Linear Predictive Coding Analysis and Synthesis of Speech using MATLAB

Muhamad Izwan bin Yaacob
Faculty of Electrical Engineering (Electronics)
Universiti Teknologi Mara, Shah Alam, Malaysia
izwan710@gmail.com

Abstract—**The speech analysis and synthesis have many applications and is essential especially in transmission of signals due to its bandwidth availability. In this paper, a Linear Predictive Coding (LPC) is used to analyze the speech parameters and synthesis the production of speech by using stored speech parameters. Two female speech samples parameters are analyzed according to Voice Activity Detection (VAD), formant frequency estimation, and autocorrelation analysis and cepstrum analysis for fundamental frequency estimation. After the speech was synthesized, the quality was measured based on Signal-to-Noise Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ). All of this was done in MATLAB software environment.**

***Index Terms*—Linear Predictive Coding (LPC), Voice Activity Detection (VAD), Signal-to-Noise Ratio (SNR), Mean Opinion Score (MOS), Perceptual Evaluation of Speech Quality (PESQ).**

I. INTRODUCTION

THE rapid development of wireless communication has challenged the service provider to optimize the number of subscriber as well as to maintain the quality of service with limited allocated bandwidth. Speech coding might be the solution to provide a low bit rate and low bandwidth with a reasonable level of speech quality. [1]

In general, speech coding is a modification technique to represent a digitized speech signal using as few bits as possible. Technically, speech coding is a lossy type of coding where the inputs and the output signal can be distinguished to be different because it is a data encoding method which compresses data by discarding some of it. The advantage of lossy methods over lossless methods is that in some cases a lossy method can produce a much smaller compressed file than any lossless method, while still meeting the requirements of the application. Usually, the telephony speech signal is coded with 0 to 8Hz frequency range with 16 kHz sampling frequency as it is regarded as sufficient for speech recognition and synthesis. [1], [2]

In this project, LPC technique is used to perform analysis and synthesis on two female speech samples. The main features of the female speech samples are extracted, and split into segments to initialize the speech synthesis and analysis which are to determine the pitch of the signal (basic formant frequency estimation), Voice Activity Detection (VAD) and also the LPC coefficients. Later on the synthesized speech quality was measured based on Signal-To-Noise Ratio (SNR) and the predicted Mean Opinion Score (MOS).

This paper is organized as follows: In section II. Speech Theory, we will discuss briefly on the theories that applies in speech coding. Section III. Analysis and Synthesis Method discusses on the methods taken to perform analysis and synthesis on the speech sample. Section IV. Results and Discussions provides the results obtained from the analysis and synthesis of the speech signal and a throughout discussions on it. Section V. Conclusions provides the conclusions on this project, and Section VI. References provide the list of references used for this project to materialize.

II. SPEECH THEORY

A. Speech Model

The speech sample needs to be modeled so that analysis and synthesis can be performed. Figure 1 shows a schematic representation of the physiological mechanism of speech production.

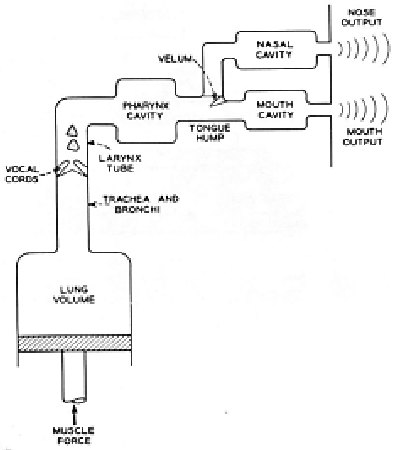

Figure 1: Schematic Representation of the physiological mechanism of speech production.

The human speech production system can be modeled using a simple structure where the lungs generate the air or energy to excite the vocal tract which is represented by a white noise source. The acoustic path composed by the larynx, vocal cords, pharynx, mouth and nose, can be modeled by a time-varying digital filter [3] which is the LPC filter.

