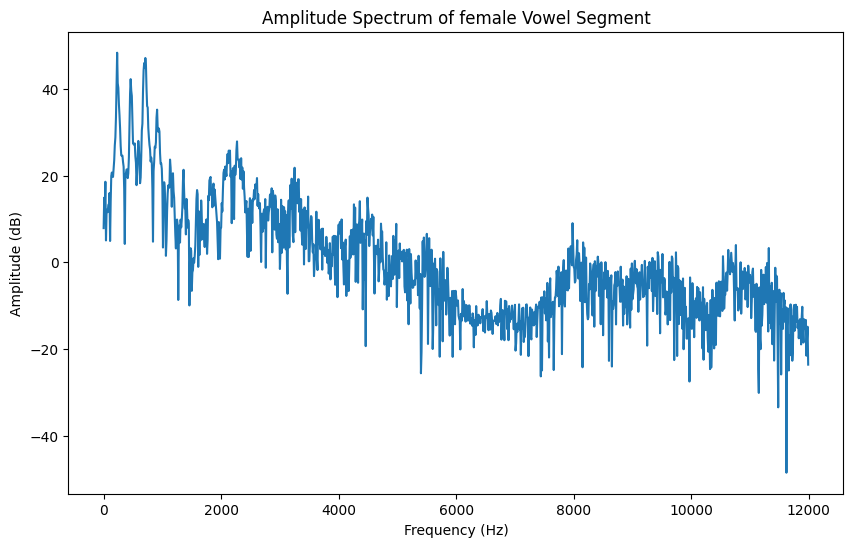


Figure 7: Amplitude Spectrum of Female Vowel Segment

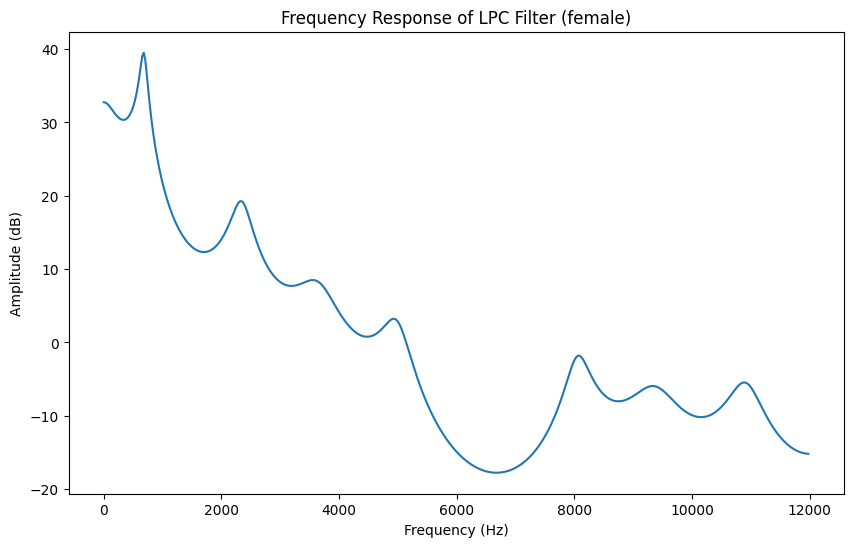


Figure 8: Frequency Response of Female Vowel Segment

Formant frequencies, the resonant frequencies that define vowel quality, were obtained by identifying prominent peaks in the frequency response. This was achieved using the ***‘find\_peaks’*** function in scipy, applied to the magnitude of the frequency response converted to decibels using ‘***20\*np.log10(abs(h\_male))****’* and ‘***20\*np.log10(abs(h\_female))****’* (Code Block 6). The *np.log10* function computes the *logarithm base 10* of the absolute magnitude values, which is then multiplied by *20* to express the amplitude in decibels (dB). This conversion to a dB scale is standard in signal processing to enhance the contrast of peaks, making formant identification clearer. Using ‘***find\_peaks***’, the first three prominent peaks were identified, representing the primary formants for each vowel.

The resulting formant frequencies were:

***Male Vowel Formant Frequencies: [328.125 Hz, 1875 Hz, 3492.1875 Hz]***

***Female Vowel Formant Frequencies: [679.6875 Hz, 2320.3125 Hz, 3562.5 Hz]***

These values reflect the expected formant distribution for typical male and female vowel sounds, with higher frequencies observed in the female segment. Formant estimation provides critical information on how accurately the LPC model replicates the natural vowel characteristics, demonstrating that the synthesised signal aligns well with human vowel structure.

**Simulation of the Impulse Train**

A periodic impulse train was generated to simulate the glottal pulse in voiced speech. Using the estimated *F0* values, a one-second impulse train was created. Slight jitter was introduced to the impulse positions using `numpy` to avoid an overly mechanical sound and mimic natural variability in human speech production, this was added after multiple failed (robotic sound) attempts and improved the performance and the natural quality of the sound. The impulse train generation code and parameters are provided in the appendix (Code Blocks 7 & 8), including the incorporation of jitter.

**The Final Synthesis**

The final synthesis was achieved by filtering the impulse train with the LPC-derived filter using `***scipy.signal.lfilter***`. Initial attempts yielded a robotic buzz rather than a vowel, prompting further tuning. Adjustments such as increasing the LPC order, adding jitter, and normalising the synthesised signal significantly improved the sound quality. The output was assessed subjectively, revealing a recognisable vowel sound with reduced artefacts and improved naturalness. The informal quality assessment and results are discussed further in the conclusion, along with potential improvements, and a summary of rejected attempts. (Code Blocks 9, 10, & 11)

**Conclusion**

This assignment successfully synthesised two distinct vowel sounds using the Source-Filter model and LPC-based methods in Python. A structured approach was implemented and various parameters were experimented with, the synthesised sounds closely approximated natural vowel sounds. Listening tests confirmed that both the male and female synthesised vowels matched the original recordings reasonably well, capturing the intended vowel quality and essential acoustic characteristics.

