EE473 Final Project

Enes Arda 2017401111

Linear Production Model

Linear Production Model is one of the most fundamental tools in speech analysis. It suggests that a sample of a speech signal at a time k can be approximated as a linear combination of its past values k-1, k-2… and an input function. Here the idea of representation relies on the assumption that a speech signal is produced by exciting a filter, which represents the vocal-tract, by a random noise (for unvoiced speech) or by an impulse train (for voiced speech) [5].

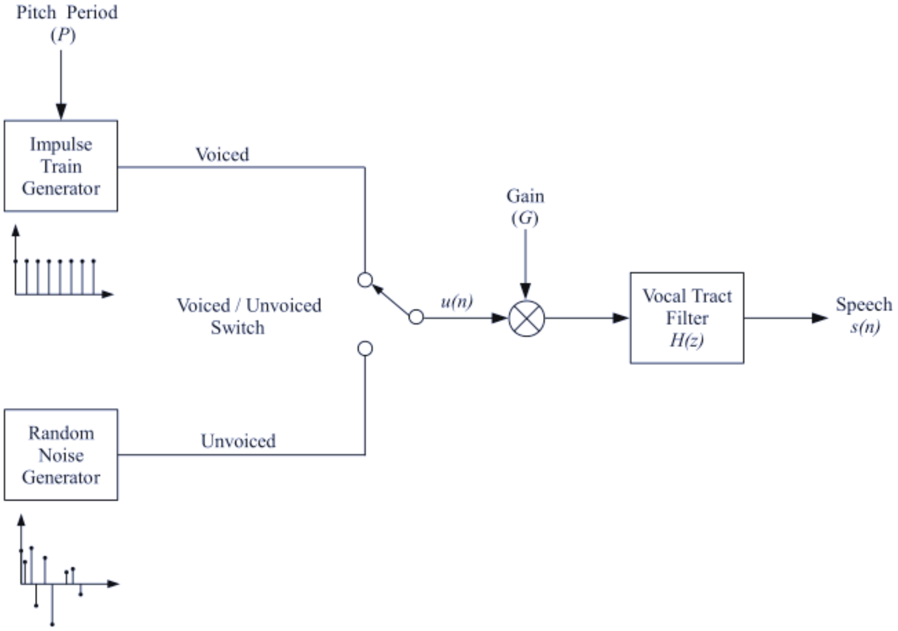


Fig1. Linear Production model for speech

A general predictor form is the ARMA(p, q) model where the current output depends on the p most recent output, current input and q most recent inputs. Autoregressive-moving average process ARMA(p, q) can be written in the following form:

$$s(n) = \sum_{i=1}^{p} \phi_{p,i} s(n-i) + G \sum_{i=1}^{q} \theta_{q,i} u(n-i) + Gu(n) \qquad \text{where G is the gain of the filter}$$

By minimizing the energy of the error between predicted values and actual values over a finite interval a unique set of coefficients for linear combination can be determined. In frequency domain the transfer function of this model can be written as:

$$H(z) = \frac{1 + \sum_{i=1}^{q} \theta_{q,i} z^{-i}}{1 - \sum_{i=0}^{p} \phi_{p,i} z^{-i}}$$

H(z) in this form is referred to as pole-zero model. Here the zeros represent the nasals and poles represent the resonances of the vocal tract. Since one needs to solve a set of non-linear equations to get the coefficients for a pole-zero model LP model in this form is prohibitive. Instead an all-pole model is preferred [1]. This is achieved when $\theta_{q,i} = 0$ for all i. Hence ARMA(p, q) process becomes AR(p) process that can be written in the following form:

$$s(n) = \sum_{i=1}^{p} \phi_{p,i} s(n-i) + Gu(n)$$

This model is preferred because it is computationally much faster as it only requires to solve a set of linear equations. Furthermore it can model sounds such as vowels well enough as the location of poles are considerably more important than the location of zeros [2]. Hence the transfer function becomes:

$$H(z) = \frac{G}{1 - \sum_{i=1}^{p} \phi_{p,i} z^{-i}}$$

The residual signal is defined as the difference between real speech and predicted speech.

$$e(n) = s(n) - \sum_{i=1}^{p} \phi_{p,i} s(n-i)$$

The coefficients of AR(p) that minimizes the energy of the residual can be found by solving Yule-Walker equations. However, in order to use this approach process should be wide-sense stationary. Unfortunately speech signals are not wide-sense stationary as a whole. Nevertheless a short frame of a speech signal can be considered to be wide-sense stationary. Therefore windowing is utilized to split the speech signal into 25ms frames, an interval which the speech signal can be considered WSS. Hamming window is used as the windowing function and each frame is overlapped to the next frame. Using this windowing technique short term filter coefficients can be calculated for each frame.

