International Journal of Engineering & Technology, 7 (3.29) (2018) 580-585

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Evaluation of pitch estimation in clean and noisy speech

A. Satyanarayana Murthy [1] *, P. Sairam [1], B. Sai Kumar [1]

[1] *B. V. Raju Institute of Technology, Narsapur, Medak, TS*
**Corresponding author E-mail: satyanarayanamurthy.a@bvrit.ac.in*

Abstract

Every human being has a distinct voice due to pitch association and it is almost like a finger print. Pitch is one of the important parameter which is used in many speech processing applications. In reality speech is a complex combination of both voiced and unvoiced sounds and cannot be separated subjectively. For the voiced speech, pitch is defined as the rate of change of vocal folds vibrations. In practice, pitch is a subjective quantity and cannot be measured directly from the voice. It is a non-linear quantity, depends upon the spectral and temporal content of the signal. Many pitch estimation methods have been developed but none can work efficiently in the presence of additive noise. It is very essential to understand the effect of noise on the pitch estimation in dealing effectively with many speech processing applications. Speech processing systems should be robust enough to counter the presence of noise to produce good quality sounds. In non-intrusive speech quality measurement algorithms, pitch is one of the quality parameter for speech assessment. The accuracy of this feature in noisy speech is correlated with the subjective quality of speech. In this paper we have been evaluated the performance of auto-correlation and cepstrum algorithms for pitch estimation and tracking.

Keywords: *Auto-Correlation; Cestrum; Pitch; Speech Processing.*

1. Introduction

Voice generation is like a sound from a musical instrument. Voice is used for many applications i.e. identify people, verbal communications, singing, emotions, allows communication via computers etc. Every human being has a unique voice and it is almost like a finger print. Every voice is a complex combination of many different characteristics like pitch, formants, tone, speaking rate etc. Voice is generated due to complex movements of muscles dictated by the brain. Pitch is the most important critical component in the voice characteristics. The pitch associated with the voice is defined as the rate of vocal folds vibrations. The voice is dependent on the pitch value. Higher the pitch value results more sound effect compared to lower pitch values. The pitch is dependent on the geometry of the vocal folds and on the strength of the muscles around them. The females have more pitch values because of shorter vocal folds compared to males. Interestingly our pitch is also affected due to other factors like emotions, moods, and inflection etc. For example if the person gets frightened, the muscles around the voice box getting tense and produce higher pitch unconsciously. So there is relation between feelings and pitch values. So people try to control the vocal folds to produce different sounds.

Pitch [1], [2] is perceived by human beings with the help of ears and brain and is continuously varying quantity. Speech is a non-stationary signal and has a bandwidth of up-to 8000 Hz approximately. Hearing loss of a person means unable to hear some of the frequencies of the speech; it is usually comes in the old age. The harmonic content produced by the vocal folds can be changed by the shape of the vocal folds, which is controlled by many factors. The harmonics considered to be the source of speech sounds. The higher the harmonic frequency if the vocal folds get closer. The harmonic content inside the vocal tract resonates due to the continuous movement of tongue, jaws etc. The resonant frequencies are called as Formants. The modeling of Formants plays an important role in speech processing applications such as coding, recognition, synthesis and enhancement etc. They contain higher energies and they are slowly varying in time. They are difficult to estimate rather than to find the peaks in the spectrum. They are estimated from Short-Time Fourier Transform (STFT) or from Linear Predictive Coding (LPC) analysis [3], [4].

The sounds i.e. speech, music or noise, are classified based on the variations of pitch and intensity. Vowels are the low frequency sounds which add richness, sexiness and identity to the speech where as the consonant sounds are the higher frequencies which add clarity to the speech. In this paper the pitch of male and female has been estimated using auto-correlation and cepstrum methods. The effect of pitch in the presence of 0dB train noise is estimated.

2. Short-term autocorrelation

The auto-correlation and cross-correlation tools are used to determine the similarity between the two sequences. The cross-correlation is performed on two different sequences where as the auto-relation is performed on the same sequence. Auto-correlation computation is performed at different time lags and it tells the similarity of the sequence with respect to the reference.

