Linear Predictive Speech Synthesizer

Coursework - 1

Submitted in partial fulfillment of requirement
for the module of

Speech & Audio Processing & Recognition

Submitted by

MANISH PANDA(6614056)

Submitted to

PROF. WENVU WANG

University of Surrey

School of Computer Science and Electronic Engineering

Department of Electrical and Electronic Engineering

United Kingdom

November 2023

Abstract

The source-filter model of speech generation is used in this coursework to investigate the synthesis of male and female vowels.Using linear predictive coding (LPC), the formant structure of each vowel is directly calculated from actual male and female vowel speech samples. By running a periodic impulse train through an all-pole filter that was acquired from the LPC analysis, synthetic vowels are produced.It is advised to utilize MATLAB to finish this coursework because of its built-in functions for filters, audio input/output, AR models, and evaluation of the synthesized speech. This task, which aims to provide an informative study on the synthesis of vowels in the context of speech production, demands a basic understanding of the source-filter model, LPC analysis, and audio signal processing.

Contents

Chapter 1

Introduction

1.1 Speech

Said another way, speech is the medium of communication. Another method to describe speech is in terms of the acoustic waveform, which is the signal that carries the message data. It's a method of using vocal sounds to convey ideas. It is a blend of consonant and vowel sounds. Voiced and unvoiced speech are the two categories into which speech signals are separated.

1.1.1 Voiced and Unvoiced speech

Phonemes, which are produced by the vocal cords and vocal tract (which includes the mouth and lips), are the building blocks of speech. When a phoneme is spoken, vibrations in the vocal chords result in spoken signals. In contrast, unvoiced signals do not require the vocal cords to be used. Voiced signals, such as the vowels /a/, /e/, /i/, /u/, and /o/, are typically louder. Conversely, unvoiced signals, such as the stop consonants /p/, /t/, and /k/, are typically more abrupt.

1.2 Speech production

The process of producing speech involves several intricate steps. The brain chooses the words first, and then the articulation process takes place. Vowels are generated with a relatively open vocal tract and no considerable constriction, while consonants are formed by restricting airflow at different locations in the vocal tract. Air is released from the lungs into the vocal tract via the trachea during speech. When breath leaves the lungs, vocal chords vibrate, creating a buzzing sound. Pitch and voice quality are influenced by the vocal cords' rate of vibration and tension. The sounds are then shaped by the vocal tract by amplifying and filtering specific frequencies.

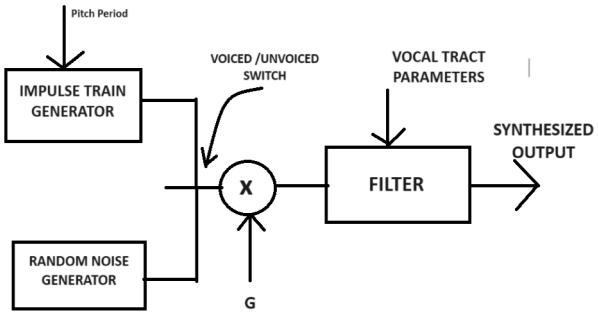

Figure 1.1: Block diagram of simplified model for speech production. [1]

1.3 Speech Synthesis

Production of speech is known as speech synthesis. Speech synthesis can be divided into two categories: parametric and concatenative. Concatenative synthesis creates the final speech by combining pre-recorded voice segments. Pitch, duration, and spectral properties are only a few of the parameters that are employed to influence the generation of a speech in parametric synthesis. Applications for synthesized speech include text-to-speech (TTS) systems, voice assistants, and communication aids. [5]

Chapter 2

Methodology

2.1 Vowel Selection

The first stage in estimating and synthesizing is choosing one male and one female vowel from the sample audio set. The male and female vowel audio sets for our experimental observation are the hood_m and hood_f audio sets.