**Informal Assessment**

The assignment’s success was assessed using informal criteria, focusing on sound quality and naturalness, and relying on visualisation to debug and understand the differences between the original waveform and the synthesised one. Early attempts at synthesis produced a robotic buzz, despite various approaches, including using Yule-Walker and covariance methods, which initially did not yield satisfactory results, this is covered in the next section. Incremental improvements were achieved by introducing a *slight jitter* to the impulse train, *visualising and stabilising mean F0*, *normalising* the synthesised sound, and *progressively increasing the LPC order* to enhance formant clarity. Informal listening tests indicated that the synthesised sounds were similar to their respective original samples. Switching between the male and female synthesised sounds, while listening, revealed clear distinctions in pitch and formant structure, highlighting the vowel characteristics perceptible to the ear.

**Debugging and Rejected Attempts**

Several alternative methods, including the Yule-Walker and covariance approaches, were explored for estimating LPC coefficients. The Yule-Walker method, based on autocorrelation, aims to fit an autoregressive (AR) model to the signal by minimising the prediction error in a least-squares sense. It was chosen because of its common usage in spectral estimation. However, this method produced a coefficient set that yielded inconsistent formant peaks, making the synthesised sound unnatural and overly metallic.

The covariance method, on the other hand, uses a covariance matrix of the signal, which can handle non-stationary signals better than the Yule-Walker approach. Despite this, it still failed to accurately capture the formant structure for the vowel sounds in this assignment. Both methods ultimately deviated significantly in their output from the `***librosa.lpc***` approach. To diagnose these issues, visual comparisons of the amplitude spectra were performed (Figures 9, 10, & 11). The plot generated using librosa showed a closer resemblance to the original, whereas Yule-Walker and covariance spectra displayed irregularities and artefacts. Ultimately, librosa.lpc, with appropriate parameter adjustments, was found to provide a more realistic vowel synthesis, underscoring the effectiveness of the selected approach.

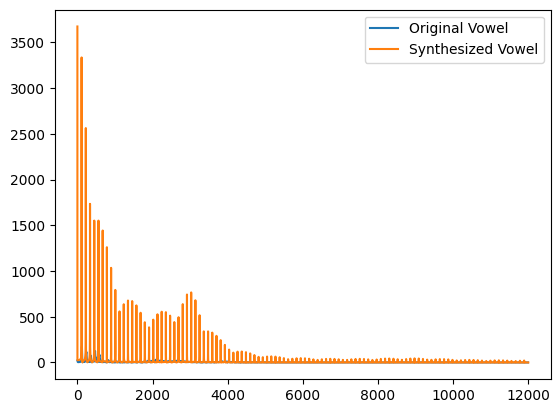


Figure 9: Waveform Comparison (Librosa)

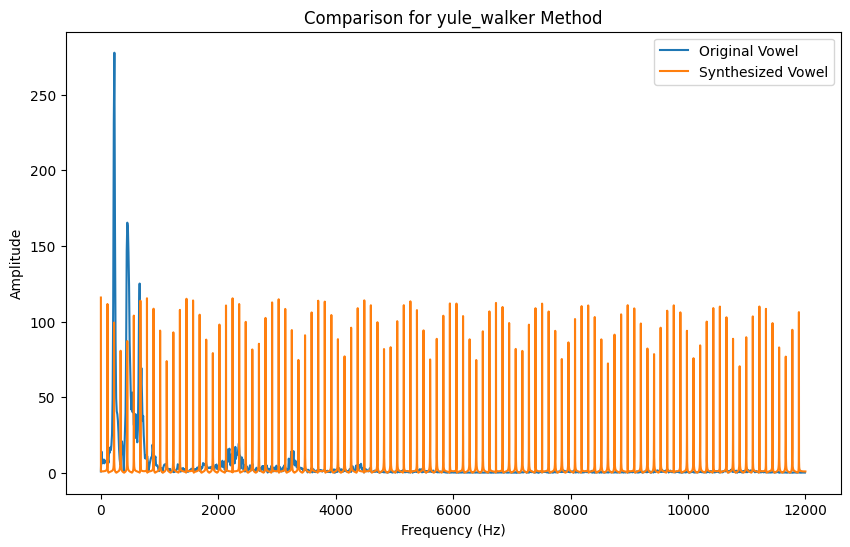


Figure 10: Waveform Comparison (Yule-Walker)

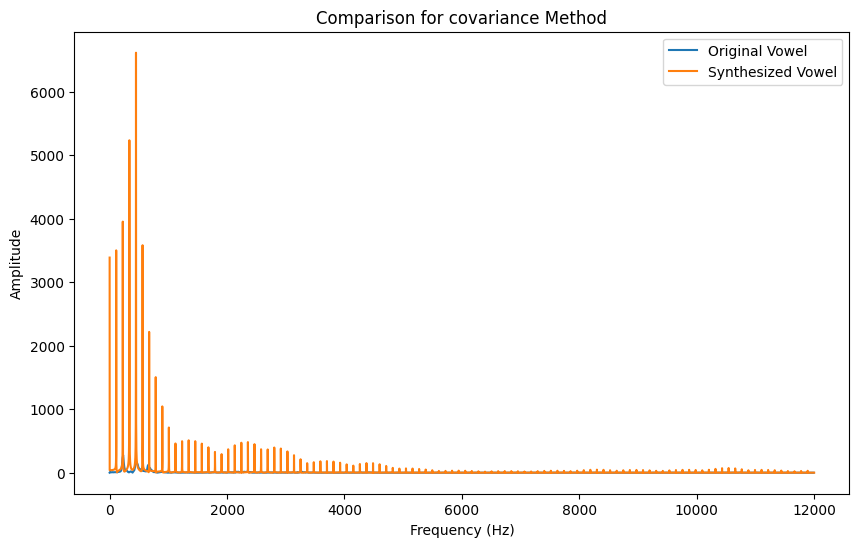
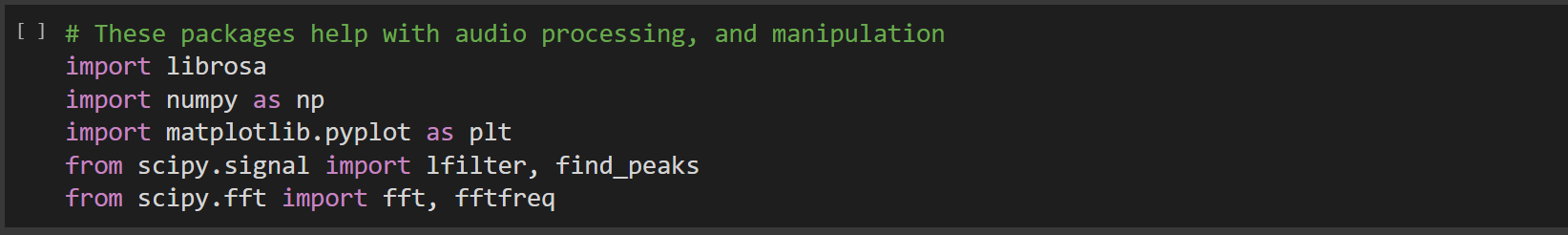