The auto-correlation is widely used in speech processing applications; however because of non-stationary nature of speech, a short-time auto-correlation [5] is used. The auto-correlation of a sequence, s(n) is defined by Eqn. (2.1), whereas the short-time non-stationary auto-correlation sequence given by Eqn. (2.2)

$$r_{ss}(k) = \sum_{m=-\infty}^{\infty} s_w(m).s_w(k+m) \quad (2.1)$$

$$r_{ss}(n,k) = \sum_{m=-\infty}^{\infty}(s(m)w(n-m).s(k+m).w(n-k+m))) \quad (2.2)$$

Copyright © 2018 A. Satyanarayana Murthy et. al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Where s_w (n)=s(m)w(n-m) is the window segment of speech s(n). The short-time auto-correlation data is different for voiced and unvoiced speech and is used to differentiate the voiced and unvoiced. A minimum frame size of two cycles of voiced speech signal is needed to auto-correlation computation. The frame size should be in the range of 10-50 msec. In the case of voiced speech the auto-correlation data shows periodic nature where as in unvoiced speech no such periodicity can be observed. So from the auto-correlation plot we can see the nature of the voiced and unvoiced behavior. The pitch period is observed from the plot from where we find a strong peak.

3. Cepstrum method

Speech is the result of excitation and vocal tract system components. In many speech processing applications the models of excitation and vocal tract components are necessary. The cepstrum tool [6]-[8], is helpful to achieve the separation of these two components independently without the prior knowledge of either the source or system. Mathematically, speech can be considered as the convolution of excitation and vocal tract components. If the excitation and vocal tract sequence is represented by e(n) and h(n) respectively then the speech sequence is denoted as s(n). The time and frequency domain representations are shown in Eqn.(3.1) and (3.2).

$$S(n) = e(n) * h(n) \tag{3.1}$$

$$S(\omega) = E(\omega).H(\omega) \tag{3.2}$$

Eqn. (3.2) shows the direct multiplication of excitation and system components. It is possible to convert the two components in the frequency domain into a linear combination of two time domain components via the cepstral analysis. So the cepstral analysis transforms the source and systems components into cepstral domain (time-domain) as follows:
Take the magnitude spectrum and apply the logarithm on both sides to linearly separate the two components. The magnitude and logarithm representations are shown in Eqn.(3.3) and (3.4).

$$|S(\omega)| = |E(\omega)|.|H(\omega)| \tag{3.3}$$

$$log|S(\omega)| = log|E(\omega)| + log|H(\omega)| \tag{3.4}$$

The log transforms the speech spectrum into linear combination (summation) of excitation and vocal tract components. The inverse discrete Fourier transform (IDFT) of Eqn. (3.4) results the excitation and vocal tract information in cepstral domain or quefrency domain, similar to the time domain. This is shown in Eqn. (3.5). The vocal tract represents the lower quefrency and the excitation represents the higher quefrency region component.

$$c(n) = IDFT(log|S(\omega)|) = IDFT(log|E(\omega)| + log|H(\omega)|) \tag{3.5}$$

The block diagram in Fig. 3.1 represents the various steps to be followed to obtain the cepstrum of the given speech segment.

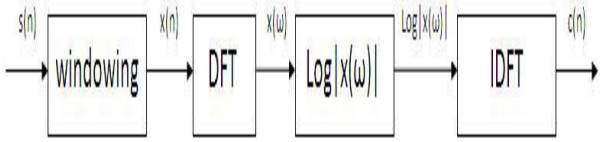


Fig. 3.1: Block Diagram Representing Computation of Cepstrum.

Methods have to devise to separate the vocal tract and excitation components in quefrency region independently. This is achieved by using the liftering operation which is similar to the filtering operation. The required frequency portion will be selected by liftering similar to filtering operation. The vocal tract information is selected by low-time liftering where as high-time liftering is performed to get the excitation information.

4. Experimental conditions

We performed experiments on speech signals which are free from noise. We took the Male and Female utterances "A good book informs what you want to know" and "She has a smart way of wearing clothes" using Matlab programming. The sampling frequency selected as 8 kHz. The speech data took from the data base NOIZEUS [9]. We selected different voice segments to determine the pitch for male and female signals.

a) Pitch estimation and tracking via auto-correlation

The auto-correlation is performed by selecting the samples from the male voice: sp25.wav file. The selected samples as follows:

i) 3600 to 4600
ii) 7000 to 7200
iii) 12500 to 13500 and the corresponding results are shown in Fig. 4.1- 4.3. The observed pitch frequencies are as follows: 186 Hz, 111Hzand 103 Hz.