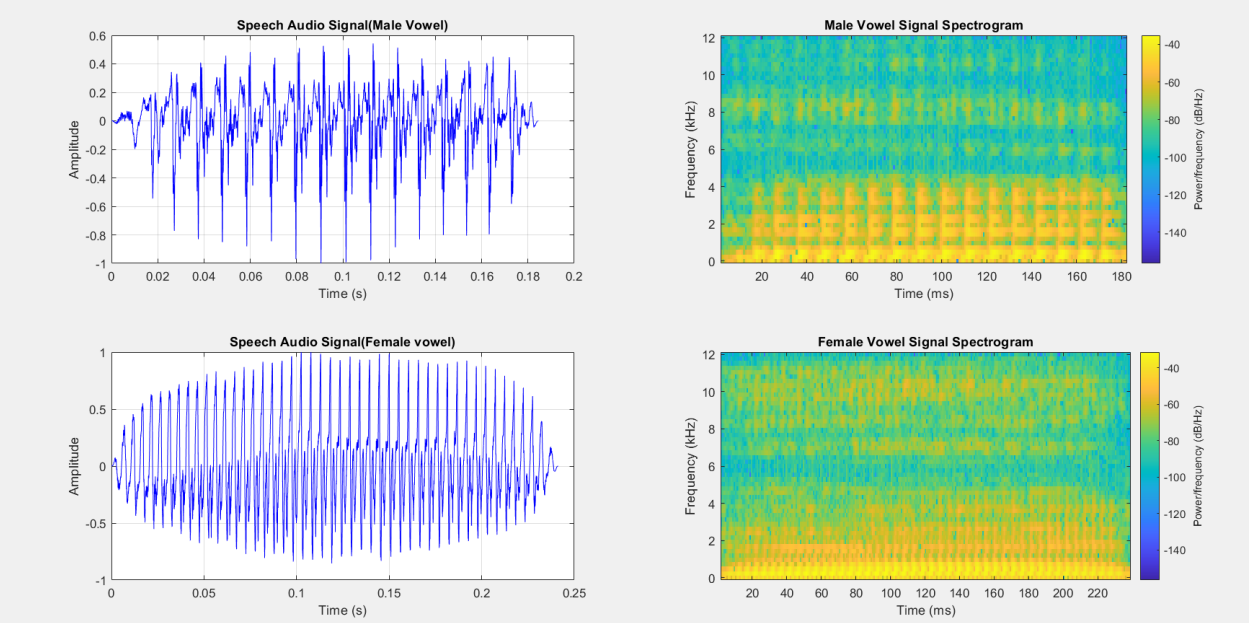

Figure 2.1: hood_m and hood_f.

2.2 Quasi-Stationary Segment Selection

Let's define quasi-stationary signals before we start the selection process. A subset of nonstationary signals known as quasi-stationary signals have statistics that are locally static across brief time intervals and vary from one time frame to the next. The vocal tract functions as a linear time-invariant system during speech generation, and these

time intervals are represented by 100 millisecond segments. The segment selection is crucial since it enables us to record the fundamental qualities of the vowels.