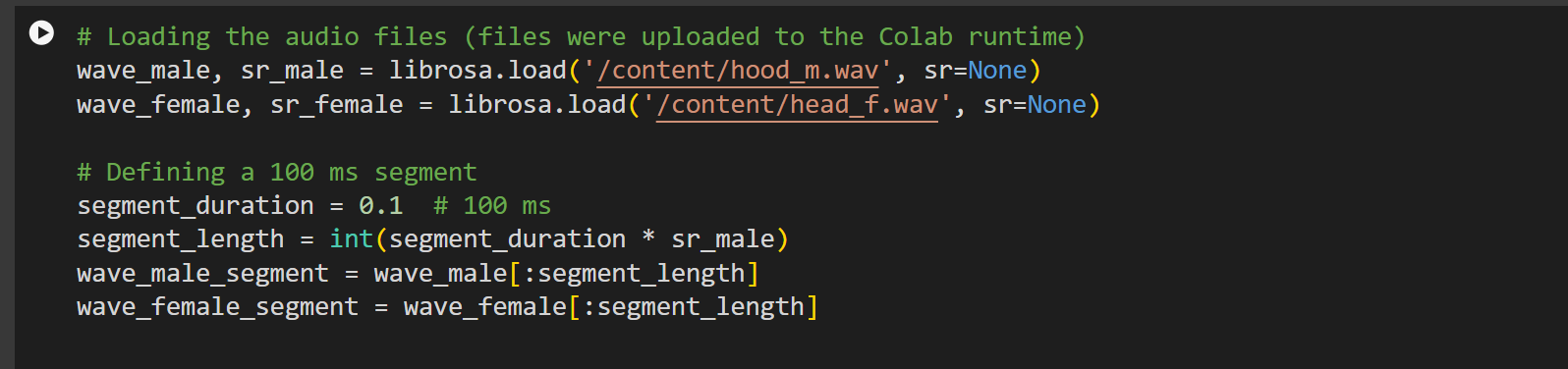
Figure 11: Waveform Comparison (Covariance)