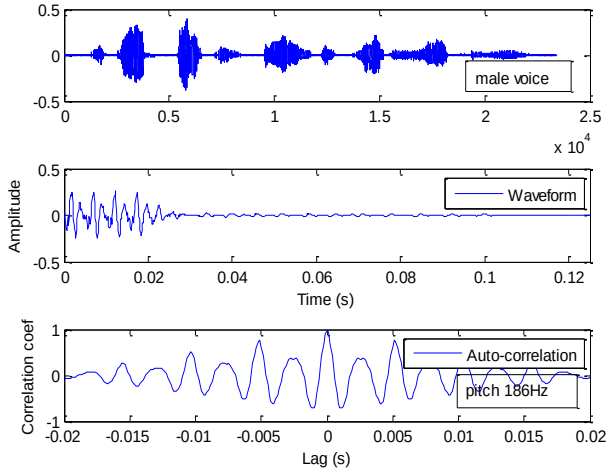


Fig. 4.1: (I): Male Speech, Voiced Segment and Auto-Correlation (Top to Bottom).

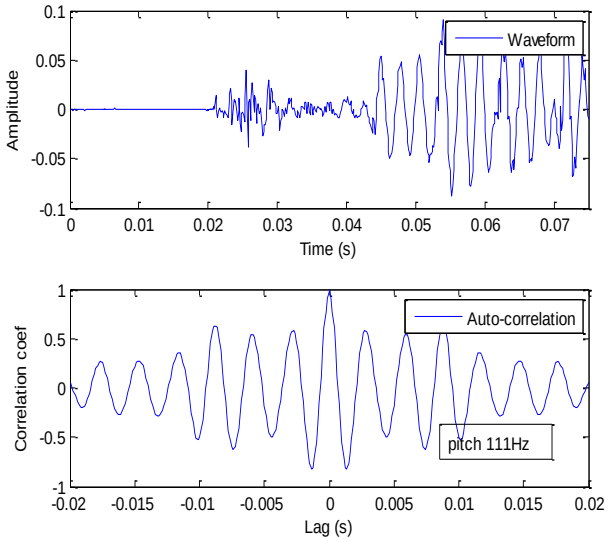


Fig. 4.2: (II): Male Voiced Segment and Auto-Correlation (Top to Bottom).

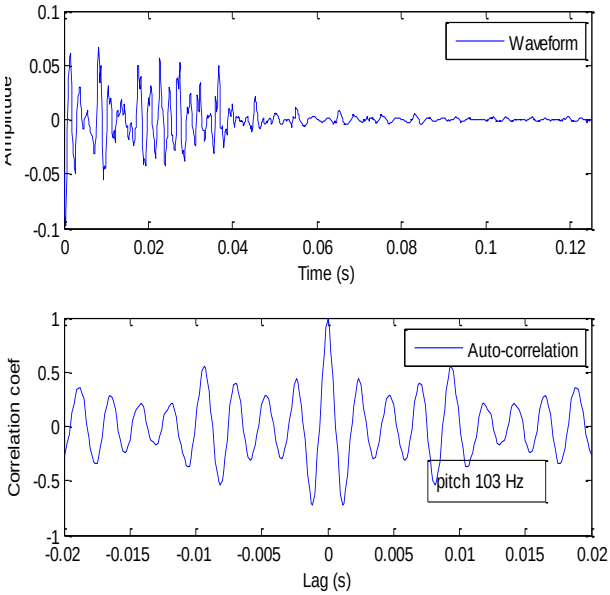


Fig. 4.3: (III): Male Voiced Segment and Auto-Correlation (Top to Bottom).

For the female utterance: sp26.wav, the experiments are conducted for the voiced segments (i) 4800 to 5600 (ii) 9600 to 10200 (iii) 12500 to 13500 and the corresponding results are shown in Figs. 4.4-4.6. The observed pitch frequencies are 421Hz, 210 Hz, and 200 Hz.

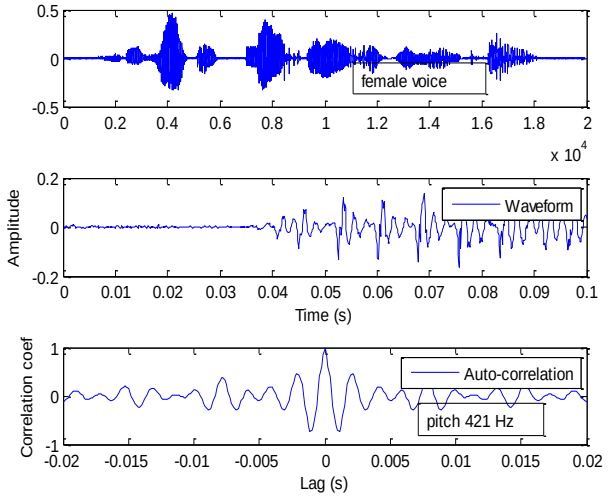


Fig. 4.4: (I): Female Speech, Segment and Auto-Correlation (Top to Bottom).

