2012 International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India

Pitch Detection in Time and Frequency Domain

Savitha S Upadhya[#1]

[#]*EXTC Department, Fr. C R I T*
Sector 9A Vashi, Navi Mumbai, Maharashtra State, India
[1]savivashi@gmail.com

Abstract— **The beauty of human speech lies in the complexity of the different sounds that can be produced by a few tubes and muscles. This intricacy, however, makes speech processing a challenging task. One defining characteristic of speech is its pitch. Detecting this Pitch or equivalently, fundamental frequency detection of a speech signal is important in many speech applications. Pitch detectors are used in vocoders, speaker identification and verification systems and also as aids to the handicapped. Because of its importance many solutions to detect pitch has been proposed both in time and frequency domains. One such solution is pitch detection is by using Autocorrelation method and Average Magnitude Difference Function (AMDF), method which are analyses done in the time domain and the other is detecting the harmonic nature in the frequency domain . This paper gives the implementation results of the pitch period estimated in the time and frequency domains for vowel and fricative speech sounds, both for male and female speakers.**

Keywords— **Pitch, Autocorrelation function, Center-clipping, FFT, Voiced speech signals, Unvoiced speech signals.**

I. Introduction

Speech can be classified into two general categories, voiced and unvoiced speech. A voiced sound is one in which the vocal cords of the speaker vibrate as the sound is made, and unvoiced sound is one where the vocal cords do not vibrate. Therefore in voiced sound as the vocal chords vibrate, "Pitch" refers to the percept of the fundamentally frequency of such vibrations or the resulting periodicity in the speech signal. It is a primary acoustic cue to intonation and stress in speech, and is crucial to phoneme identification in tone languages. Most low rate voice coders requires accurate pitch estimation for good reconstructed speech, and some medium rate coders use pitch to reduce transmission rate while preserving high quality speech. Pitch patterns are useful in speaker recognition and synthesis. Real time pitch displays can also give feedback to the deaf learning to speak [1].

Various pitch detection algorithms (PDAs) have been developed in the past: Autocorrelation method [2], AMDF method [2], CPD [3]. Most of them have very high accuracy for voiced pitch estimation, but the error rate considering voicing decision is still quite high. Moreover, the PDAs performance degrades significantly as the signal conditions deteriorate [4]. Pitch detection algorithms can be classified into the following basic categories: time-domain based tracking, frequency domain based tracking or joint time-frequency domain based tracking. This paper discusses both time domain and frequency domain based pitch period detection for vowel and fricative speech sounds, both for male and female speakers.

II. Pitch Detection Algorithms

The pitch can be determined either from periodicity in time or from regularly spaced harmonics in frequency domain. Time domain pitch estimators require - a preprocessor to filter and simplify the signal via data reduction, basic pitch estimator and a post processor to correct errors. It generally locates the quasi periodic time structure of the speech signal or an alternation of high and low amplitudes, or points of discontinuities. Other approaches have speech examined over a short term window. Autocorrelation method and AMDF are two such techniques [1].In frequency domain, the pitch is determined by operating on a block (short-time frame) of speech samples, transforming them spectrally to enhance the periodicity information in the signal. Periodicity appears as peaks in the spectrum at the fundamental and its harmonics[1].

A. Time Domain Pitch Detection Algorithms

I) Autocorrelation Method

The autocorrelation approach is the most widely used time domain method for determining pitch period of a speech signal [3]. This method is based on detecting the highest value of the autocorrelation function in the region of interest. For given discrete signal $x(n)$, the autocorrelation function is generally defined as in (1)

$$R(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m) \qquad 0 \le m \le M_0 \quad (1)$$

where N is the length of analyzed sequence and M_0 is the number of autocorrelation points to be computed. For pitch detection, if we assume that x*(n)* is periodic sequence i.e. *x(n)=x(n+P)* for all *n,* it is shown that the autocorrelation function is also periodic with the same period ,*R(m)=R(m+P).* Conversely, the periodicity in the autocorrelation function indicates periodicity in the signal. For a non-stationary signal, such as speech, the concept of a long-time autocorrelation measurement given by (1) is not really meaningful. In practice, we operate with short speech segments, consisting of finite number of samples. That is why in autocorrelation based PDAs short-time autocorrelation function, given by (2) is used.

$$R(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} x(n)x(n+m) \qquad 0 \le m \le M_0 \quad (2)$$

The variable m in (2) is called lag or delay, and the pitch is equal to the value of m which results in the maximum *R(m)*.