**Appendix**

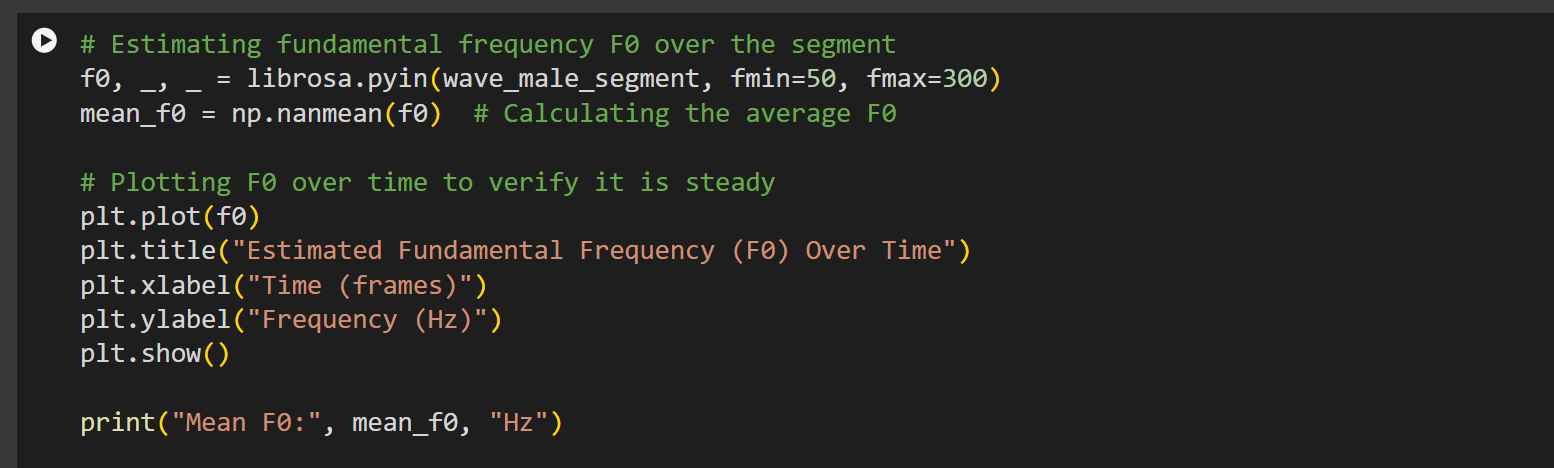
1. **Code Implementation**

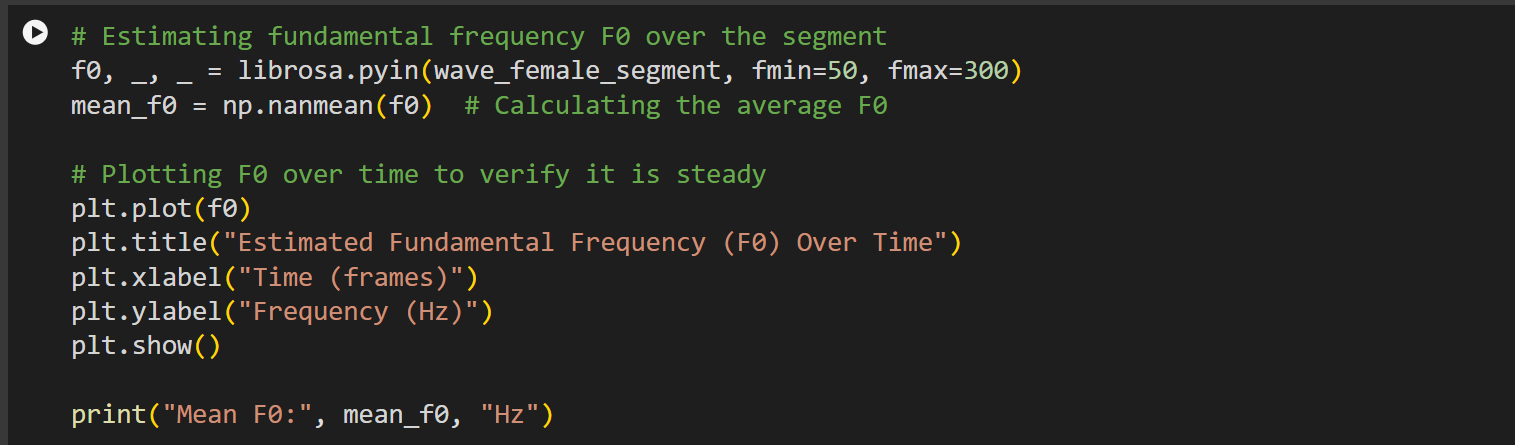
****

Code Block 1: Importing all the libraries

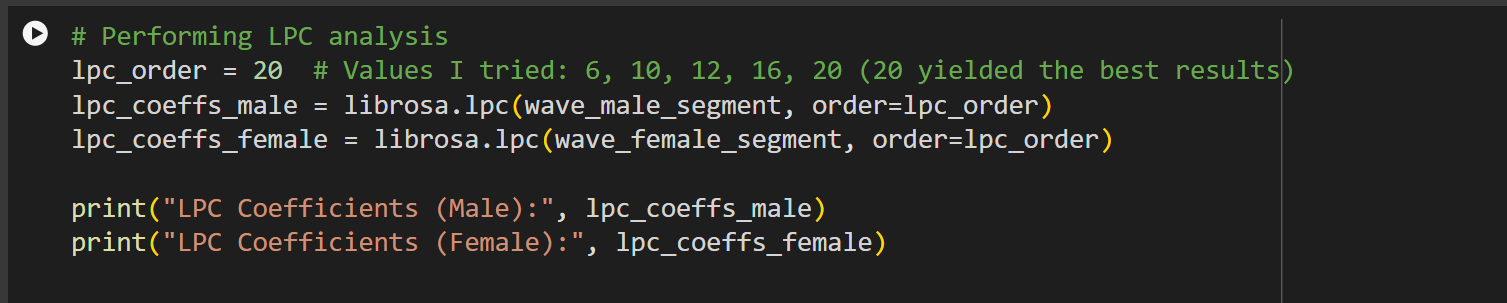