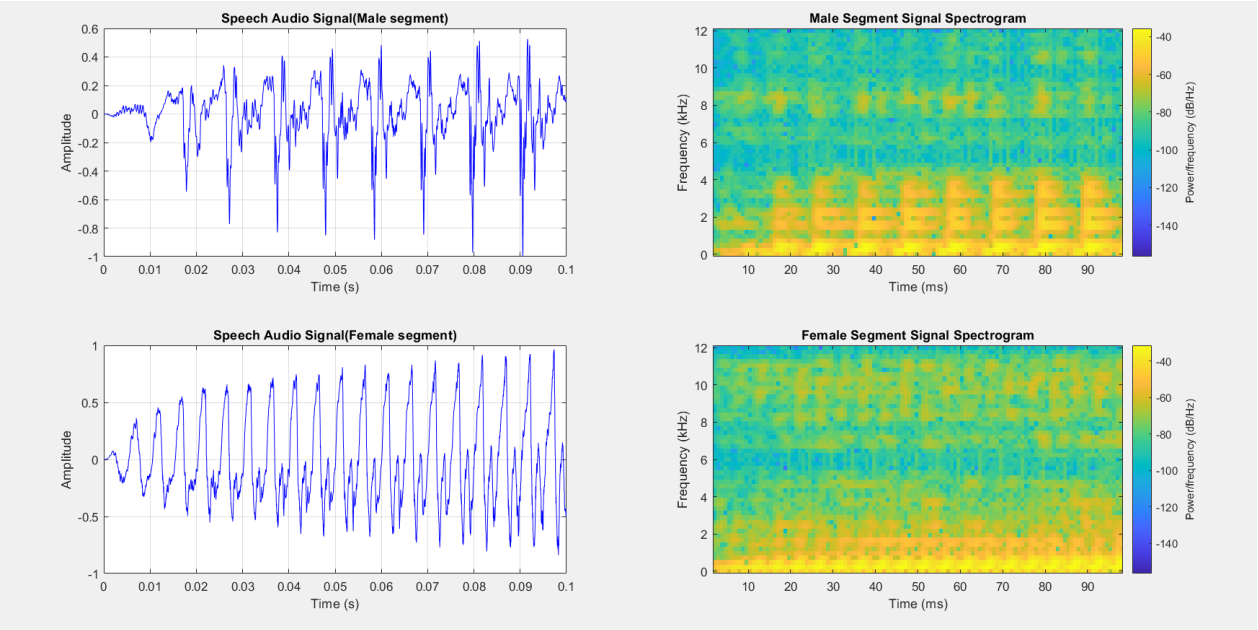

Figure 2.2: Male and Female Vowel Quasi-Stationary Segment of 100 ms.

Chapter 3

Linear Predictive Coding Estimation

3.1 Linear Predictive Coding

By simulating the vocal tract and the way sounds are formed during speech, a method known as linear predictive coding synthesizer can produce speech that is similar to that of a human. Speech synthesis and processing both make extensive use of the technique known as linear predictive ccoding (LPC). The speech production source-filter model serves as the foundation for LPC. It is assumed that the vocal tract filters the speech signal after the vocal cords create it. Every voice sample in the LPC model is a linear mixture of earlier samples. Using a linear function of earlier samples, it attempts to forecast the present sample. The linear prediction function's coefficients that minimize prediction error are estimated by LPC during analysis. Algorithms like the covariance approach and autocorrelation method are used for this. The vocal tract response is represented by the prediction coefficients. Thus, the vocal tract filter that converts the original sound into speech is efficiently modeled by LPC analysis. LPC generates speech during synthesis by utilizing the prediction coefficients. Synthetic speech is generated by passing an LPC vocal tract filter over a sound source signal.

3.2 Frequency Response Analysis

An LPC (Linear Predictive Coding) filter's frequency response simulates the features of the vocal tract for various speech sounds. Speech processing and synthesis can be used to explain the frequency response of a Linear Predictive Coding (LPC) filter for a male vowel or any other vocal sound. The vocal tract is modeled by LPC, a popular technique in speech analysis and synthesis, as a filter that modifies the sound spectrum during speech. The LPC filter's frequency response explains how various frequencies in the input speech signal are affected by this filter. The way the input signal's various frequencies are amplified or attenuated by the LPC filter is measured

by its frequency response. The way the filter acts over the audible spectrum is usually expressed as a function of frequency. In the frequency response of the LPC filter, the male vowels have distinct formants (resonant peaks) that change according to the vowel being said. There are variations in the frequency response between male and female speakers because male speakers often have different vocal tract layouts.

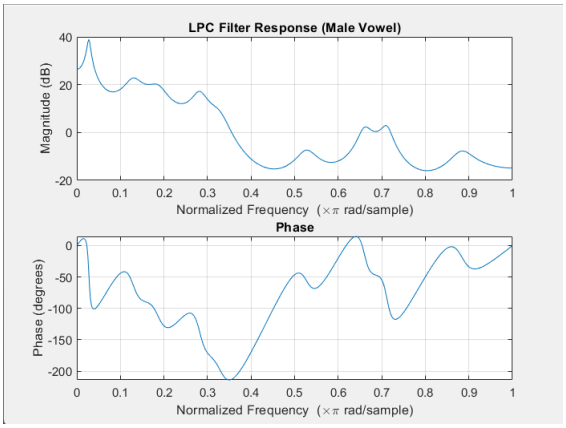

Figure 3.1: Frequency Response of LPC Filter for Male Vowel for LPC order 20.

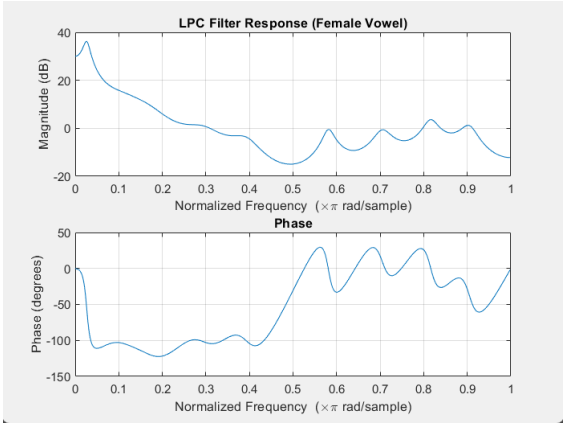

Figure 3.2: Frequency Response of LPC Filter for Female Vowel for LPC order 20.