II) Autocorrelation Pitch Detector based on Center Clipping Method

One of the major limitations in the common Autocorrelation method discussed in Section II- A-I is that it retains too much of the information in the speech signals and hence the autocorrelation function has many peaks. Most of these peaks can be attributed to the damped oscillations of the vocal tract response which are responsible for the shape of each period of the speech wave. Therefore in cases when the autocorrelation peaks due to the vocal tract response are bigger than those due to the periodicity of the vocal excitation, the simple procedure of picking the largest peak in the autocorrelation function will fail [2].

To avoid this problem it is again useful to process the speech signal so as to make the periodicity more prominent while suppressing other distracting features of the signal. Here center-clipping technique is used in the pre-processing stage. This removes the effects of the vocal tract transfer function and fewer peaks will appear in the autocorrelation function as compared to many peaks in the common autocorrelation method [2]. This will help us to estimate the pitch more accurately. In this technique, the relation between the input signal $x(n)$, and the center-clipped signal $y(n)$ is given as in (3)

$$y(n) = clc[x(n)] = \begin{cases} (x(n) - C_L), x(n) \ge C_L \\ 0 \qquad\quad , |x(n)| < C_L \\ (x(n) + C_L), x(n) \le C_L \end{cases} \quad (3)$$

where C_L is the clipping threshold.

Generally, C_L is about 30% of the maximum absolute signal value within the signal frame [2]. Non-linear operations on the speech signal such as center-clipping tend to flatten the spectrum of the signal passed to the candidate generator. This results in the increase of the distinctiveness of the true period peaks in the autocorrelation function.

III) AMDF Method :Average Magnitude Difference Function

The average magnitude difference function (AMDF) is another type of autocorrelation analysis. Instead of correlating the input speech at various delays (where multiplications and summations are formed at each value), a difference signal is formed between the delayed speech and original, and at each delay value the absolute magnitude is taken. For the frame of N samples, the short-term difference function AMDF is defined as in (4)

$$D_x(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} |x(n) - x(n+m)| \qquad 0 \le m \le M_0 \quad (4)$$

where *x(n)* are the samples of analyzed speech frame, *x(n+m)* are the samples time shifted on m samples and N is the frame length. The difference function is expected to have a strong local minimum if the lag m is equal to or very close to the fundamental period.

PDA based on average magnitude difference function has advantage in relatively low computational cost and simple implementation. Unlike the autocorrelation function, the AMDF calculations require no multiplications. This is a desirable property for real-time applications. For each value of delay, computation is made over an integrating window of N samples. The average magnitude difference function is computed on speech segment at lags running from 16 to 160 samples. The pitch period is identified as the value of the lag at which the minimum AMDF occurs [2].

In extractors of this type, the limiting factor on accuracy is the inability to completely separate the fine structure from the effects of the spectral envelope. For this reason, decision logic and prior knowledge of voicing are used along with the function itself to help make the pitch decision more reliable [5].

B. Frequency Domain Pitch Detection Algorithm

The short time Fourier transform of the speech signal is computed. The voiced speech signal is manifested in sharp peaks that occur at integer multiples of the fundamental frequency. Another approach is to measure the separation of adjacent harmonics.

III. Implementation Results and Discussion

A. Autocorrelation Method

Fig.1 shows the block diagram of pitch detection using autocorrelation method. At the beginning of processing, the speech signal must be segmented into frames. Speech recordings with 8 kHz sampling frequency were used for experiments. Therefore, input speech signal must be segmented into frames of 240 samples (30ms).