Code Block 2: Loading the Audio Files & Defining a 100ms Segment

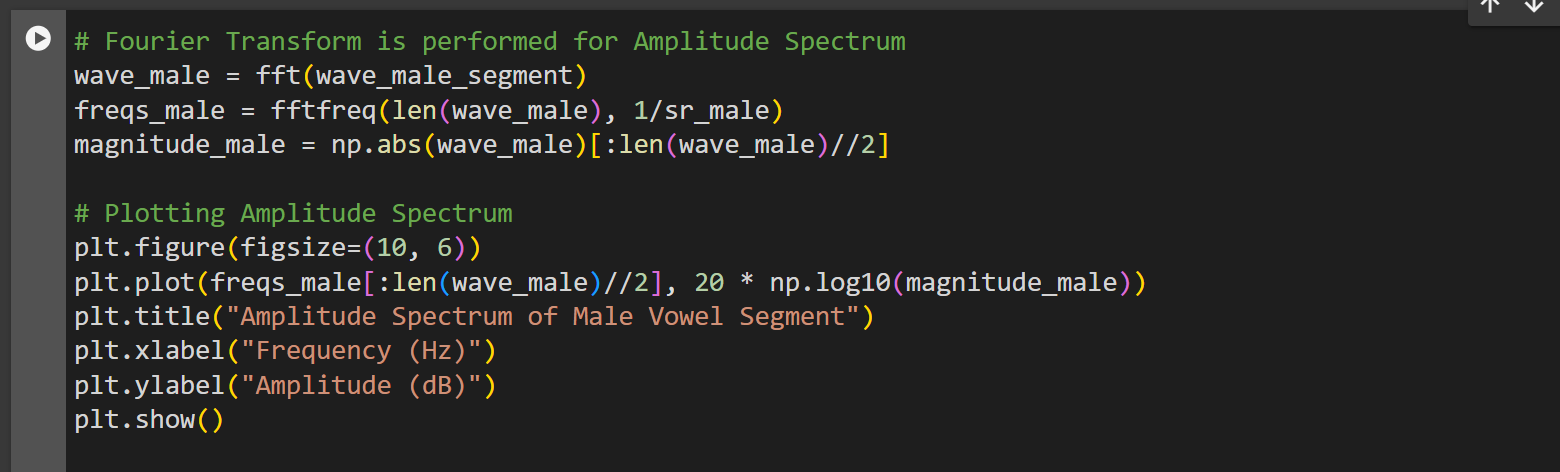


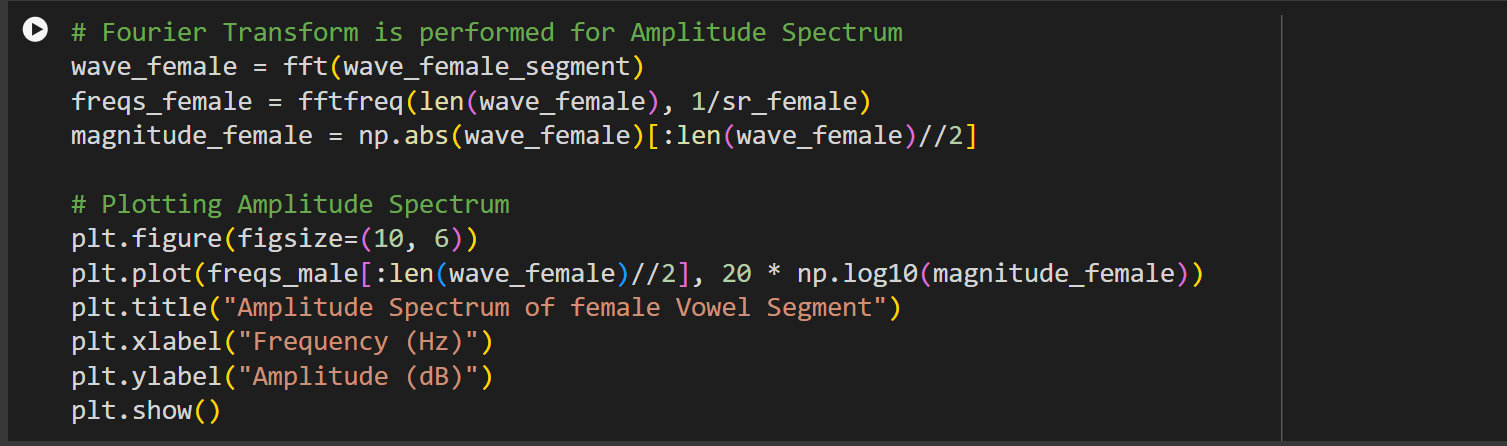


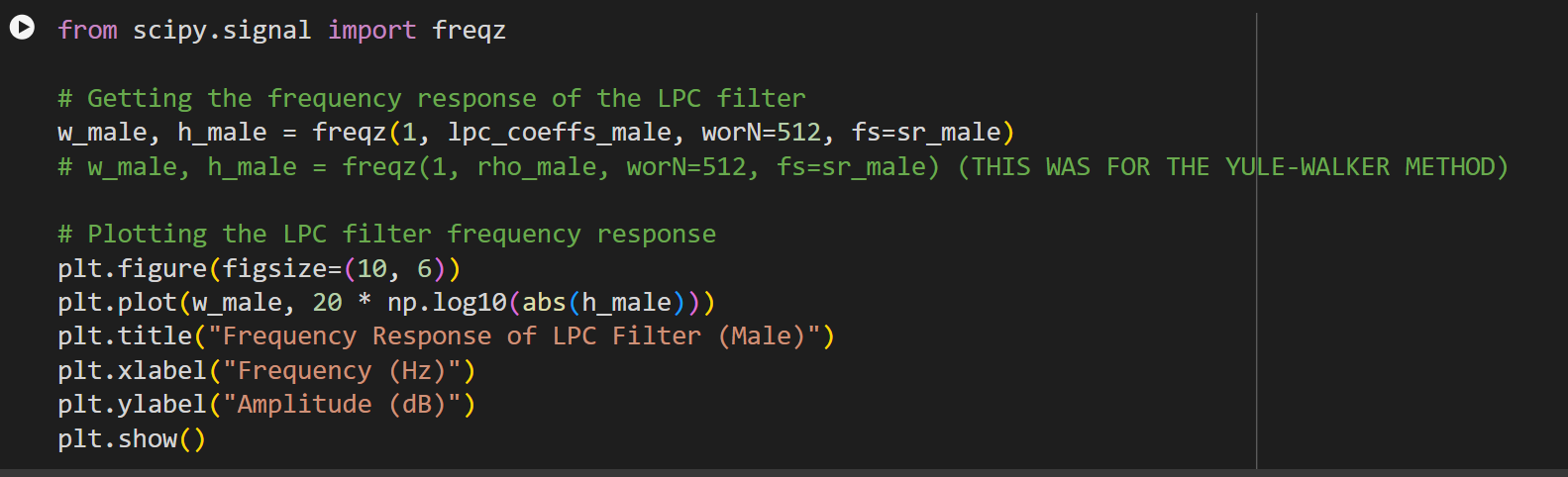
Code Block 3: Fundamental Frequency over time

