International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072

Pitch Detection Algorithms in Time Domain

Dr. Chavan Madhukar S[1], Sutar Akshay A[2]

[1]*Associate Professor & PG Co-ordinator, Dept. of Electronics and Telecommunication Engineering, P.V.P. Institute of Technology, Budhgaon-416304, Maharashtra, India*
[2]*PG Student, Dept. of Dept. of Electronics and Telecommunication Engineering, P.V.P. Institute of Technology, Budhgaon-416304, Maharashtra, India*

Abstract - *Speech signal can be classified into voiced, unvoiced and silence regions. The near periodic vibration of vocal folds is excitation for the production of voiced speech. Pitch is an important parameter in speech processing. It plays a major role in many speech processing applications. Speech signal is affected by background noise and it degrades the performance. Estimation of pitch for noisy speech signal is important task in many applications. The paper proposes a pitch detection algorithm based on the short-time average magnitude difference function (AMDF) and the short-term autocorrelation function (ACF). 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. This paper gives the implementation results of the pitch period estimated in the time for samples of speech sounds, for different voice samples.*

Key Words: **Pitch, Autocorrelation function, LPF, Center-clipping, Center peak-clipping, Energy of Clipped Signal.**

1. 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.

The pitch estimation plays a very important role in speech compression, speech coding, speech recognition and synthesis, as well as in voice translation. A good estimation of the pitch period is crucial to improving the performance of speech analysis and synthesis systems.

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.

Various pitch detection algorithms (PDAs) have been developed in the past. 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. 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 time domain based pitch period estimation.

2. PITCH DETECTION ALGORITHMS

Speech signal varies with time and so for extracting features properly, we need to identify a smaller portion or window segment of speech. Then the recorded speech signal is downsampled and read using 'wavread' function in MATLAB. This function reads a small window of the downsampled signal and returns the sampled values as well as the sampling frequency. For pitch detection, time domain analyses are done in this work. The pitch can be determined either from periodicity in time domain. Time domain pitch analysis includes the Autocorrelation method and AMDF techniques

3. TIME DOMAIN PITCH DETECTION ALGORITHM

3.1 Autocorrelation Method

The autocorrelation approach is the most widely used time domain method for determining pitch period of a speech signal. 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)

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

Where, N is the length of analyzed sequence and n is index in frame 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.

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

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.

3.2 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

$$R(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} x(n)x(n+m) \qquad 0 \le m \le M_0 \tag{3}$$

Where x (n) are samples of analyzed speech frame, x (n+m) are samples and N is 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 frequency.

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.

4. IMPLEMENTATION RESULTS AND DISCUSSION

4.1 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).

- **FRAMING AND WINDOWING FOR ACF**

Perform frame blocking such that a stream of audio signals is converted to a set of frames. The time duration of each frame is about 20~30 ms if the frame duration is too big, the time-varying characteristics of the audio signal cannot extracted. On the other hand, if the frame duration is too small, cannot extract valid acoustic features. In general, a frame should be contains several fundamental periods of the given audio signals. Usually the frame size in terms of sample points is equal to the powers of 2 such as 256, 512, 1024, etc. such that it is suitable for fast Fourier transform.

To reduce the difference between neighboring frames, overlap between them is done. Usually the overlap is 1/2 to 2/3 of the original frame. The more overlap, the more computation is needed.