3.3 Formant Frequency Estimation

Formants, which are prevalent in vowels, are high-energy frequency peaks in the spectrum. Every formant has a corresponding vocal tract resonance. You can think of formants as filters. The vocal cavities work together to produce a resonance effect that filters sound, amplifying certain frequency components and attenuating others.

```
Male Vowel Formant Frequencies (Hz):
Formant 1: 331.50 Hz
Formant 2: 1547.62 Hz
Formant 3: 2243.08 Hz
Female Vowel Formant Frequencies (Hz):
Formant 1: 318.25 Hz
Formant 2: 1341.95 Hz
Formant 3: 2056.32 Hz
```

Figure 3.3: Formant Frequencies of male and female vowel Quasi-stationary segments of 100ms for LPC order 20 .

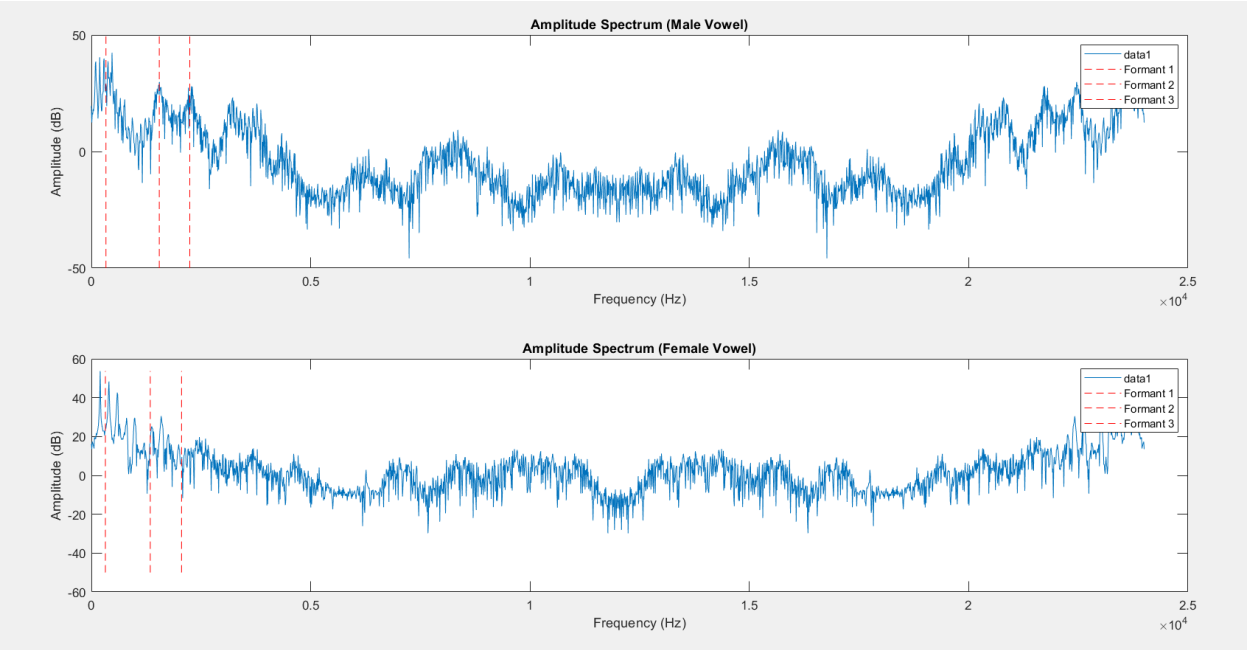

Figure 3.4: Amplitude spectrum of male and female vowel.

3.4 Fundamental frequency Estimation

The male pitch fundamental (F0) is lower, around 85-155Hz [3]. This affects the harmonic structure of the male vowel spectrum. The clustering of formants in the low frequency region and the lower F0 enhances the lower frequency content in the male vowel spectrum.

```
Estimated F0 for male (mean): 95.4706 Hz
Estimated F0 for female (mean): 200.6835 Hz
```

Figure 3.5: Fundamental Mean Frequency of male and female vowel for LPC order 20.