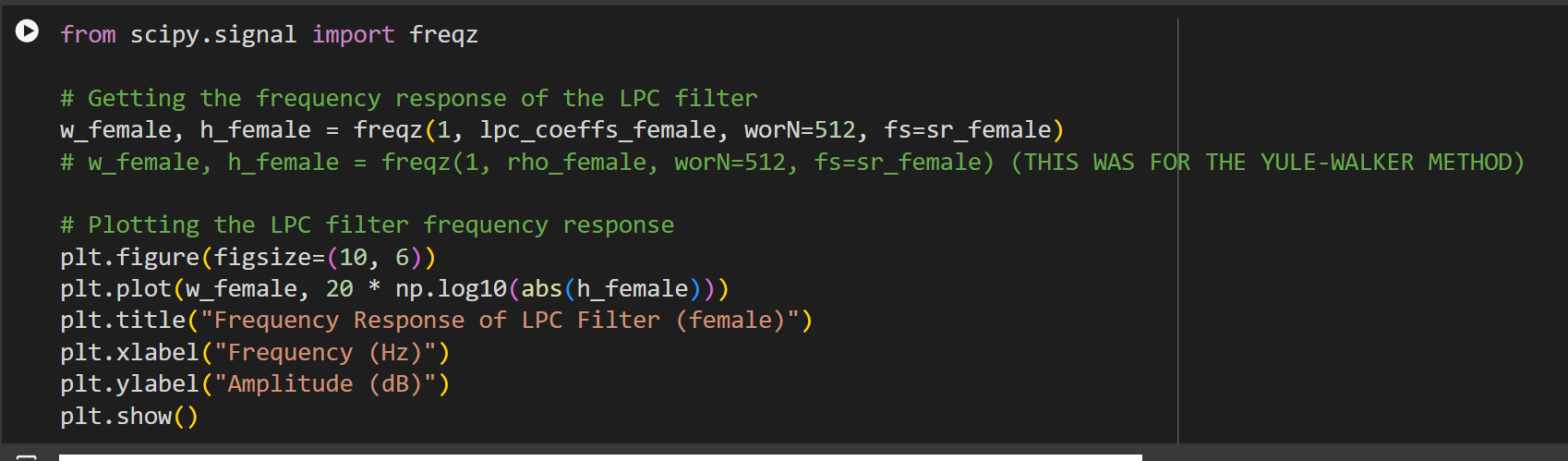


Code Block 4: Performing LPC Analysis

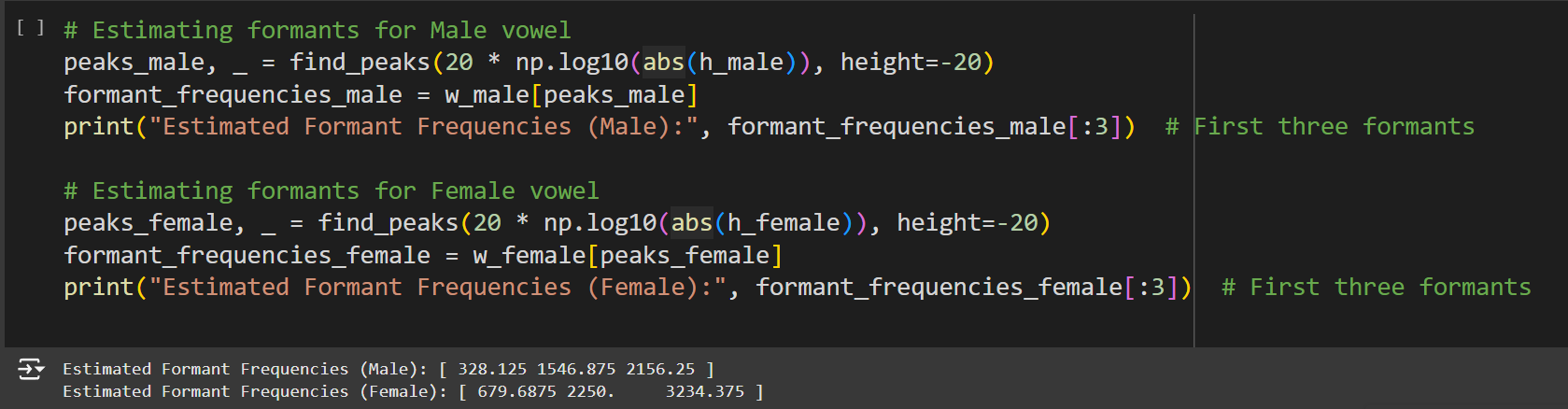




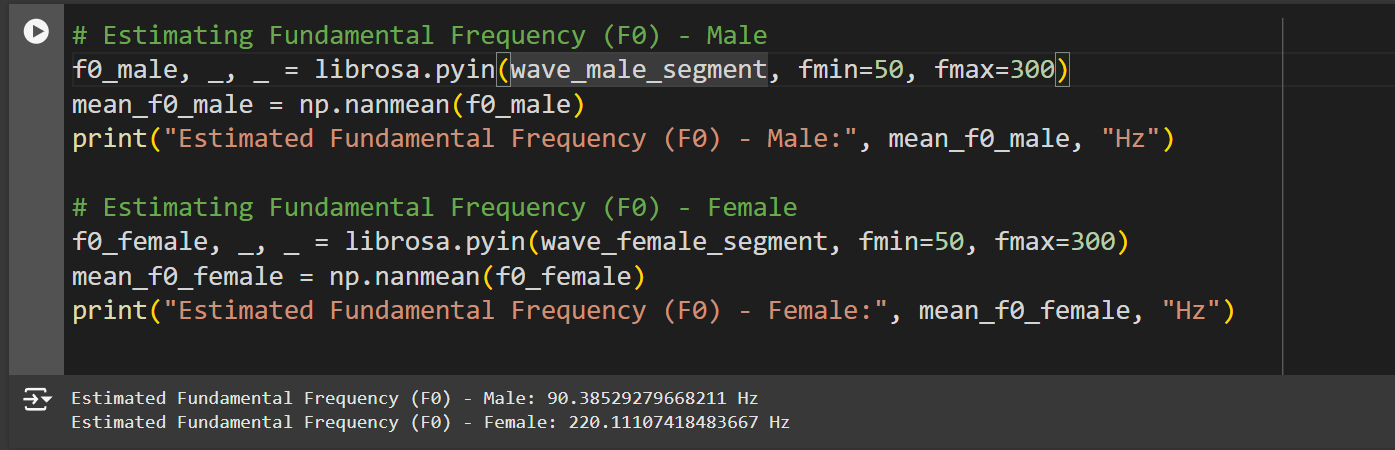