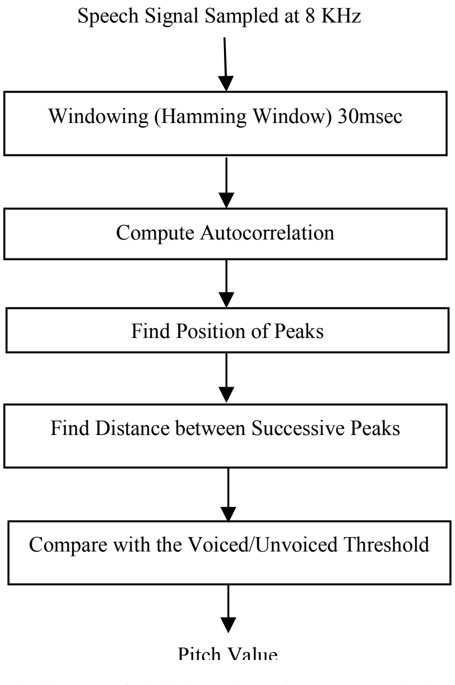
Speech Signal Sampled at 8 KHz
Windowing (Hamming Window) 30msec
Compute Autocorrelation
Find Position of Peaks
Find Distance between Successive Peaks
Compare with the Voiced/Unvoiced Threshold
Pitch Value

Fig. 1 Block diagram of pitch detection using Autocorrelation method

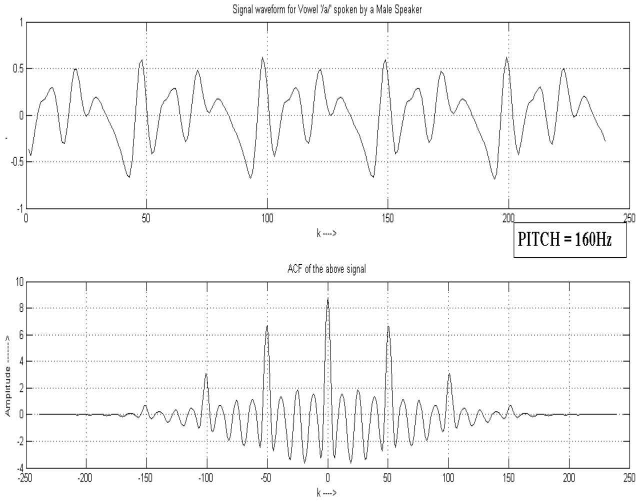

Fig. 2 Autocorrelation of voiced frame

After windowing and computing the autocorrelation over the range of lags, the peaks are searched from the autocorrelation function. The positions (index) of the peaks are obtained. The distance between successive peaks is measured. If these distances are within a threshold value then the frame is classified as voiced. Otherwise, the section is classified as unvoiced. Then the value of fundamental frequency can be computed from the pitch period.

Fig. 2 shows the experimental results of the autocorrelation method obtained for a voiced frame. The first waveform of the figure is the signal waveform and the second is the autocorrelation function of the signal. The periodicity of the signal waveform and also the uniform time lag difference between the peaks of the autocorrelation function explains the fact that the input speech signal is voiced. The figure is the result of the vowel '/a/' spoken by a male speaker. The pitch estimated is 160Hz.

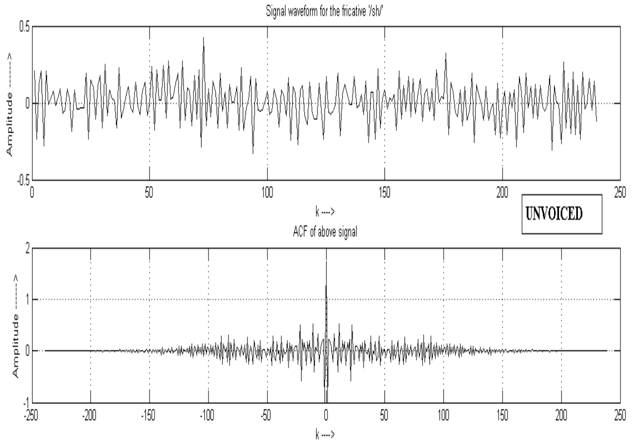

Fig. 3 Autocorrelation of unvoiced frame

Similarly Fig. 3 shows the experimental results of the autocorrelation method obtained for an unvoiced frame. The aperiodicity of the signal waveform and also the non uniform time lag difference between the peaks of the autocorrelation function explains the fact that the input speech signal is unvoiced. The figure is the result of the fricative '/sh/' spoken by a female speaker.