Pitch fundamental (F0) for females is usually higher than for males; it can range from about 165-255 Hz or even higher [3]. The spectral properties and harmonic structure of female vowel spectra are affected by this elevated F0 value.

Chapter 4

Synthesis

4.1 Impulse Train

An impulse train is a series of impulses. Each impulse is represented as an instantaneous and unit amplitude pulse. It is frequently employed in signal processing theory and analysis. It is also referred to as Dirac delta function. Mathematically, it is represented as a sum of shifted impulses :

$$\delta(t) + \delta(t - T) + \delta(t - 2T) + \delta(t - 3T) + \dots$$

where $\delta(t)$ is Dirac delta function, and T is the period of impulse train.

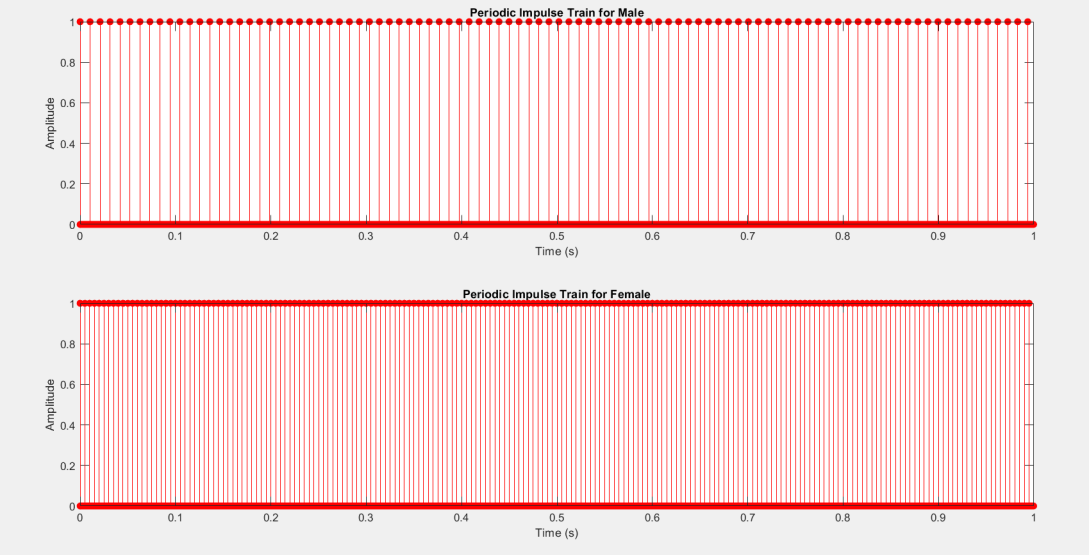

Figure 4.1: Impulse Train for Male and Female Vowel.

4.2 Filtering

It is an essential component in producing natural speech, which involves modeling the vocal tract's characteristics to mold the speech signal that is produced. Frequently

referred to as vocal tract modeling or formant filtering. The synthesized speech signal is created by passing the excitation signal—that is, the sound source—through the LPC filter. By convolving the excitation signal with the LPC transfer function, the synthesized or filtered signal can be generated. Expressed mathematically as: [2]

$$y[n] = \sum_{k=0}^{p} a_k x[n-k]$$

where, y[n] is the output speech signal , a_k is the LPC coefficients and $x[n-k]$ is the excitation signal delayed by k samples.