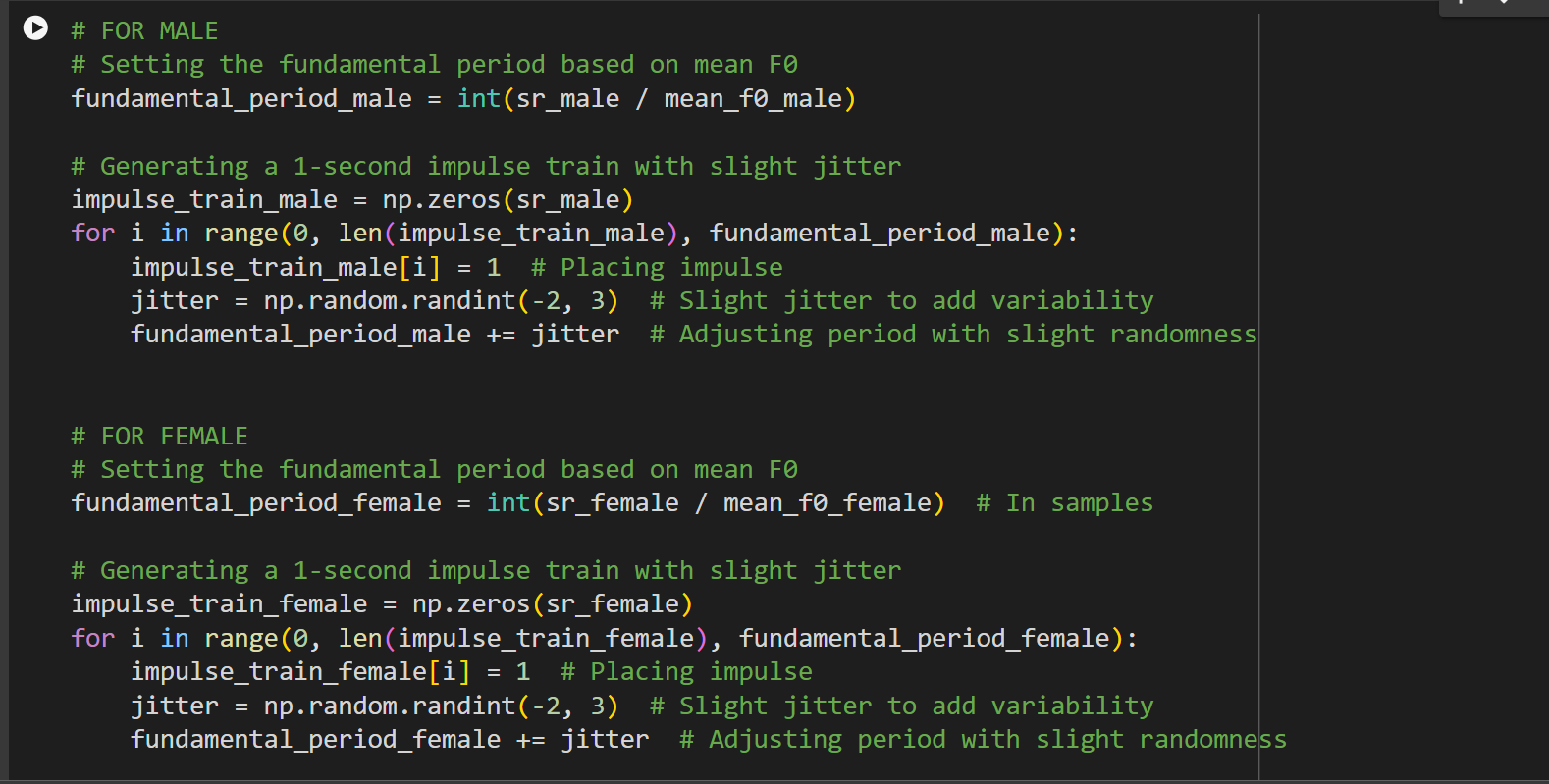
Code Block 5: Calculating and Visualising Amplitude Spectrum and Frequency Response