There are several terminologies that are used often:

- Frame size: The sampling points within each frame
- Frame overlap: The sampling points of the overlap between consecutive frames.
- Frame step (or hop size): This is equal to the frame size minus the overlap.
- Frame rate: The number of frames per second, which is equal to the sample frequency divided by the frame step

For instance, if a stream of audio signals with sample frequency fs=16000, and a frame duration of 25 ms, overlap of 15 ms, then

- Frame size = fs*25/1000 = 400 (sample points)
- Frame overlap = fs*15/1000 = 240 (sample points)
- Frame step (or hop size) = 400-240 = 160 (sample points)
- Frame rate = fs/160 = 100 frames/sec

fs = 16,000 samples/second
Frame rate (overlap percentage) = 10 ms
Window length (Frame length) = 25 ms
=> (25 ms * 16,000 = 4000 sample/frame)

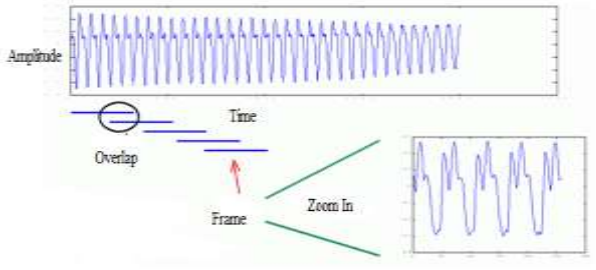

Fig 1. Process of frame blocking for ACF method

Speech is a time varying signal, and some variations are random. Usually during slow speech, the vocal tract shape and excitation type do not change in 200 ms. But phonemes have an average duration of 80 ms. most changes occur more frequently than the 200 ms time interval. Signal analysis assumes that the properties of a signal usually change relatively slowly with time. This allows for short-term analysis of a signal. The signal is divided into successive segments, analysis is done on these segments, and some dynamic parameters are extracted. The signal s(n) is multiplied by a fixed length analysis window w(n) to extract a particular segment at a time. This is called windowing. Choosing the right shape of window is very important, because it allows different samples to be weighted differently. The simplest analysis window is a rectangular window of length Nw:

$$w(n) = \begin{cases} 1, & 0 \le n \le N_w - 1, \\ 0 & otherwise \end{cases} \quad (4)$$

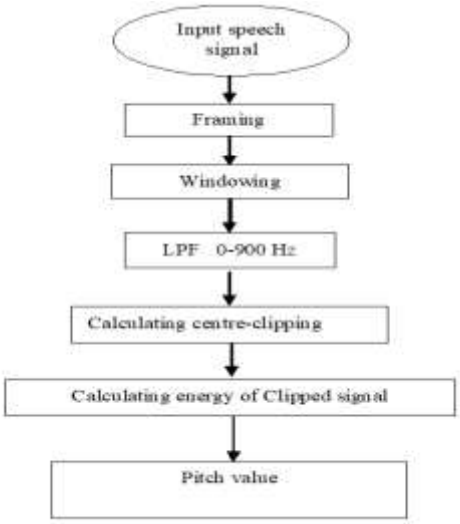

Fig. 2 Flowchart of pitch detection using ACF method

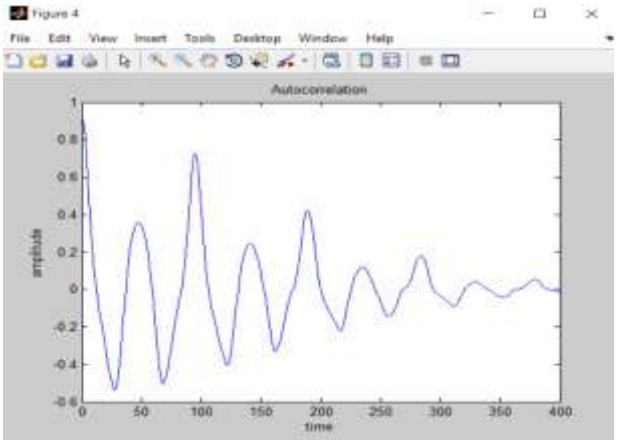

Fig. 3 Autocorrelation of voiced frame

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. 3 shows the experimental results of the autocorrelation method obtained for a voiced frame. The waveform in the figure 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.

4.2 Average Magnitude Difference Function Method

Procedure of processing operations for AMDF based pitch detector is quite similar to the Autocorrelation method.