B. Linear Predictive Coding

In this section, we will discuss on how LPC model can be used as the speech model. LPC model can be assumed as a linear combination of the past speech signal. This idea can be represented as in Eq. (1).

$$s(t) \approx a_1 s(t-1) + \cdots + a_p s(t-p) \qquad (1)$$

Where *s(t)* is the speech signal at time, *t*, while *p* represents the order of the LPC filter, and *a* is the LPC coefficient. The coefficients $a_1, a_2, \dots, a_p$ are assumed as constant over the speech analysis frame.

In this project, we are using a female speech sample with a sampling frequency, *fs* of 16 kHz with an LPC order, *p* of 10.

We convert Eq. (1) to an equality by including an excitation term, *Gu(n)*, giving:

$$s(t) = Gu(t) + \sum_{i=1}^{p} a_i s(t-1) \qquad (2)$$

Where *u(t)* is a normalized excitation and *G* is the gain of the excitation. By expressing Eq. 2 in the z-domain we get the relation:

$$S(z) = GU(z) + \sum_{i=1}^{p} a_i z^{-1} S(z) \qquad (3)$$

Leading to the transfer function of the LPC filter which is:

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \qquad (4)$$

As the LPC order is 10, the summation is computed starting at *i*=1 up to 10, this means that only the first 10 coefficients are transmitted to the LPC synthesizer. [4] The actual excitation function for speech is essentially a quasi-periodic pulse train (for voiced speech sounds) or a random noise source (for unvoiced sounds. [3]

III. ANALYSIS AND SYNTHESIS METHODS

In this section, the methods taken to accomplish the desired objectives are going to be explained. Linear Predictive Coding (LPC) is being applied for the speech analysis and synthesis, and the system is to be implemented by using MATLAB software. The implemented system will be achieved based on the flowchart in Figure 2.

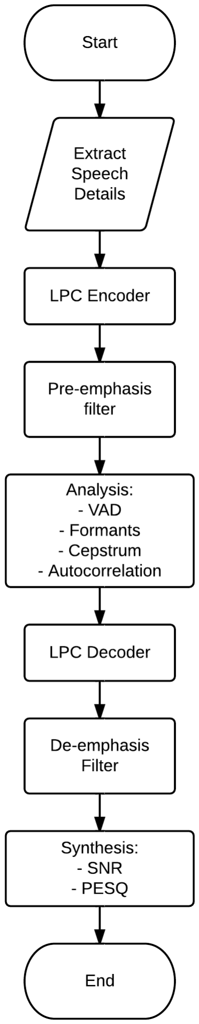

Figure 2: LPC Analysis and Synthesis Flowchart

From the flowchart, the analysis and synthesis methods taken on the speech signal will be discussed. In this project, the speech sample was obtained from two female speech samples file. The first process is to extract the details from the sample and then the data was then entered into MATLAB for the speech sample to be digitized. The details are the value of the sampling frequency and also the speech data itself which are represented as *fs* and *x* respectively.

The obtained features of the speech samples were then transferred to the encoder. As in an LPC, the digitize signals must be split into segments. And also in this LPC encoder, we have determined the LPC coefficients by using the Levinson-Durbin algorithm. Once the voiced and unvoiced section are determined, encoding consists of deriving the optimal LPC coefficients ($a_1 \dots a_p$) for the vocal tract model so as to minimize the mean-square error between the predicted signal and the actual signal. [3]

Next is the analysis part of the speech sample. Figure 3 shows the block diagram of the analysis part of the LPC system.

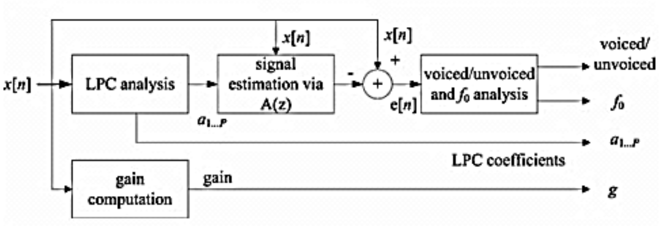

Figure 3: Analysis part of the LPC system

In this analysis part, we are able to determine the parameters required for the speech signal analysis such as the VAD and the formant frequencies.

To observe the speech as voiced or unvoiced signal, the voice activity detector (VAD) classifies a frame as voiced or unvoiced. Typically voiced sounds are of a higher magnitude in energy compared unvoiced sounds.