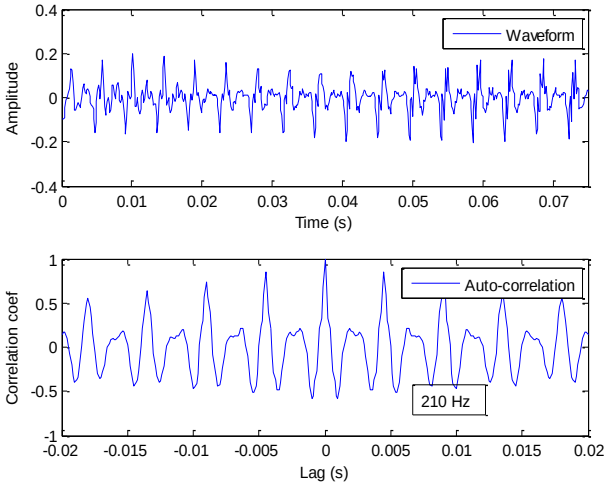


Fig. 4.5: (II): Female Voice Segment and Auto-Correlation (Top to Bottom).

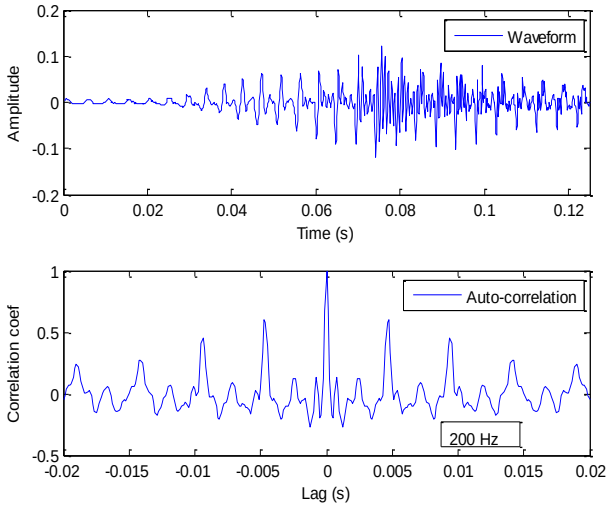


Fig. 4.6: (III): Female Voice and Auto-Correlation (Top to Bottom).

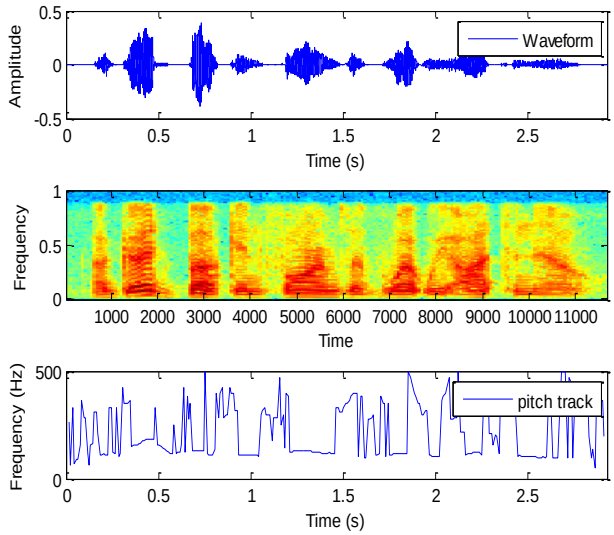


Fig. 4.7: Male Speech, Spectrogram and Pitch Tract (Top to Bottom).

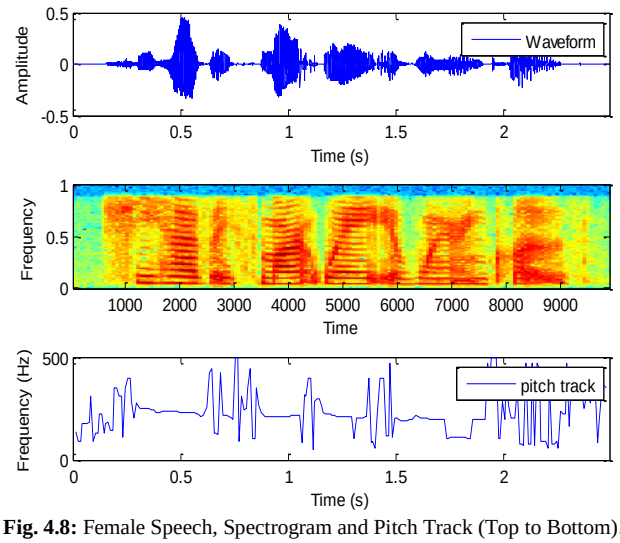


Fig. 4.8: Female Speech, Spectrogram and Pitch Track (Top to Bottom).