The Yule-Walker equations that gives the AR coefficients to minimize the energy of the residual are as follows:

$$\begin{bmatrix} R_s(0) & R_s(1) & R_s(2) & R_s(3) & \dots & R_s(p-1) \\ R_s(1) & R_s(0) & R_s(1) & R_s(2) & \dots & R_s(p-2) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ R_s(p-1) & R_s(p-2) & R_s(p-3) & R_s(p-4) & \dots & R_s(0) \end{bmatrix} \begin{bmatrix} \phi_{p1} \\ \phi_{p2} \\ \vdots \\ \phi_{pp} \end{bmatrix} = \begin{bmatrix} R_s(1) \\ R_s(2) \\ \vdots \\ R_s(p) \end{bmatrix}$$

$$\mathbf{R}_s\phi_p = \mathbf{r}_s$$

The solution to this equation is $\phi_p = \mathbf{R}_s^{\dagger}\mathbf{r}_s$. Hence the coefficients can be found by solving linear equations.

First Approach to Pitch Detection

As stated in [3] pitch is the perception of how high or low an audio signal sounds, which can be regarded as frequency that corresponds closely to the fundamental frequency of the speech signal. The pitch detection algorithms therefore aim to detect the fundamental frequency of a quasi-periodic speech signal. Although it may seem simple there are many different methods for pitch detection and none of them have been reported to be error free for any signal [4]. One of the most important reasons for that is that the assumption that speech signals are quasi-periodic is often erroneous. Therefore as stated before a small frame of the speech signal where it can be considered to be periodic and WSS is used to detect the pitch.

As stated in [7] speech can be divided into two categories as voiced and unvoiced speech. In the voiced speech the vocal cords of the speaker vibrate as the sound is made. However in the unvoiced speech vocal cords do not vibrate. This is the reason why in Fig1 the linear prediction model uses an impulse train as the input for the voiced speech and a random noise for the unvoiced speech. Hence the fundamental frequency of a speech signal, its pitch, is the fundamental frequency of the vibration of the vocal cords.

Pitch detection algorithm can also be used to detect if a signal is voiced or unvoiced. As stated before unvoiced speech signals are noise-like. Therefore their periods are expected to be very low. On the other hand voiced speech signals are periodic with the period of the vibration of the vocal cords and therefore the periods of the voiced signals are expected to be higher than the unvoiced signals. Hence by determining a threshold value a speech signal can be classified as voiced or unvoiced using its period.

Pitch detection algorithms can work in time domain, in frequency domain or both. In this project first of all the autocorrelation method is used that is suggested in [7], which is a time domain approach. If the small frame of a speech signal is periodic $s(n) = s(n+P)$ its autocorrelation function is also periodic with the same period $R_s(n) = R_s(n+P)$. This can be observed in the following figure.

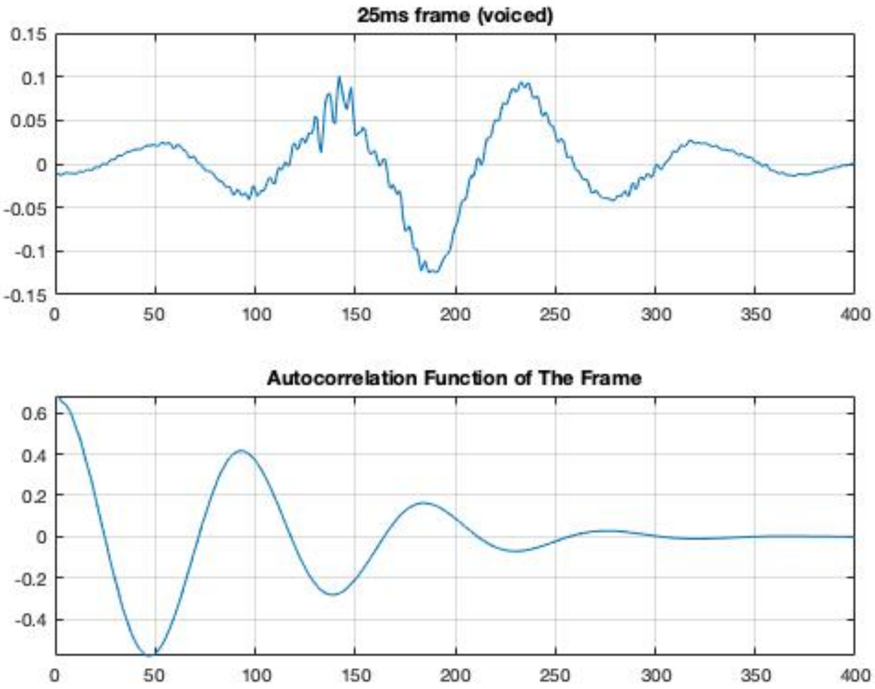


Fig2. 25ms voiced frame of a speech signal and its autocorrelation

As one can see from Fig2 the autocorrelation of a speech signal can be used to find the fundamental frequency of a frame. The peaks of the autocorrelation function can be detected and the difference between the subsequent peaks can be regarded as the period. However this approach has some problems. For example in the following figure:

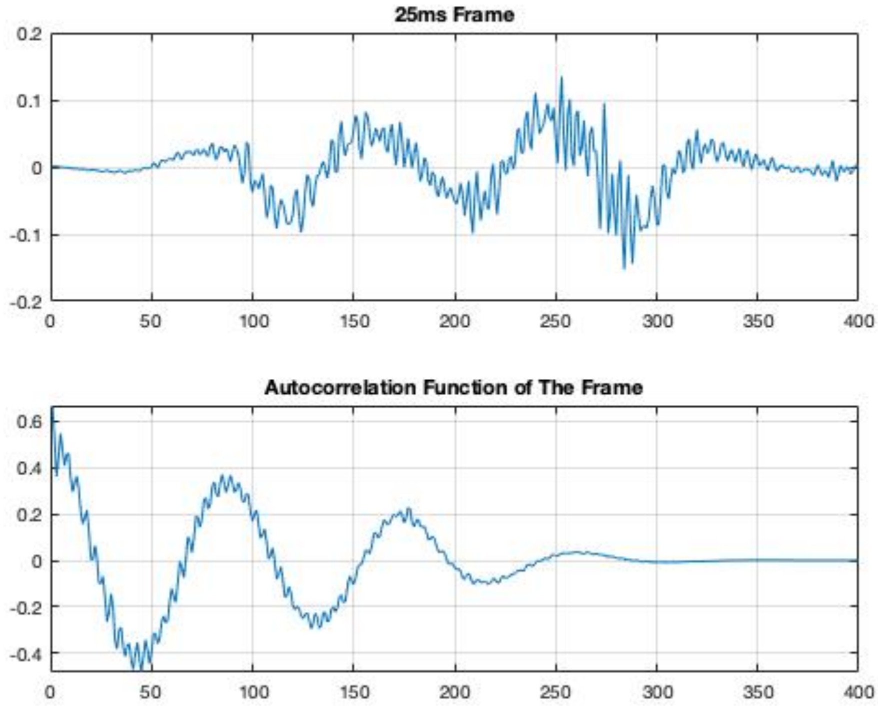


Fig3. 25ms frame of a speech signal and its autocorrelation

It is very hard to detect the peaks of the autocorrelation function in Fig3. To overcome this problem the autocorrelation of the autocorrelation function is taken to smoothen the function.

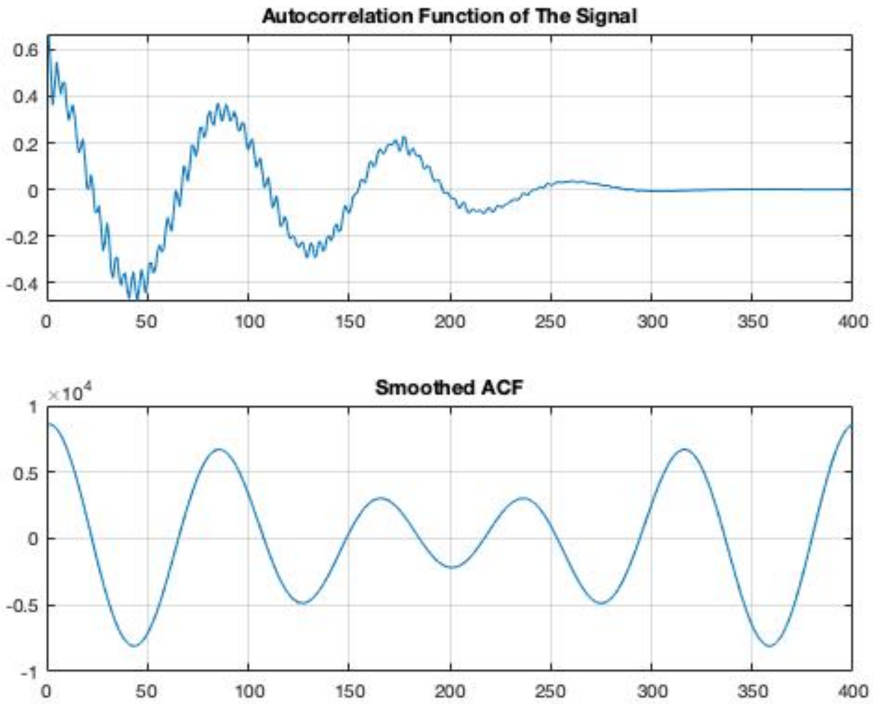
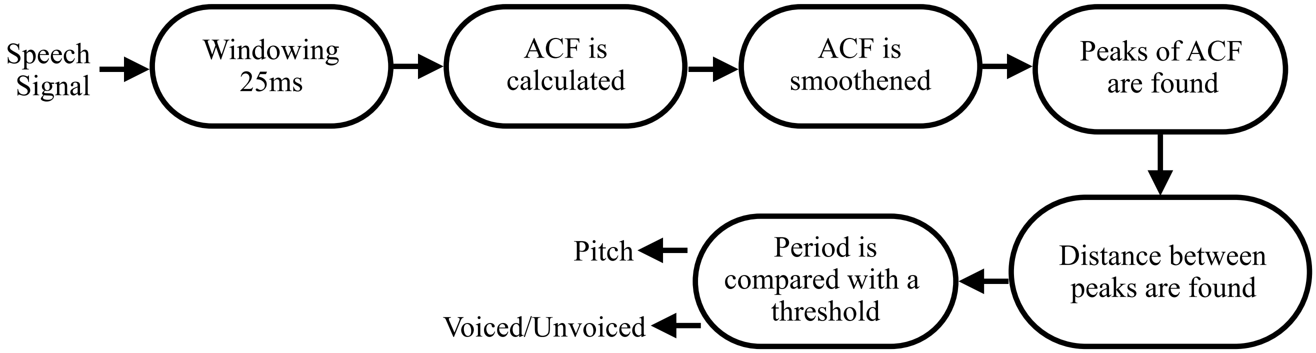