- **FRAMING AND WINDOWING FOR AMDF**

Perform frame blocking such that a stream of audio signals is converted to a set of frames. The time duration of each frame is about 20~30 ms if the frame duration is too big, the time-varying characteristics of the audio signal cannot extracted. On the other hand, if the frame duration is too small, cannot extract valid acoustic features. In general, a frame should be contains several fundamental periods of the given audio signals. Usually the frame size in terms of sample points is equal to the powers of 2 such as 256, 512, 1024, etc. such that it is suitable for fast Fourier transform.

To reduce the difference between neighboring frames, overlap between them is done. Usually the overlap is 1/2 to

2/3 of the original frame. The more overlap, the more computation is needed.

There are several terminologies that are used often:

- Frame size: The sampling points within each frame
- Frame overlap: The sampling points of the overlap between consecutive frames
- Frame step (or hop size): This is equal to the frame size minus the overlap.
- Frame rate: The number of frames per second, which is equal to the sample frequency divided by the frame step

For instance, if a stream of audio signals with sample frequency fs=16000, and a frame duration of 25 ms, overlap of 15 ms, then

- Frame size = fs*25/1000 = 400 (sample points)
- Frame overlap = fs*15/1000 = 240 (sample points)
- Frame step (or hop size) = 400-240 = 160 (sample points)
- Frame rate = fs/160 = 100 frames/sec

fs = 16,000 samples/second

Frame rate (overlap percentage) = 10 ms

Window length (Frame length) = 25 ms

=> (25 ms * 16,000 = 4000 sample/frame)

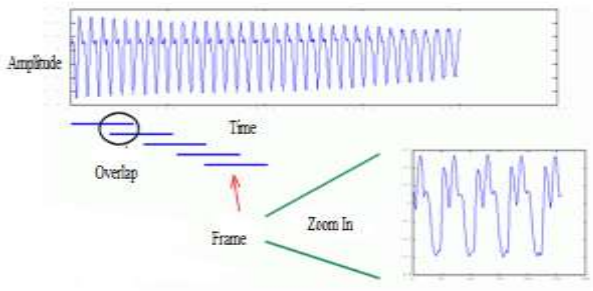

Fig 4 Process of frame blocking for AMDF method

Speech is a time varying signal, and some variations are random. Usually during slow speech, the vocal tract shape and excitation type do not change in 200 ms. But phonemes have an average duration of 80 ms. most changes occur more frequently than the 200 ms time interval. Signal analysis assumes that the properties of a signal usually change relatively slowly with time. This allows for short-term analysis of a signal. The signal is divided into successive segments, analysis is done on these segments, and some dynamic parameters are extracted. The signal s(n) is multiplied by a fixed length analysis window w(n) to extract a particular segment at a time. This is called windowing. Choosing the right shape of window is very important, because it allows different samples to be weighted differently. The simplest analysis window is a rectangular window of length Nw:

$$w(n) = \begin{cases} 1, & 0 \le n \le N_w - 1, \\ 0 & otherwise \end{cases} \quad (6)$$

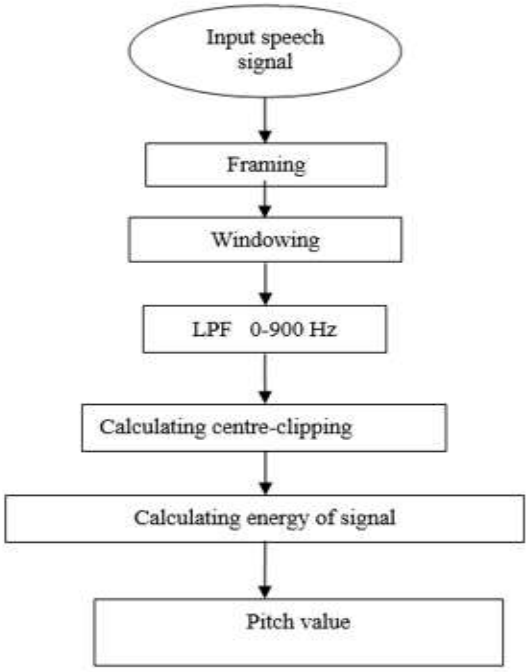

Fig. 5 Flowchart of pitch detection using AMDF method

After windowing, the average magnitude difference function is computed on the speech segment as defined in equation 3. The pitch period is identified as the value of the lag at which the minimum AMDF occurs. This is the pitch period. Fig. 5 shows the flowchart of AMDF method.