B. Autocorrelation Pitch Detector based on Center-Clipping Method

Fig. 4 shows the block diagram of Autocorrelation pitch detector based on center clipping method. The speech signal sampled at a frequency of 8KHz is windowed. The first stage of processing is the computation of a clipping threshold C_L for 30-ms section of speech. The clipping level is set at a value which is 50 percent of the maximum absolute value of the signal in that windowed segment. Following the determination of the clipping level C_L, the 30-ms section of speech is center

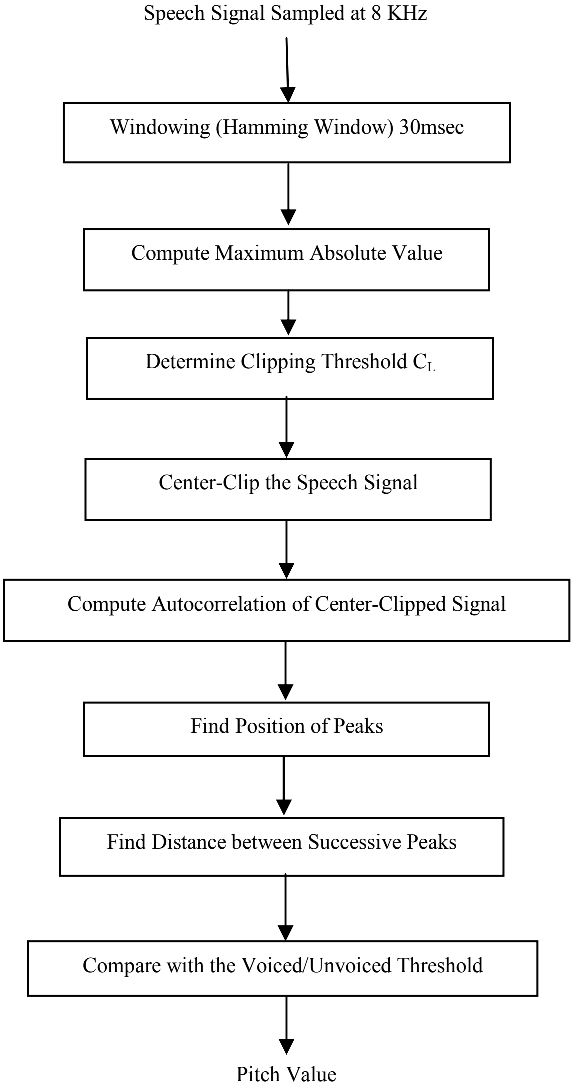
Speech Signal Sampled at 8 KHz
Windowing (Hamming Window) 30msec
Compute Maximum Absolute Value
Determine Clipping Threshold C_L
Center-Clip the Speech Signal
Compute Autocorrelation of Center-Clipped Signal
Find Position of Peaks
Find Distance between Successive Peaks
Compare with the Voiced/Unvoiced Threshold
Pitch Value

Fig. 4 Block diagram of Autocorrelation pitch detector based on center clipping method

clipped as per (3). Following clipping, the autocorrelation function for the 30-ms section is computed over a range of lags and the procedure is repeated as in Section II A.

Fig. 5 shows the experimental results of the Autocorrelation pitch detector based on center clipping method obtained for a voiced frame. The first waveform of the figure is the signal waveform and the second is the center-clipped signal waveform and the third is the autocorrelation function of the center clipped signal. The periodicity of the signal waveform and also the uniform time lag difference between the peaks of the autocorrelation function explains the fact that the input speech signal is voiced. The figure is the result of the vowel '/i/' spoken by a female speaker. The pitch estimated is 266.67Hz.