Fig4. 25ms frame of a speech signal and its autocorrelation

In Fig4 one can see that taking the autocorrelation of the autocorrelation function smoothens the signal and therefore can be used to find the peaks and the period.

As stated before after the periods are found the speech signal can be classified as voiced or unvoiced using a threshold on the period. Hence the block diagram of the first pitch detection algorithm can be summarized as follows:



Second Approach to Pitch Detection

Although autocorrelation method gave some good results. Another method is also utilized to compare performances. The method suggested in [8], namely simple inverse filter tracking (SIFT) algorithm is used to detect the pitch of the signal. This process again starts with the windowing of the signal. Then the LP coefficients are used to find the residual. As stated before the residual should resemble an impulse train if the speech signal is voiced. However as stated in [8] this is usually not very reliable. A better approach is to take the autocorrelation of this residual to observe the peak, time of which should give the pitch period.

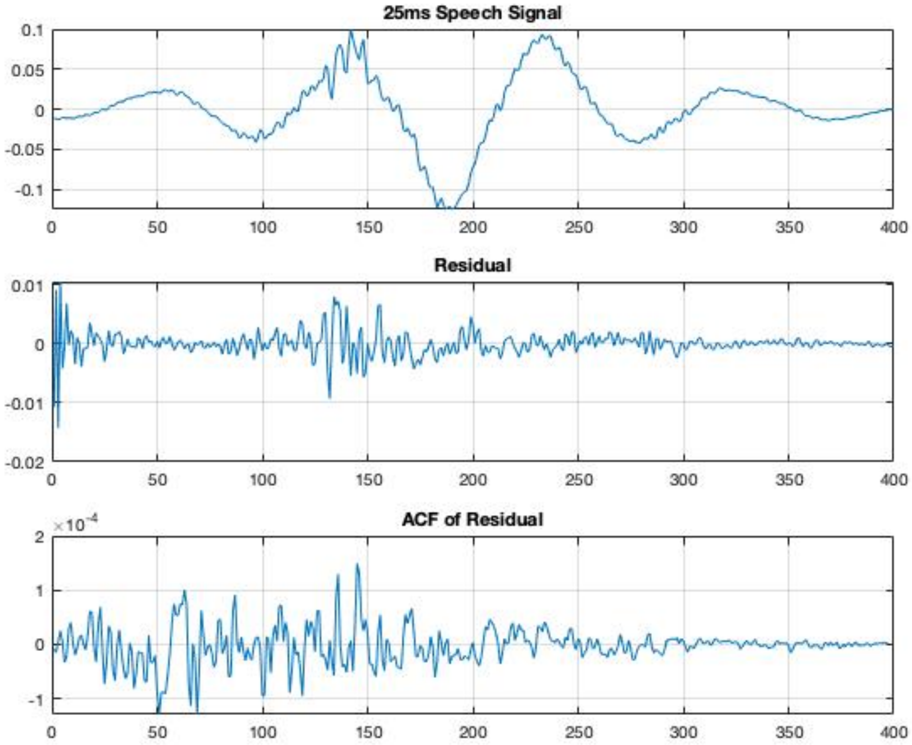
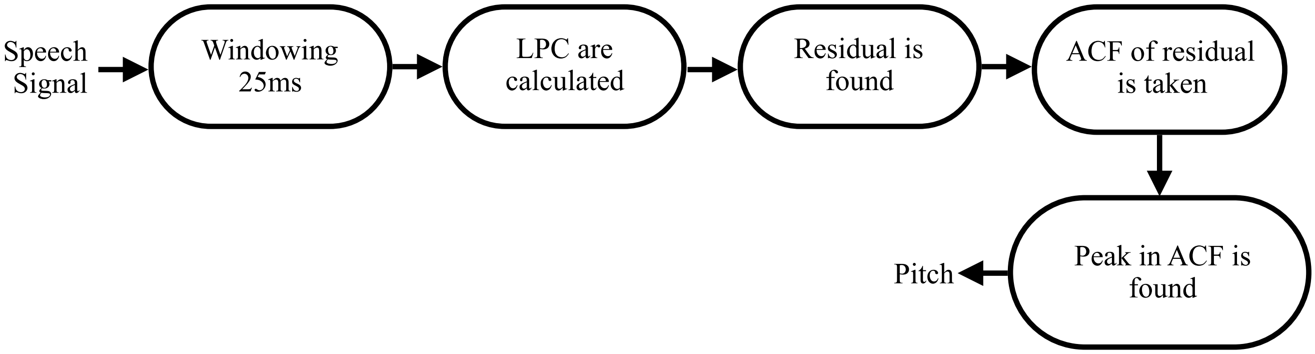