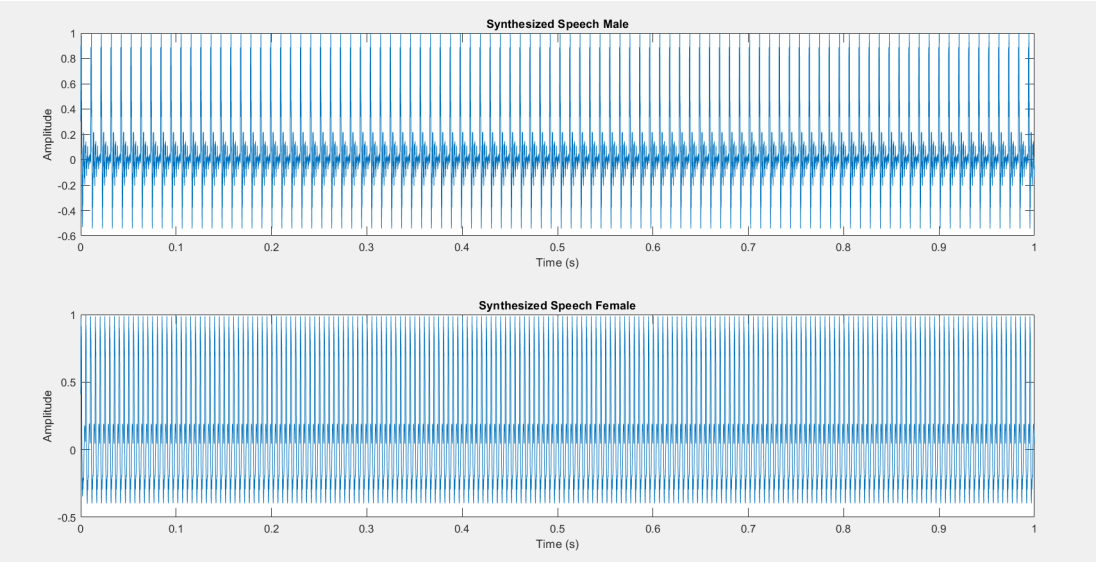

Figure 4.2: Synthesized speech for male and female vowel.

Chapter 5

Experimenting with AR Model Orders and Segment Lengths

Let's split it into four scenarios to experiment with segment lengths and AR model orders.

Case 1: Model order = 20, segment length = 125 ms

```
Male Vowel Formant Frequencies (Hz):
Formant 1: 340.51 Hz
Formant 2: 1537.38 Hz
Formant 3: 2252.14 Hz
Female Vowel Formant Frequencies (Hz):
Formant 1: 333.90 Hz
Formant 2: 1088.90 Hz
Formant 3: 2135.71 Hz
Estimated F0 for male (mean): 94.329 Hz
Estimated F0 for female (mean): 198.6012 Hz
```

Figure 5.1: Impulse Train for Male and Female Vowel.

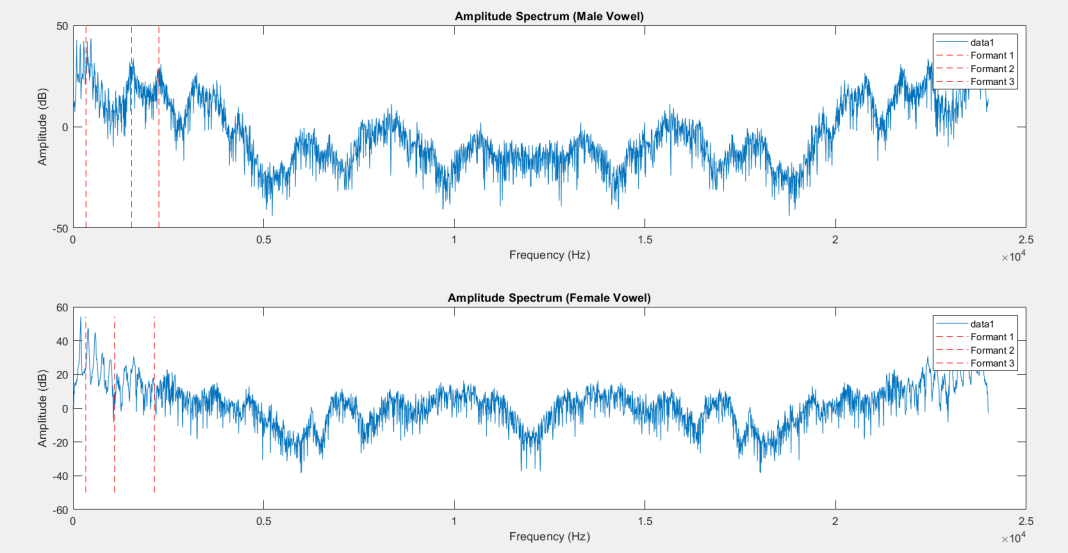

Figure 5.2: Impulse Train for Male and Female Vowel.

Case 2: Model order is 40 and segment length is 125 ms.

```
Male Vowel Formant Frequencies (Hz):
Formant 1: 226.19 Hz
Formant 2: 465.15 Hz
Formant 3: 1538.97 Hz
Female Vowel Formant Frequencies (Hz):
Formant 1: 249.71 Hz
Formant 2: 595.40 Hz
Formant 3: 1188.63 Hz
Estimated F0 for male (mean): 94.329 Hz
Estimated F0 for female (mean): 198.6012 Hz
```

Figure 5.3: Impulse Train for Male and Female Vowel.