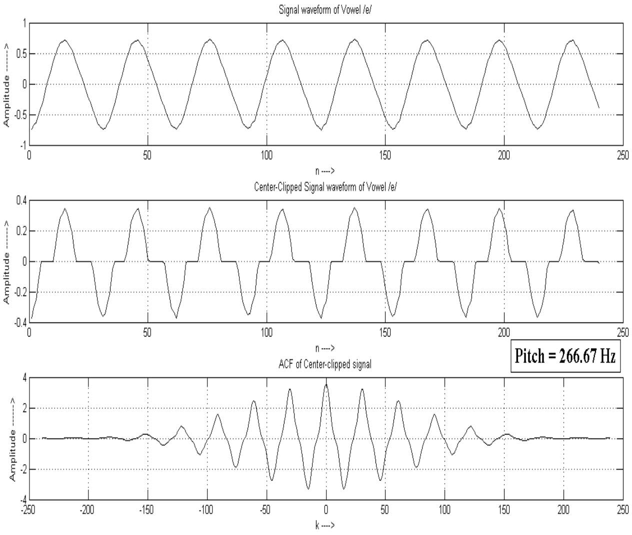

Fig. 5 Autocorrelation of center-clipped voiced frame

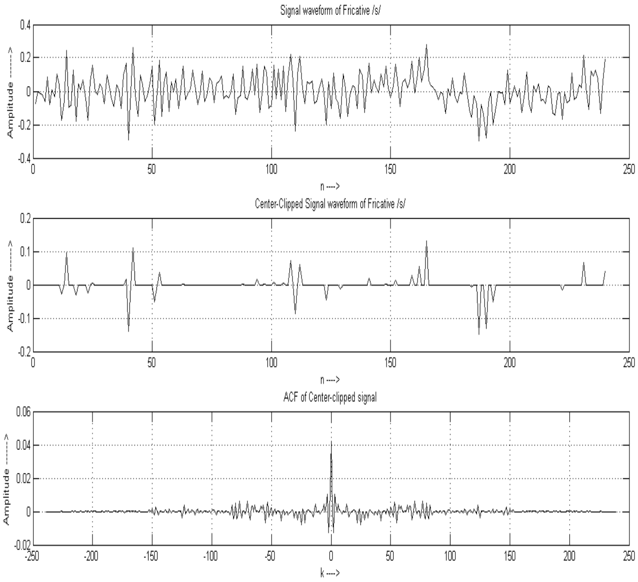

Fig. 6 Autocorrelation of center-clipped unvoiced frame

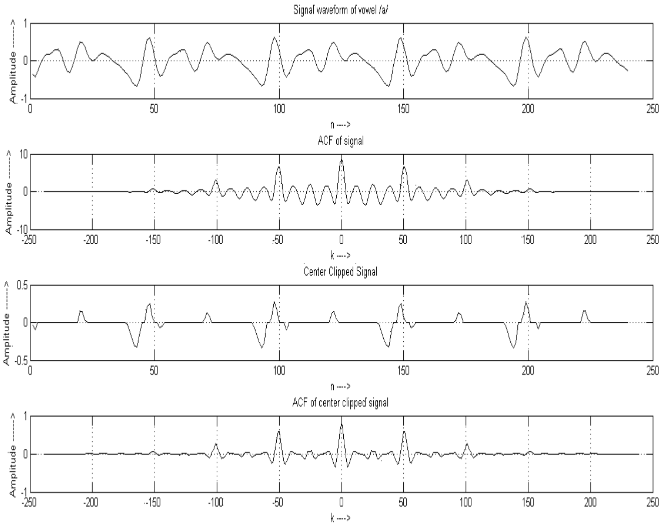

Fig. 7 Comparison of autocorrelation function calculated from voiced frame and its center-clipped version.