Fig5. Same speech signal in Fig2, its residual and residual's ACF

One can see from Fig5 that the ACF of the residual gives peak at the period of the speech signal, albeit not very large peak. Hence this method can also be used to detect the pitch of the speech signal. The SIFT algorithm in block diagram:



Voiced/Unvoiced

As stated before the pitch period can be used to detect if a speech signal is voiced or not. However as stated in [9] there are other criteria that can be checked too. For example the energy of a voiced speech is usually greater than the energy of an unvoiced speech. Moreover, zero-crossing rate of a signal can also be checked. Zero crossing rate is the at which the zero axes is crossed by a given signal. Therefore it gives an indirect information about its frequency and can be used to decide on whether a speech signal is voiced or not. Voiced speech is produced due to the excitation of vocal tract by the periodic air flow. Hence voiced speech results in a small zero-crossing rate. On the other hand unvoiced speech is produced by the constriction of vocal tract to result in turbulent air flow, which can be resembled as noise and therefore has a high zero crossing rate [10]. The zero-crossing rates and the energies of the frames of a sample speech signal can be observed in the following figure:

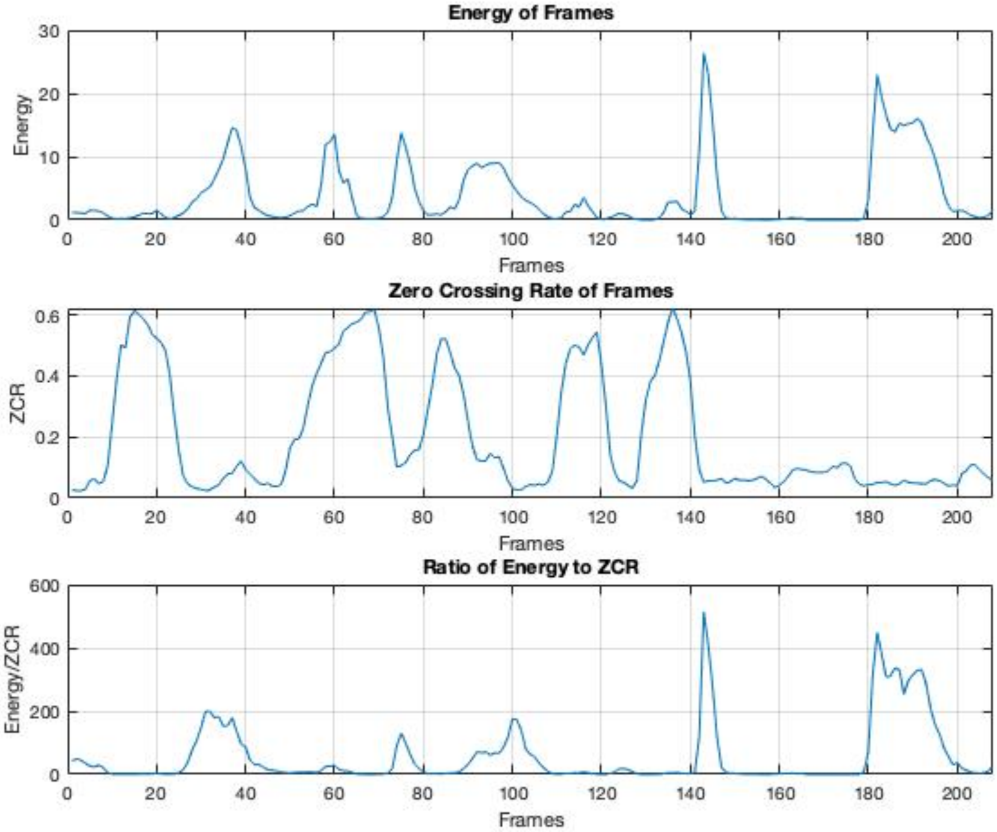


Fig6. Energy and ZCR values of frames of a 2 second speech signal.

One can see from the Fig6 that speech signal consists of frames that vary in energies and zero crossing rates. As stated before the voiced speech has high energy and low ZCR, whereas unvoiced speech has low energy and high ZCR. Therefore their ratios can be used to compare with a threshold to make the decision of voiced/unvoiced.