Subsequent is to determine the formant frequency estimation. To find the formant frequencies from the filter, we need to find the locations of the resonances that make up the filter. This involves treating the filter coefficients as a polynomial and solving for the roots of the polynomial. From this estimation, we are able to determine the pitch of the signal.

Next is to examine the fundamental frequency estimation in frequency domain or also known as the Cepstrum Analysis. The cepstrum method for pitch analysis is also based on the DFT or rather the inverse-DFT of the log of the DFT to be more accurate. The algorithm itself is quite straightforward and is shown in Eq. (4).

$$\hat{s}[n] = DFT^{-1}\{\log DFT(x[n])\} \quad (4)$$

$\hat{s}[n]$ is the real cepstrum component of the inverse DFT having the units of quefrency corresponding to the unit of time, which can then be analyzed for peaks rendering the fundamental frequency. [5]

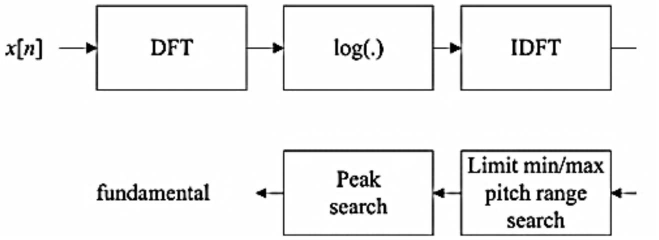

Figure 4: Fundamental frequency computation flowchart via the cepstrum.

Also part of the analysis is the fundamental frequency estimation in time domain by using autocorrelation method. The autocorrelation function of a random signal describes the general dependence of the values of the samples at one time on the values of the samples at another time. Consider a random process x(t) (i.e. continuous-time), its autocorrelation function is written as:

$$R_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} x(t)x(t+\tau)dt \quad (5)$$

Where T is the period of observation and $R_{xx}(\tau)$ is always real-valued and an even function with a maximum value at $\tau = 0$. **[10]**

And finally is to perform the short-time frequency analysis by displaying the spectrogram of the speech sample. The use of spectrograms and a pattern playback for research on speech has the unique advantage that it permits the study of isolated acoustic cues for speech perception. This method has shown that the consecutive sounds of the language are usually so intimately connected that they cannot be separated and recombined in a different order without serious loss in intelligibility. [5]

As the LPC analysis is performed on the speech sample, we are now moving into the encoder function. The values needed (*aCoeff, pitch_plot, voiced,* and *gain*) for the next stage which is the decoder was obtained from the encoder. The purpose of the decoder is to reconstruct the original speech sample based on the LPC coefficients, the pitch and other parameters encoded by the LPC encoder. [3] By doing so, the speech sample is being synthesized. The block diagram of the synthesis part of the speech sample is as shown in Figure 5.

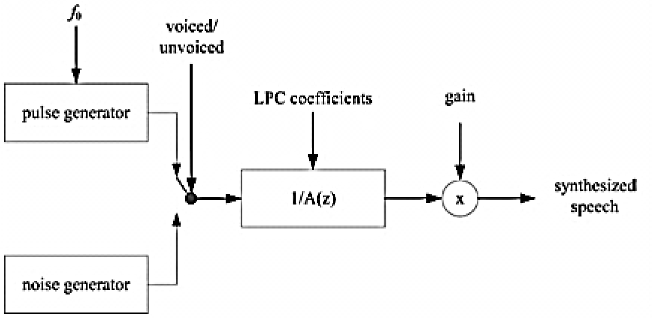

Figure 5: Synthesis part of LPC system

From the block diagram, the methods taken to synthesize the speech sample are shown. From the encoder, we have determined the necessary parameters, which are the voiced/unvoiced, LPC coefficients, and the gain. These parameters are essential in performing synthesis on the speech samples.

The synthesized speech samples were then measured to assess the quality. The speech quality assessments were based on SNR and the predicted MOS score from Perceptual Evaluation of Speech Quality (PESQ) method.