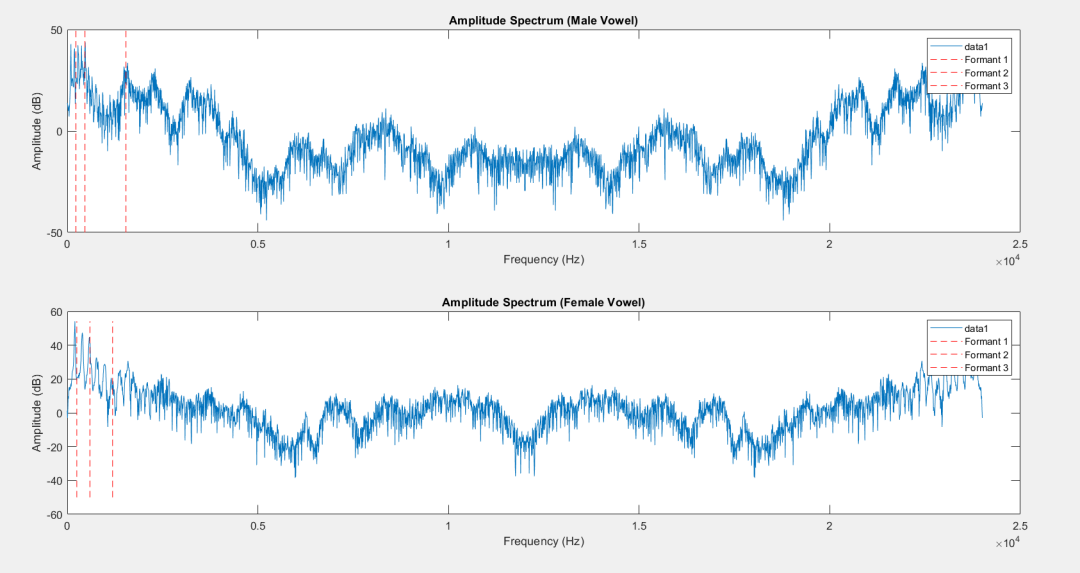

Figure 5.4: Impulse Train for Male and Female Vowel.

Case 3: 150 ms for the segment length and 20 for the model order

```
Male Vowel Formant Frequencies (Hz):
Formant 1: 342.47 Hz
Formant 2: 1544.41 Hz
Formant 3: 2263.54 Hz
Female Vowel Formant Frequencies (Hz):
Formant 1: 374.20 Hz
Formant 2: 1822.83 Hz
Formant 3: 2431.38 Hz
Estimated F0 for male (mean): 94.1041 Hz
Estimated F0 for female (mean): 196.9762 Hz
```

Figure 5.5: Impulse Train for Male and Female Vowel.