Gain

Our AR model was $s(n) = \sum_{i=1}^{p} \phi_{p,i} s(n-i) + Gu(n)$. Since u(n) differs depending on whether the signal is voiced or not the gain should also differ depending on the type of the speech signal. Multiplying both sides of the model by s(n) and taking the expected values one can get

$$R_s(0) = \sum_{i=1}^{p} \phi_{p,i} R_s(i) + GR_{us}(0)$$

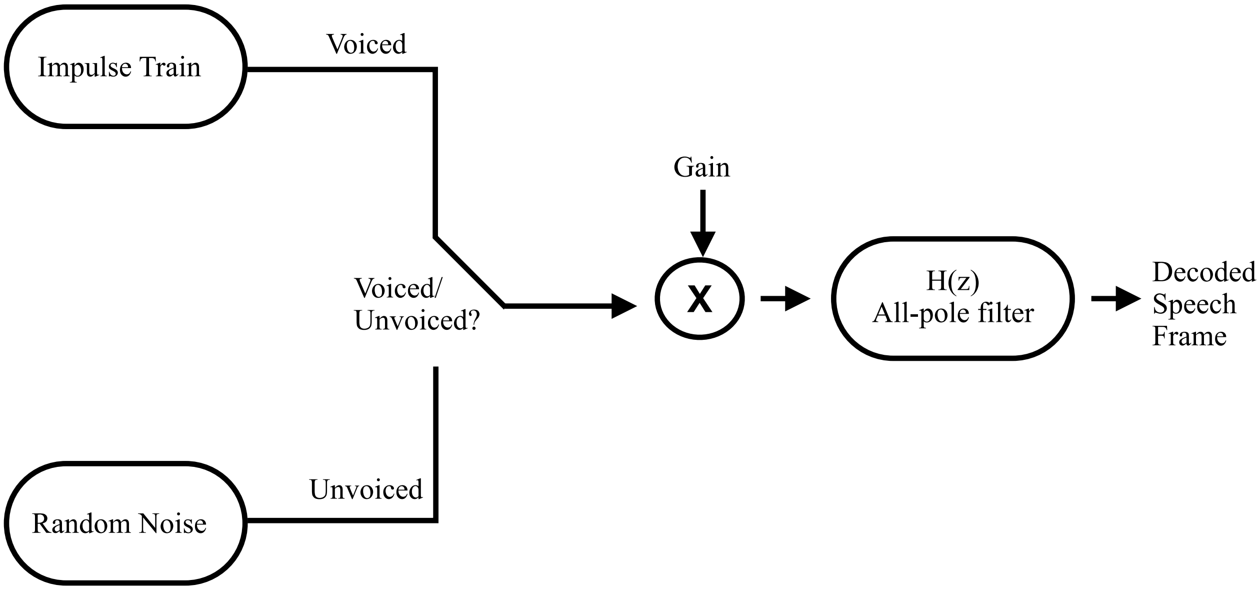
As stated in [8] $R_{us}(n) = G\delta(n)$ for the unvoiced case and $R_{us}(n) = \frac{G}{P}\delta(n)$ for the voiced case where P is the pitch period in samples. Therefore the gain parameters can be stated as:

$$G = \sqrt{R_s(0) - \sum_{i=1}^{p} \phi_{p,i} R_s(i)}$$ for the unvoiced case and

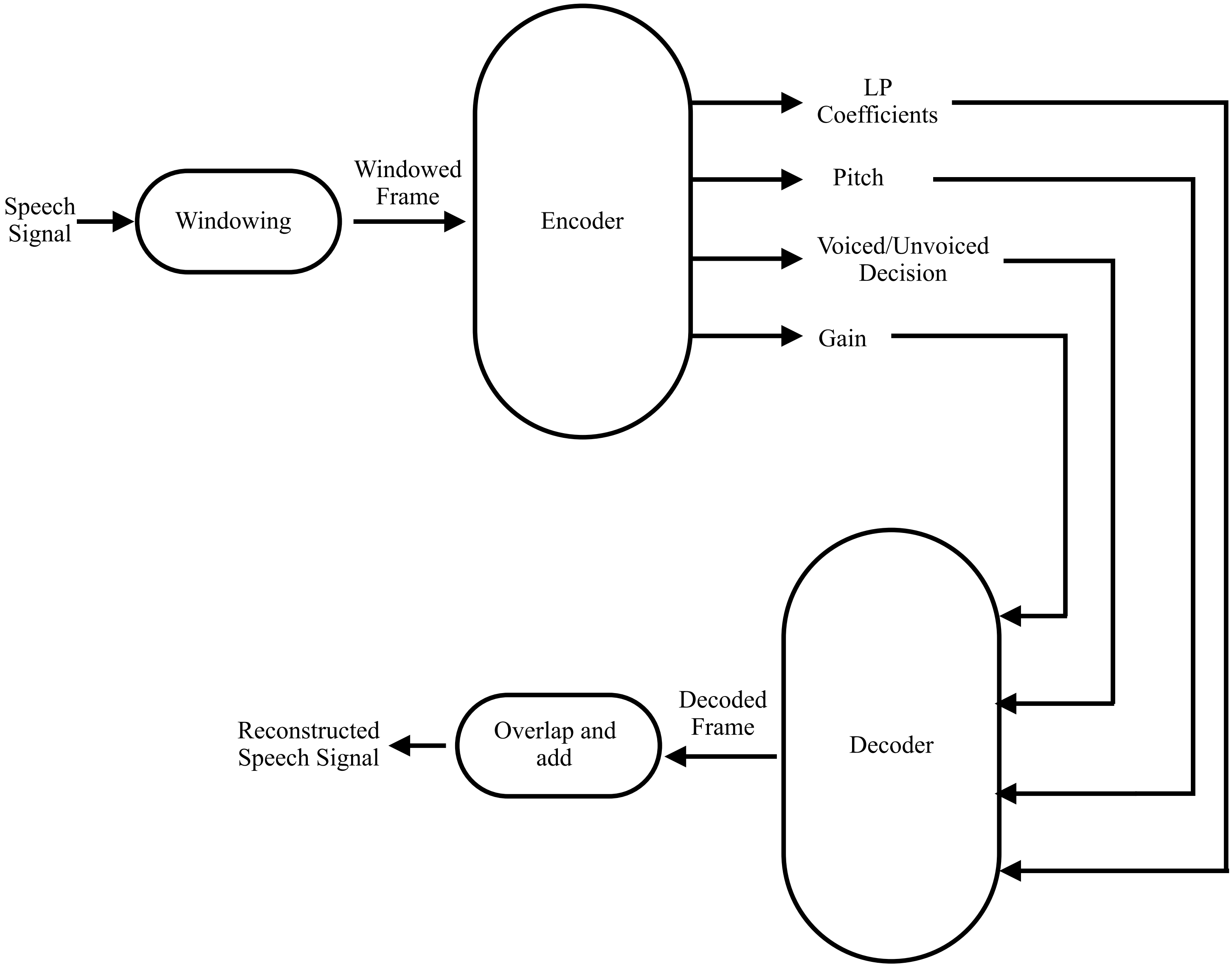
$$G = \sqrt{P \times (R_s(0) - \sum_{i=1}^{p} \phi_{p,i} R_s(i))}$$ for the voiced case.