To measure SNR, random noise signals are needed to be added to the speech samples. Eq. (6) was utilized to calculate the resulting SNR for both the original and synthesized speech and then compared with each other's.

$$SNR_{dB} = 10\log_{10}\left(\frac{A_{signal}}{A_{noise}}\right)^2 = 20\log_{10}\left(\frac{A_{signal}}{A_{noise}}\right) \quad (6)$$

And lastly, the resulting synthesized speech signal was compared to the original speech samples to compute the predicted MOS by using the PESQ method. PESQ predicts the results of subjective listening tests as from MOS on telephony systems including from speech coding. PESQ uses a sensory model to measure the speech quality where the original uncompressed signal was compared with the degraded version of the signal at the output of the communication system.

IV. RESULTS AND DISCUSSIONS

A. *Voice activity detection (VAD)*

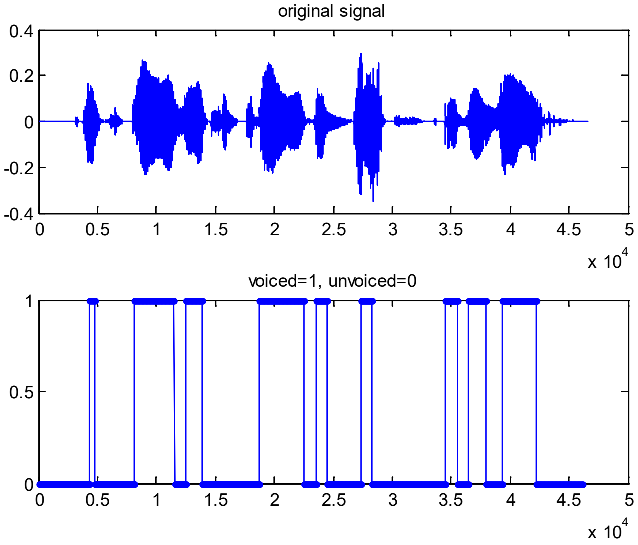

Figure 6: VAD of speech Sample 1.

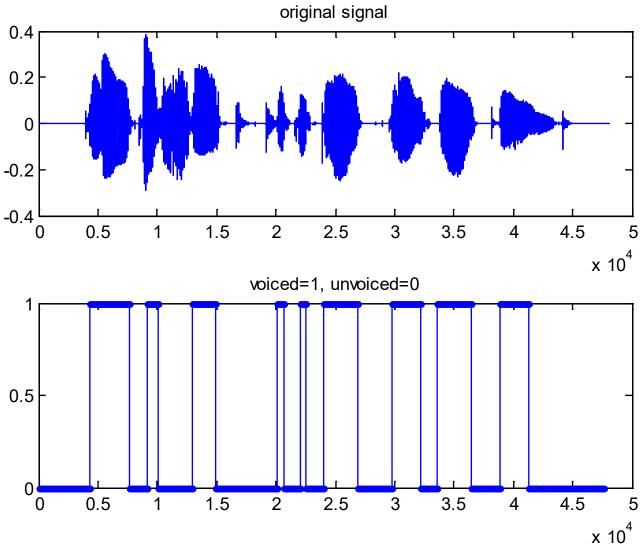

Figure 7: VAD of speech Sample 2.

Voice Activity Detection (VAD) determines which parts of the speech samples are considered voiced, and which part is considered unvoiced. From the bottom figure in Figure 6 and Figure 7, 1 indicates voiced section while 0 indicates unvoiced section. VAD can facilitate speech processing, and can also be used to deactivate some processes during non-speech section of an audio session where it can avoid unnecessary coding and transmission of silence packets, saving on computation and on network bandwidth.

B. *Cepstrum Analysis*

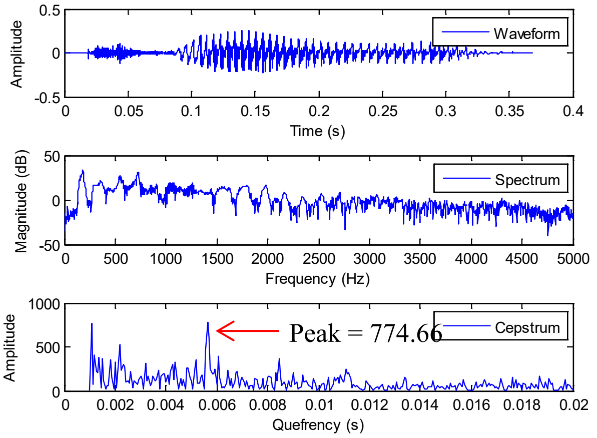

Figure 8: Time sequence (top), power spectrum (middle), cepstrum (bottom) for Sample 1.