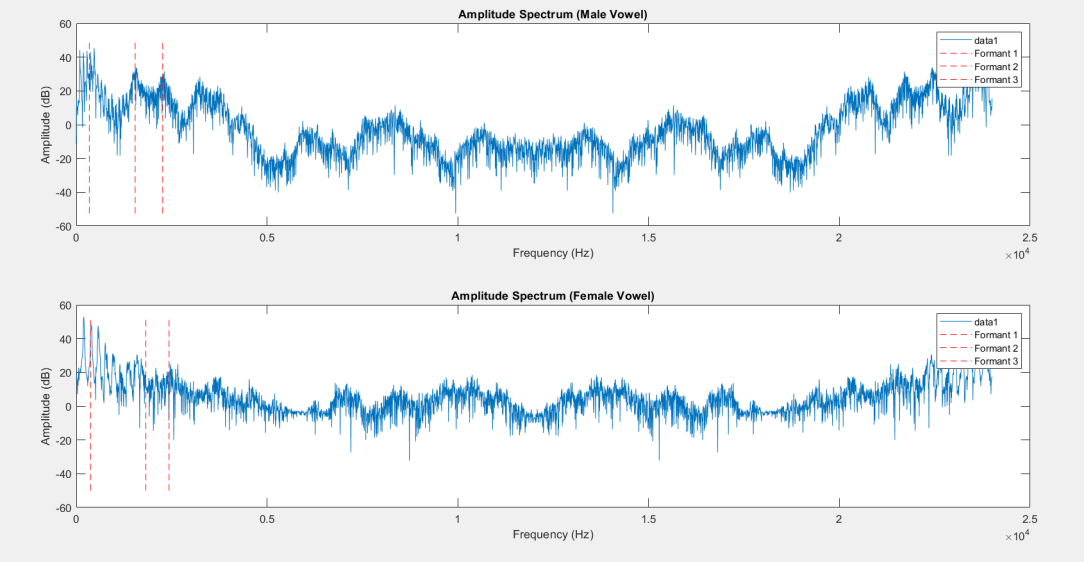

Figure 5.6: Impulse Train for Male and Female Vowel.

Case 4: Model order is 40 and segment length is 150 ms.

```
Male Vowel Formant Frequencies (Hz):
Formant 1: 224.63 Hz
Formant 2: 462.91 Hz
Formant 3: 1544.44 Hz
Female Vowel Formant Frequencies (Hz):
Formant 1: 251.51 Hz
Formant 2: 601.47 Hz
Formant 3: 1318.12 Hz
Estimated F0 for male (mean): 94.1041 Hz
Estimated F0 for female (mean): 196.9762 Hz
```

Figure 5.7: Impulse Train for Male and Female Vowel.

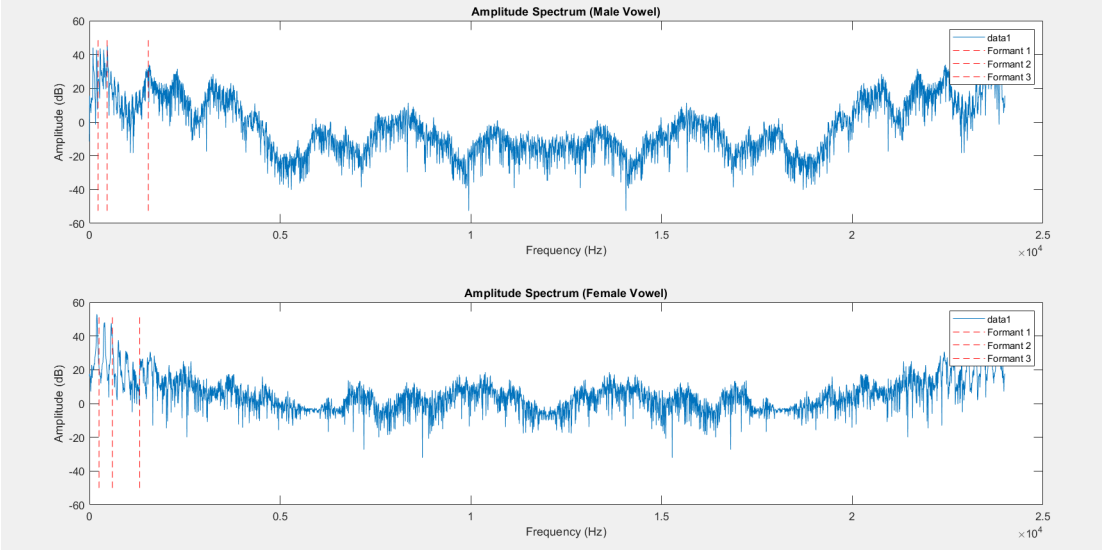

Figure 5.8: Impulse Train for Male and Female Vowel.

Chapter 6

Informal Assessment

What I discovered in the LPC filter model is that when the LPC order grows, the number of poles increases, resulting in better synthesized speech. Every pole also has a formant frequency connected with it. The synthesized output voice produces a loud beeping sound when the LPC order is decreased. The synthesized output speech is far superior to lower LPC order sounds when the LPC order is increased. It is evident that we can hear the vowel.

References

[1] L.R. Rabiner, R.W Schafer - Digital Procsesing of Speech Signals-Prentice Hall(1978) : Page 398

[2] L.R. Rabiner, R.W Schafer - Digital Procsesing of Speech Signals-Prentice Hall(1978) : Page 445

[3] https://en.wikipedia.org/wiki/Voice_frequency : Fundamental Frequency section

[4] https://en.wikipedia.org/wiki/Speech_synthesis : Synthesizer Technologies section

[5] Matlab functions given in the coursework assignment document