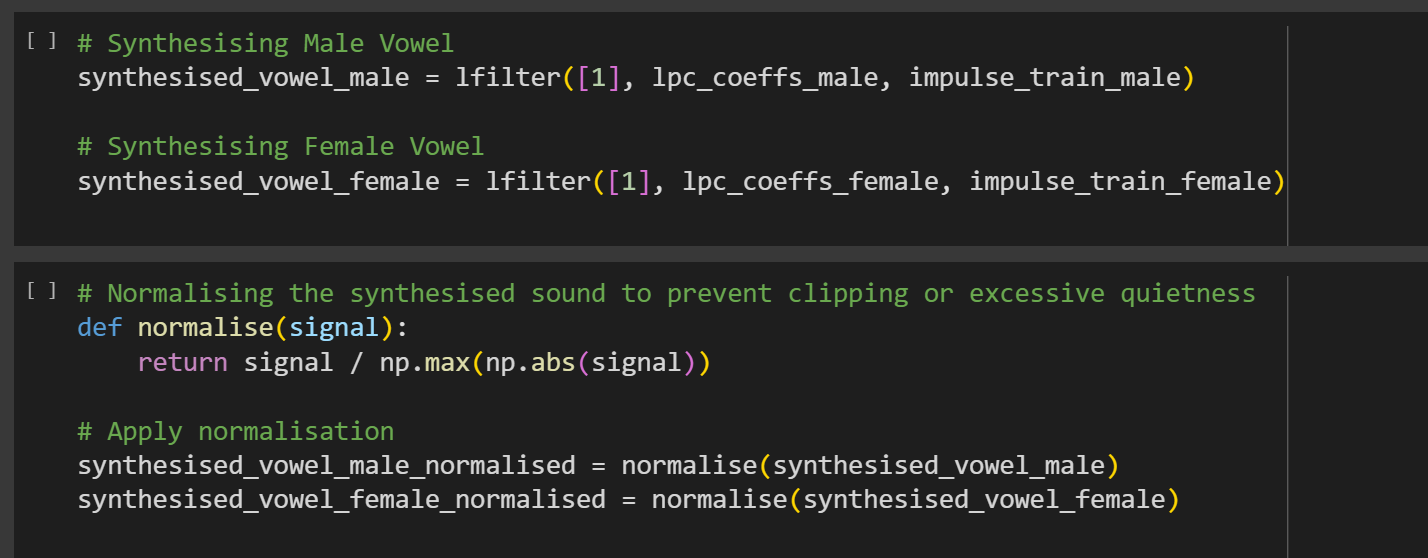
Code Block 6: Estimating Formants



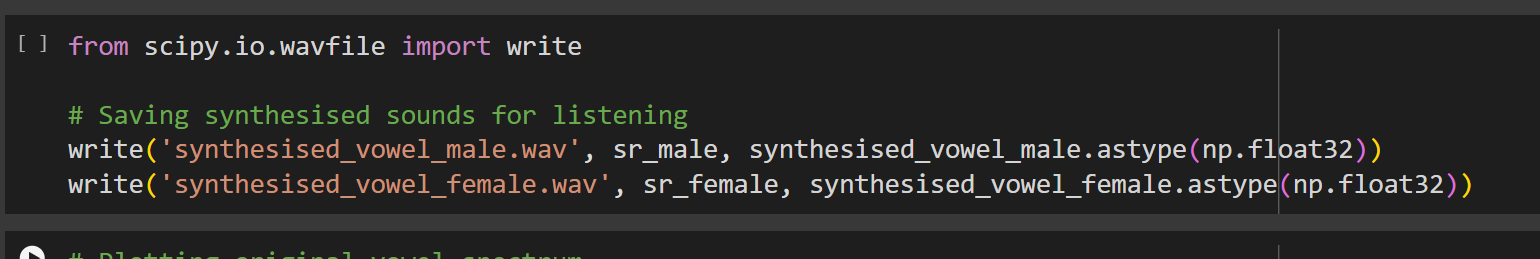
Code Block 7: Estimating Fundamental Frequency (F0)



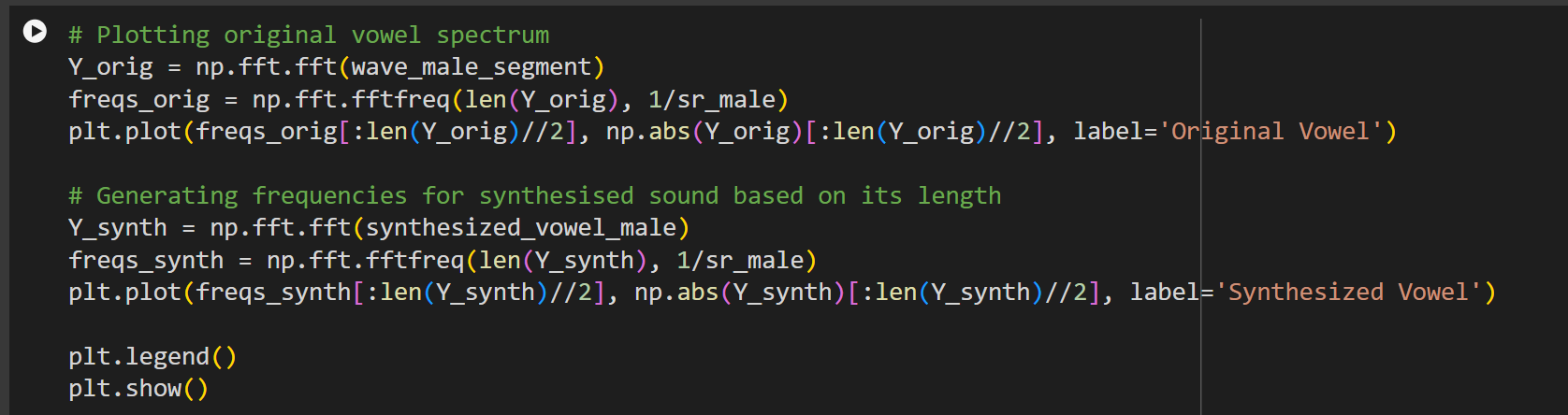
Code Block 8: Generating Impulse Train



Code Block 9: Synthesising and Normalising Vowel Sounds



Code Block 10: Saving the Synthesised Vowel Sounds



Code Block 11: Creation of Waveform Comparison graph for debugging

**B. List of Figures**

1. Figure 1: Diagram of Source-Filter Model – Demonstrates the Source-Filter model of speech production, showing glottal pulse as the excitation source and the vocal tract as the filter.
2. Figure 2: Estimated Fundamental Frequency (F0) Over Time for Male Segment – The mean F0 estimation graph is displayed, showing F0 stability for the male vowel segment.
3. Figure 3: Estimated Fundamental Frequency (F0) Over Time for Female Segment – Displays the mean F0 estimation graph, showing F0 stability for the female vowel segment.
4. Figure 4: LPC Coefficients for Male and Female Vowel Segments – Shows the LPC coefficients obtained for both male and female vowel segments, used to model the vocal tract.
5. Figure 5: Amplitude Spectrum of Male Vowel Segment – A Fourier-transformed representation of frequency distribution for the male vowel segment.
6. Figure 6: Frequency Response of LPC Filter (Male) – Visualises how the LPC filter for the male segment enhances certain frequencies corresponding to formants.
7. Figure 7: Amplitude Spectrum of Female Vowel Segment – A Fourier-transformed representation of frequency distribution for the female vowel segment.
8. Figure 8: Frequency Response of LPC Filter (Female) – Visualises how the LPC filter for the female segment enhances specific frequencies associated with formants.
9. Figure 9: Synthesized vs. Original Amplitude Spectrum Comparison (Librosa Method) – Compares the amplitude spectrum of the original vowel with the spectrum obtained from the synthesised sound using the `librosa` method.
10. Figure 10: Synthesized vs. Original Amplitude Spectrum Comparison (Yule-Walker Method) – Compares the amplitude spectrum of the original vowel with the synthesised version using the Yule-Walker method, highlighting artefacts in the synthesised sound.
11. Figure 11: Synthesized vs. Original Amplitude Spectrum Comparison (Covariance Method) – Compares the amplitude spectrum of the original vowel with the spectrum from the synthesised sound using the covariance method, demonstrating deviations and irregularities.