Figure 8 shows the time sequence, middle figure the power spectrum in dB, and the bottom figure the cepstrum for Sample 1. The fundamental frequency is commonly computed by finding the maximum peak for a given region in the cepstrum; that is, limiting the search to a minimum and maximum frequency range for pitch. For the speech sample cepstrum, the peak in the cepstrum between 1 ms and 20 ms is 774.66 and when converted to hertz is 177.778 Hz. The voiced speech of a typical adult male will have a fundamental frequency from 85 to 180 Hz, and that of a typical adult female from 165 to 255 Hz [35] [36]. So, the value of the fundamental frequency obtained for Sample 1 is located in the standard range.

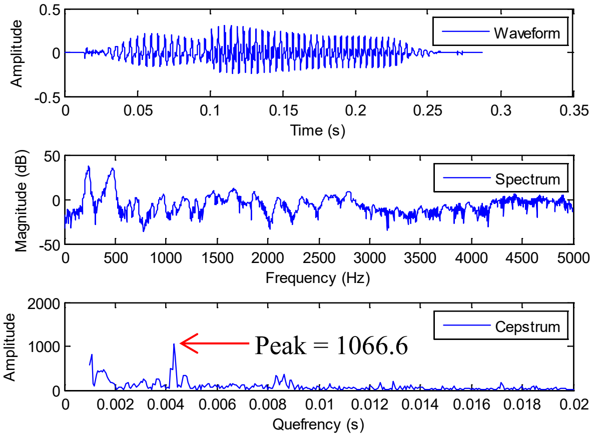

Figure 9: Time sequence (top), power spectrum (middle), cepstrum (bottom) for Sample 2.

As for Sample 2, the peak of the cepstrum is at the amplitude of 1066.6 and when converted to Hertz is 231.884 Hz which is also in the range of the theoretical fundamental frequency for female speech.

C. Autocorrelation

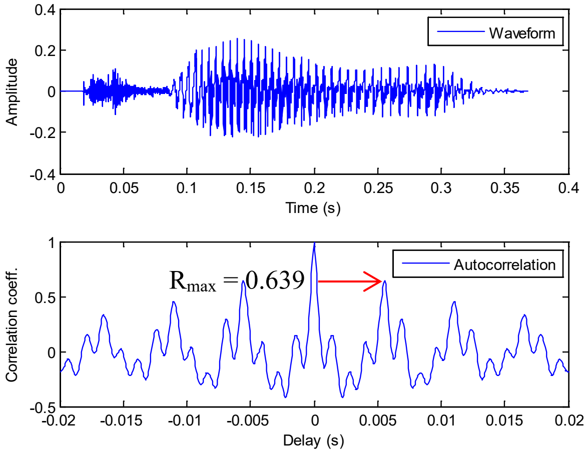

Figure 10: Autocorrelation of the speech Sample 1.

From Figure 10, the fundamental frequency in the time domain from the waveform was directly estimated. We can see that the autocorrelation function peaks at zero delay and at delays corresponding periods. We can estimate the fundamental frequency by looking for a peak in the delay interval corresponding to the normal pitch range in speech. At region corresponding to positive delays, we have determined that the maximum correlation coefficient, R_{max} is at 0.639145 and a frequency, F_0 of 177.778 Hz which is the fundamental frequency.