Similarly Fig. 6 shows the experimental results of the autocorrelation method obtained for an unvoiced frame. The aperiodicity of the signal waveform and the center-clipped one and also the non uniform time lag difference between the peaks of the autocorrelation function explains the fact that the input speech signal is unvoiced. The figure is the result of the fricative '/s/' spoken by a male speaker.

Fig. 7 presents the experimental results obtained for a voiced frame, its center clipped version, and the difference between the autocorrelation function calculated from original signal frame and center-clipped signal frame. As discussed in Section II-B the autocorrelation function contains many other peaks as compared to the autocorrelation function obtained after center-clipping the signal. The figure is the result of the vowel '/a/' spoken by a male speaker.

C. AMDF Method :Average Magnitude Difference Function

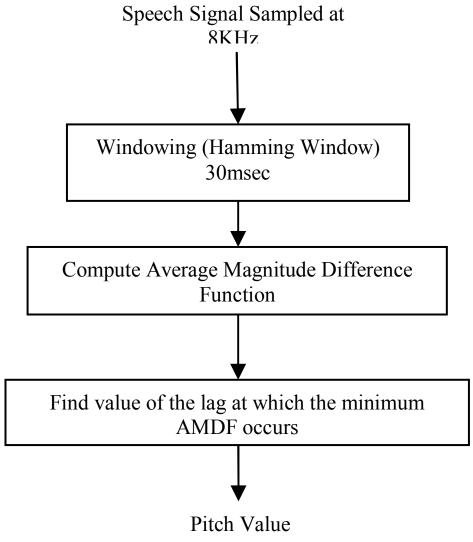
Speech Signal Sampled at 8KHz
Windowing (Hamming Window) 30msec
Compute Average Magnitude Difference Function
Find value of the lag at which the minimum AMDF occurs
Pitch Value

Fig. 8 A Block diagram of AMDF method

Procedure of processing operations for AMDF based pitch detector is quite similar to the Autocorrelation method. After windowing, the average magnitude difference function is computed on the speech segment as defined in (4). The pitch period is identified as the value of the lag at which the minimum AMDF occurs. This is the pitch period. Fig. 8 shows the block diagram of AMDF method.

Fig. 9 shows the experimental results of AMDF method. The first waveform of the figure is the Signal waveform and the second is the AMDF waveform. The minimum lag point is found to be at m=28.The pitch is then equal to 275.86Hz. The figure is the result of the vowel '/u/' spoken by a female speaker. Similarly Fig. 10 shows the AMDF function of an unvoiced frame. The figure is the result of fricative '/sh/' spoken by a female speaker