Decoder

From the encoder 4 types of information is received for each frame. LP coefficients, voiced/unvoiced decisions, pitch periods and the gains. Using these information each frame can be reconstructed. If the frame is voiced speech then impulse train is used to reconstruct. If the frame is unvoiced speech then the random noise is used to reconstruct. Then the input is multiplied by the gain and all-pole filter is used to reconstruct the frame. This can be shown in the following block diagram.



After each frame is reconstructed frames can be overlapped and added in the same fashion as in the decoder. The overall block diagram is as follows:



In MATLAB LP coefficients are calculated by solving Yule-Walker equations, 2 different pitch detection algorithms are used, voiced/unvoiced detection is done by using the ratio of energy to ZCR and the gains are calculated according to the formulas. As the AR order 50 is chosen by empirical observation. 2.1 second speech sample is used as the input and the results are as follows.

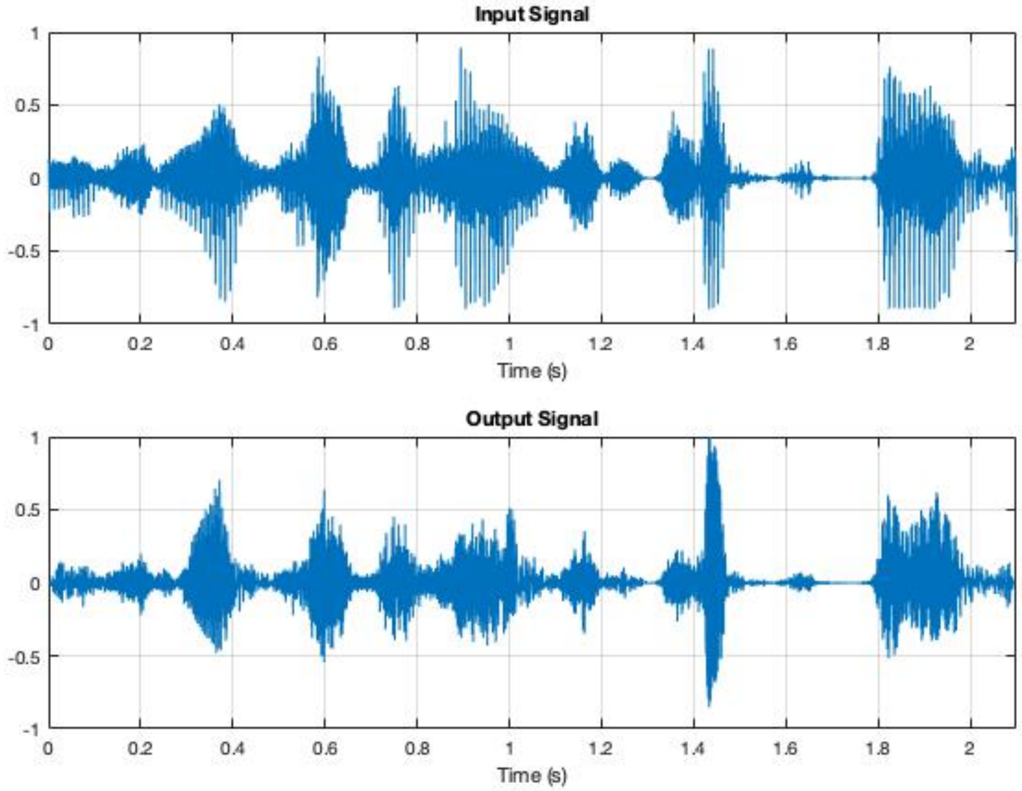


Fig7. Input-Output pair using autocorrelation method to detect pitch

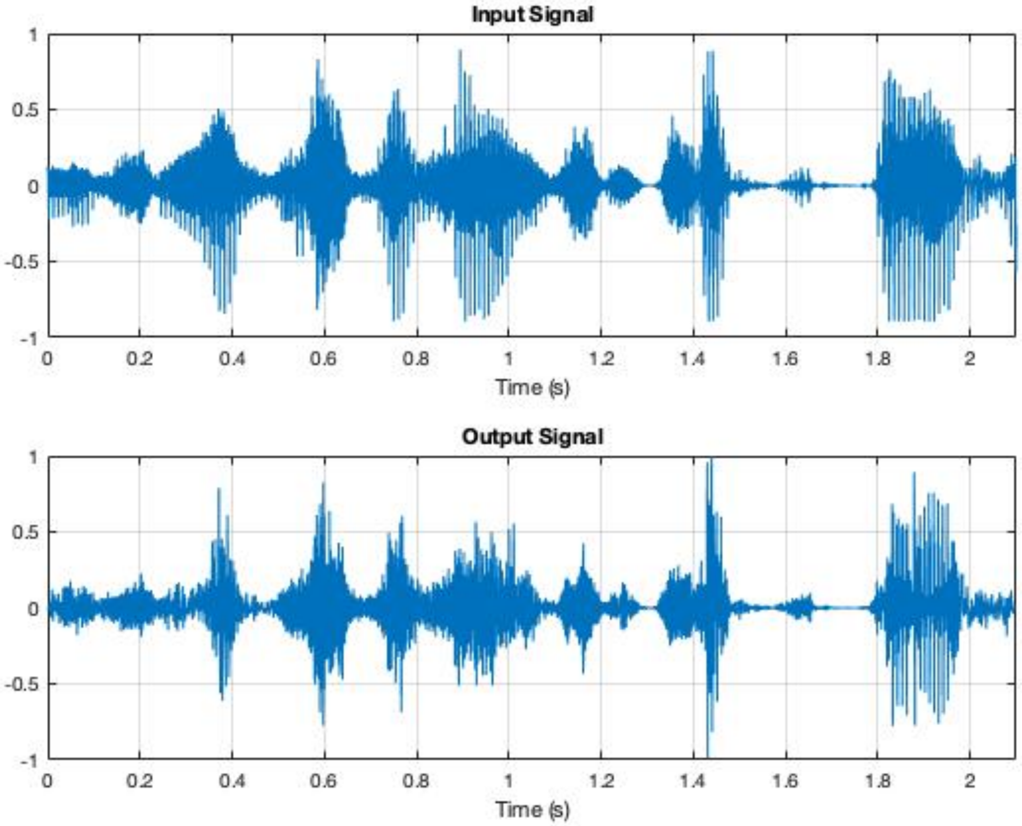


Fig8. Input-Output pair using SIFT algorithm to detect pitch

One can see that both methods work fine but the second method results in less residual energy.