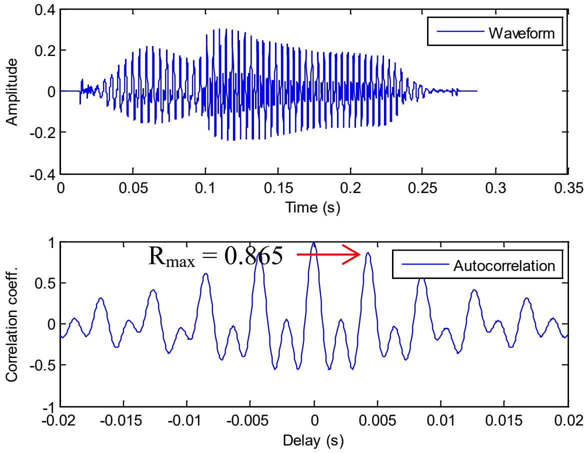

Figure 11: Autocorrelation of the speech Sample 2.

Figure 11 shows the autocorrelation analysis for Sample 2. By using the similar method as done on Sample 1, the determined maximum correlation coefficient, Rmax is at 0.86549 and when converted to Hertz is equal to 231.884 Hz which is exactly the same as when using the previous method.

D. Formant frequencies estimation

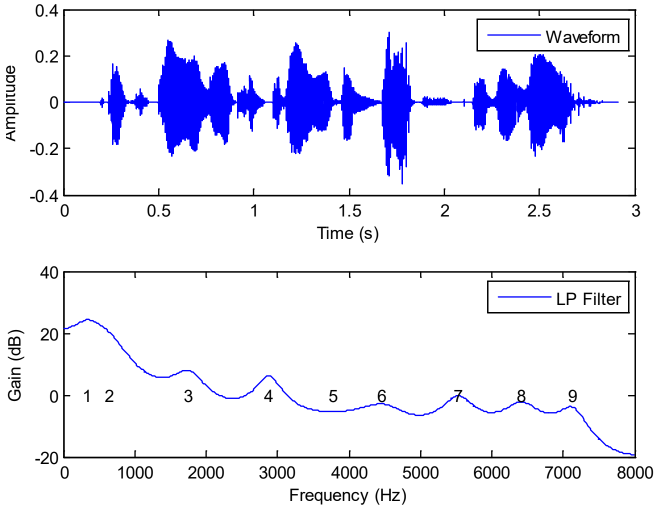

Figure 12: Formant frequencies for the speech Sample 1.

From the formant frequencies estimation, 9 points of formant frequencies which represents the pitch of the signal was estimated. Table 1 shows the values of the frequency at each formant points based on Figure 9. These points are the locations of the resonances that make up the filter.

TABLE 1
FORMANT FREQUENCIES OF SPEECH SAMPLE 1

Formant	Frequency (Hz)
1	340.9
2	662.7
3	1765.1
4	2881.6
5	3782.8
6	4475.7
7	5530.4
8	6418.8
9	7141.1

The same method was the applied for Sample 2. Figure 13 shows the position of the formants which is labeled from 1 to 8. The values at each point are then entered into Table 2. The only difference compared from the formant estimation for Sample 1 is that this sample contains 8 formant points. This is simply because the sample used is different and different samples contain different details and characteristics.

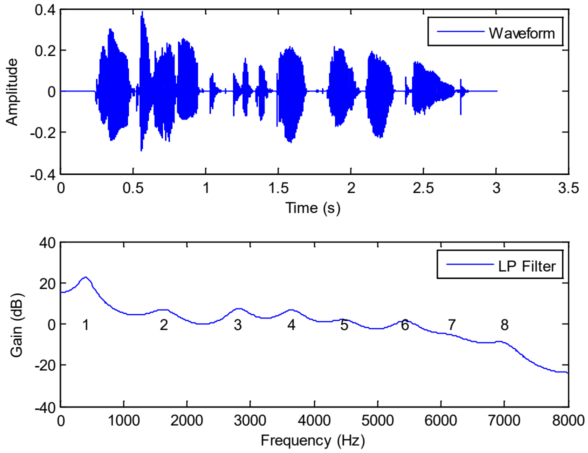

Figure 13: Formant frequencies for the speech Sample 2.

TABLE 2
FORMANT FREQUENCIES OF SPEECH SAMPLE 2

Formant	Frequency (Hz)
1	415.6
2	1651.1
3	2815.6
4	3565.2
5	4489.4
6	5441.7
7	6176.6
8	7008.8

E. *Spectrogram*

Figure 14 and Figure 15 shows the spectrogram of Sample 1 and Sample 2 respectively. Both results are obtained by using the same method.