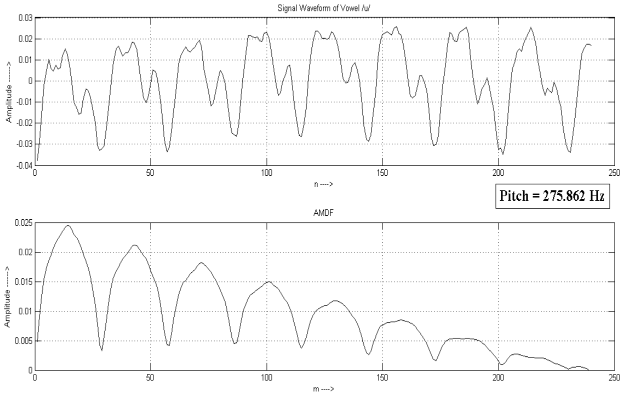

Fig. 9 AMDF Function of voiced frame

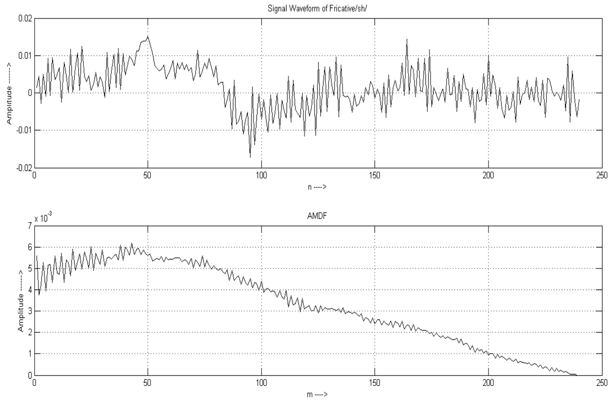

Fig. 10 AMDF function of unvoiced frame

D. Frequency Domain Pitch Detection Algorithm

Here after windowing, the FFT of the speech signal is computed .Then the fundamental frequency is calculated from the spectrum obtained. Fig. 11 shows the spectrum of a voiced frame (vowel '/u/ 'spoken by a female speaker) along with its time domain waveform. The FFT length was taken to be 2048. The pitch is then equal to 277.4Hz. Fig. 12 shows the spectrum of unvoiced frame (fricative '/sh/' spoken by a male speaker).The harmonic nature of the voiced section is observed in Fig. 11.

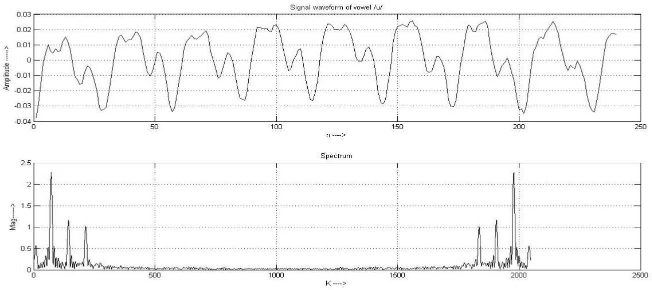

Fig. 11 Spectrum of voiced frame

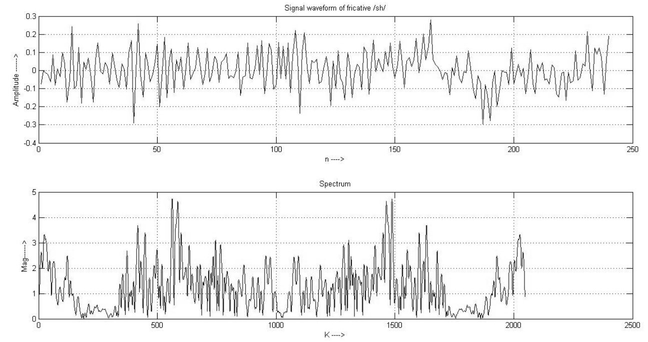

Fig. 12 Spectrum of unvoiced frame

IV. Conclusions

This paper discusses the different pitch detection algorithms for speech signals. PDAs based on the autocorrelation function – Autocorrelation method and Autocorrelation Pitch Detector based on Center-Clipping Method and the Average Magnitude Difference Function method were implemented in time domain. Each of the described algorithms has their advantages and drawbacks. From the experimental results, the Autocorrelation Pitch Detector based on Center-Clipping Method is more convenient for common usage. This algorithm exhibits accurate results of pitch detection and low computational complexity. The AMDF method has great advantage in very low computational complexity. This makes it possible to implement it in real-time applications. However this algorithm showed poor results in accuracy of pitch detection. Short time Fourier transform was implemented in the frequency domain and the fundamental frequency was calculated.

References

[1] Doughlas O'Shaughnessey , *Speech Communication Human And Machine,* 2nd Edn, Universities Press India Limited, India 2001.

[2] L. R.Rabiner and R. W.Schafer ,*Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[3] W. J. Hess, *Pitch Determination of Speech Signals*, New York: Springer, 1993

[4] H. Bořil, P. Pollák, "Direct Time Domain Fundamental Frequency Estimation of Speech in Noisy Conditions". *Proc. EUSIPCO2004*, Wien, Austria, vol. 1, pp. 1003-1006, 2004.

[5] L. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on ASSP,* vol. 24, pp. 399-417, 1976.

[6] B. Kotnik, H. Höge, and Z. Kacic, "Evaluation of Pitch Detection Algorithms in Adverse Conditions". *Proc. 3rd International Conference on Speech Prosody*, Dresden, Germany, pp. 149 -152, 2006