REFERENCES

[1] J. Makhoul, “Linear prediction: A tutorial review,” Proc. IEEE, vol. 63, pp. 561–580, Apr. 1975.

[2] D. O’Shaughnessy, “Linear predictive coding,” IEEE Potentials, vol. 7, pp. 29–32, Feb. 1988.

[3] McLeod, Philip & Wyvill, Geoff. (2005). A smarter way to find pitch.

[4] Benesty, J., Sondhi, M. M., & Huang, Y. (Eds.). (2008). *Springer handbook of speech processing* (Vol. 1). Berlin: Springer.

[5] Markel, J. D., & Gray, A. J. (2013). *Linear prediction of speech* (Vol. 12). Springer Science & Business Media.

[6] Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International journal of speech technology*, *16*(2), 143-160.

[7] S. S. Upadhya, "Pitch detection in time and frequency domain," 2012 International Conference on Communication, Information & Computing Technology (ICCICT), 2012, pp. 1-5, doi: 10.1109/ICCICT.2012.6398150.

[8] Deller, J.R., Proakis, J.G., & Hansen, J.H. (1993). Discrete-Time Processing of Speech Signals.

[9] Nandhini, S., & Shenbagavalli, A. (2014). Voiced/unvoiced detection using short term processing. *International Journal of Computer Applications*, *975*, 8887.

[10] R.G, Bachu & S., Kopparthi & B., Adapa & Barkana, Buket. (2010). Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. 10.1007/978-90-481-3660-5-47.