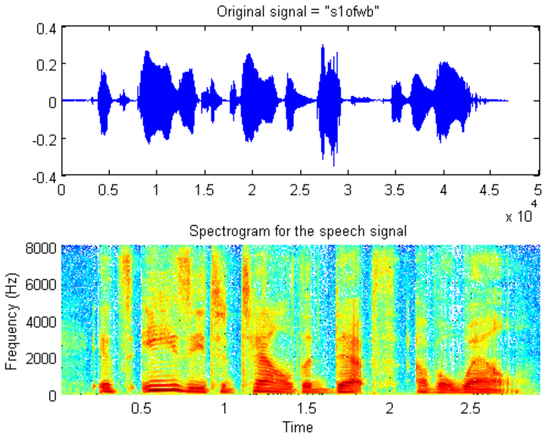

Figure 14: Spectrogram of speech Sample 1.

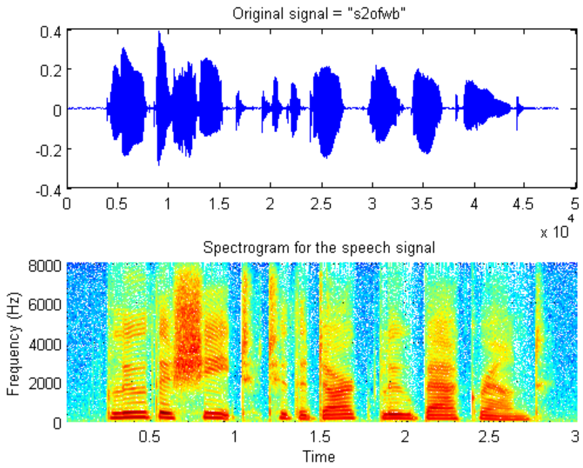

Figure 15: Spectrogram of speech Sample 1.

Blue indicates low energy portion of the spectrum, with red indicating the most energetic portions. The speech samples contain significant energy from zero frequency up to around 8 kHz. If the periodicity is sustained, we will generally see a horizontal line forming in the spectrogram. If it is not periodic (horizontal) in the time-domain, it will show up as a flat spectrum in the DFT frame and a block of noise in the spectrogram which is pattern-less and vertical. Thus, in our spectrogram plot, armed with these very basic attributes regarding signals, we can observe that non-patterned (non-periodic) attributes or noise-like sounds are present. While in the other parts, pitched signals with harmonics are observable with occasional low frequency modulation as well as amplitude modulation. [4]

F. *Speech Synthesis*

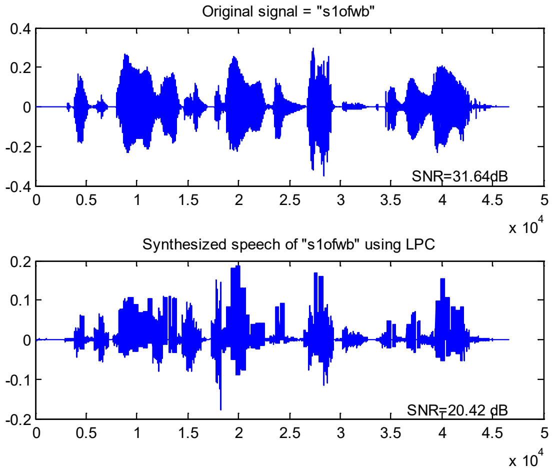

Figure 16: Speech sample of the original signal (top) and the compressed signal (bottom) for Sample 1.

From Figure 16, we are able to witness the effect on LPC synthesis on the original speech sample. The synthesized speech signal (bottom) appears to have less data and appears jagged due to the loss in the data encoding method which compresses data by discarding some of it since we are using a lossy method which is LPC.

The SNR for both original and synthesized speech Sample 1 was determined and are presented in Table 3. SNR values are usually in the range of 0 dB to 60 dB. Values outside this range will be limited prior to mapping to quality. [13] The value of SNR obtained from Sample 1 is considered good since it is in the theory range. The resulting score of PESQ is 1.60 which falls into the Poor quality based on the MOS score.