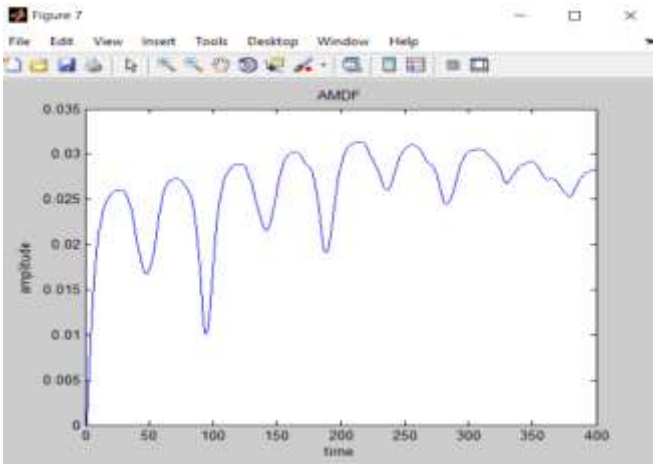

Fig 6. AMDF Function of voiced frame

Fig. 6 shows the experimental results of the AMDF method obtained for a voiced frame. The waveform in the figure is the AMDF function of an signal. 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.

5. CONCLUSIONS

This paper discusses the different pitch detection algorithms for speech signals. PDAs based on the autocorrelation function – Autocorrelation method and Autocorrelation. Each of the described algorithms has their advantages and drawbacks. The AMDF method has great advantage in very low computational complexity. This makes it possible to implement it in real-time applications.

REFERENCES

[1] Chandrashekar H. M1 and Pratibha K2, "Estimation and Tracking of Pitch for Noisy Speech Signals using EMD based Autocorrelation Function Algorithm" 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India

[2] K. Inbanila1 and E. Krishnakumar2 ''Enhancement of Substitution Voices Using F1 Formant Deviation Analysis and DTW Based Template Matching" International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) 2017.

[3] 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.

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

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

[6] 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.

[7] 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

ACKNOWLEDGEMENT

Gratitude is the hardest emotion to express and often one doesn't find adequate words to convey that entire one feels. I would like to express the deepest appreciation to my Guide and PG Coordinator Dr. Madhukar S. Chavan who has the attitude and the substance of a genius that he continually and convincingly conveyed a spirit of adventure in regard to research. Without his guidance and persistent help this dissertation would not have been possible.

I would like to thank my Head of department Dr. D. B. Kadam for his valuable suggestions and constant encouragement all through the project work.

I offer my humble and sincere thanks to Dr. D. V. Ghewade Principal, P.V.P.I.T. Budhgaon for his all possible cooperation. I express my sincere thanks to all faculty and staff members of Electronics and Telecommunication Engg. Department of P.V.P.I.T. Budhgaon for their kind co-operation and encouragement.

It gives me immense pleasure to acknowledge and thank many individuals who contributed in various ways for the successful completion of this work. This project consumed huge amount of work, research and dedication. I would like to extend my sincere gratitude to all of them.

BIOGRAPHIES

"Dr. Chavan Madhukar S.
Associate Professor & PG Co-ordinator.
Department of Electronics and Telecommunication Engineering,
P.V.P. Institute of Technology,
Budhgaon."

"Mr. Sutar Akshay A.
PG Student "Description
Department of Electronics and Telecommunication Engineering,
P.V.P. Institute of Technology,
Budhgaon"