TABLE 3
SNR FOR SPEECH SAMPLE 1

Sample 1	SNR (dB)
Original	31.64
Synthesized	20.42

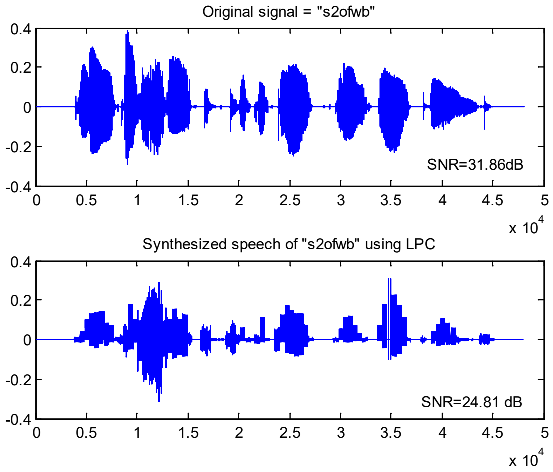

Figure 17: Original and synthesized of speech Sample 2.

The same method was also used to determine the quality for speech Sample 2. The resulting SNR values are as in Table 4. It can be observed that the value of SNR decreases for the synthesized speech signal just like in Sample 1. This is also due to the loss of data during the synthesis process.

TABLE 4
SNR FOR SPEECH SAMPLE 2

Sample 2	SNR (dB)
Original	31.86
Synthesized	24.81

The PESQ score obtained for speech Sample 2 is 1.73 which is a slight increase in number compared to the ones for speech Sample 1 but it still considered in the Poor quality.

V. CONCLUSIONS

In this project, LPC was utilized to analyze and synthesize two speech samples by using MATLAB software. From the analysis, the VAD, fundamental frequencies, formant frequencies, and also the spectrogram analysis was obtained. From this, the significant characteristics based on the parameters analyzed have been determined so that a reasonable level of speech quality after the speech sample was synthesized can be achieved. And from the synthesis part, the SNR values and also the predicted MOS scores based on the PESQ method has been determined. Based on this assessment, the synthesized speech signal will be able to be classified based on their quality. In the future, study and application of LPC can be expanded in speech recognition, underwater communications, and other applications as well.

VI. REFERENCES

[1] T. P. Barnwell III, K. Nayebi and C. H. Richardson, *SPEECH CODING, A computer Laboratory Textbook*, John Wiley & Sons, Inc. 1996.

[2] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall (Signal Processing Series), 1978.

[3] Pablo Gomez, *"Speech Coding & Linear Prediction Coding (LPC)"*, Florida International University, 2004

[4] Tae Hong Park, *Introduction To Digital Signal Processing: Computer Musically Speaking*, World Scientific Publishing, 2010.

[5] Poor, H. V., Looney, C. G., Marks II, R. J., Verdú, S., Thomas, J. A., Cover, T. M. *Information Theory. The Electrical Engineering Handbook,* 2000.

[6] A. Spanias, "*Speech coding: a tutorial review*," Proc. IEEE, vol. 82, pp 1541-1582, 1994.

[7] Levelt, WJ (1999). "*Models of word production*." Trends in cognitive sciences 3 (6): 223–232. doi:10.1016/S1364-6613(99)01319-4. PMID 10354575.

[8] R. Sproat, and J. Olive. *Text-to-Speech Synthesis, Digital Signal Processing Handbook*, 1999.

[9] Wai C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, Wiley Inter-science.

[10] Jerry D. Gibson, "*Speech Coding Methods, Standards, and Applications"*, University of California, Santa Barbara, 2000

[11] Jonathan Harrington, Steve Cassidy, "*Techniques in speech acoustics*", Springer, 1999

[12] Amol R. Madane, Zalak Shah, Raina Shah, Sanket Thakur, "*Speech Compression Using Linear Predictive Coding"*, MIR Labs, 2007

[13] Massimo Tistarelli, Mark S. Nixon, *Advances in Biometrics: Third International Conferences*, Springer, 2009.