b) Pitch estimation and tracking via cepstrum

The samples from the segments of speech file (male): sp25.wav file are as follows: (i) 3600 to 4600 (ii) 7000 to 7200 (iii) 12500 to 13500. The results are shown in Fig. 4.9-4.11. The observed pitch frequencies are as follows: 186 Hz, 500Hz, and 98Hz.

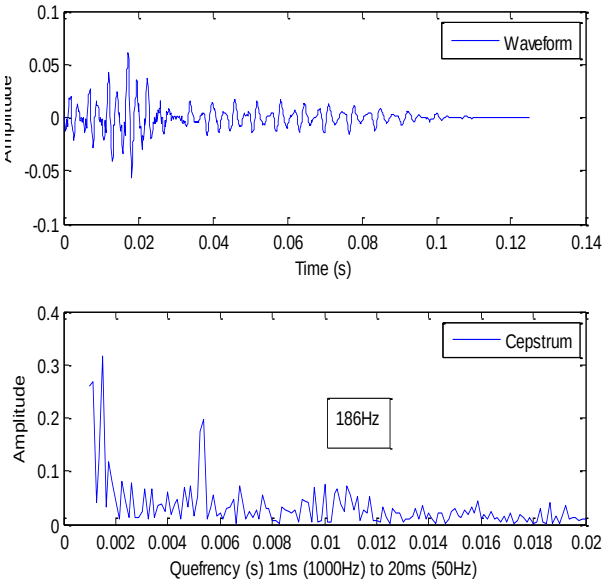


Fig. 4.9: (I): Male Voice Segment and Cepstrum (Top to Bottom).

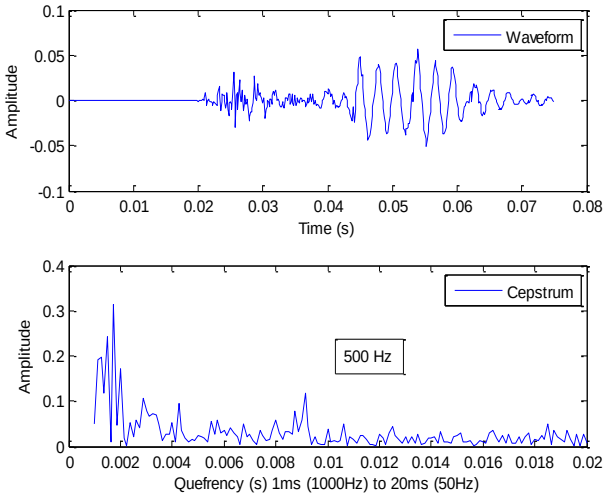


Fig. 4.10: (II): Male Voice Segment and Cepstrum (Top to Bottom).

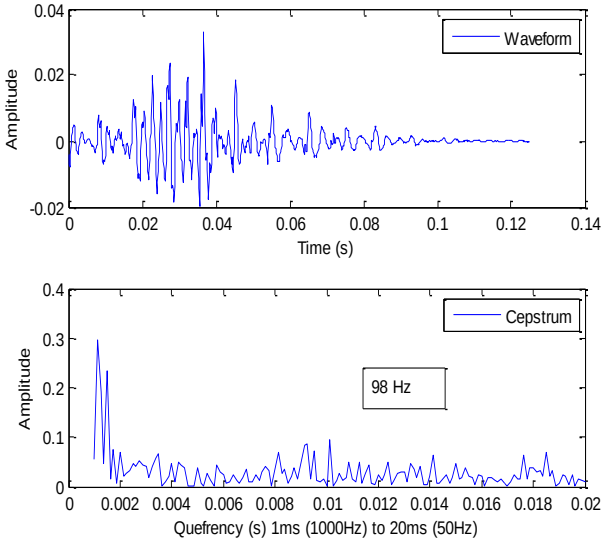


Fig. 4.11: (III): Male Voice Segment and Cepstrum (Top to Bottom).

For the female utterance (sp26.wav) the results are as follows (i) 4800 to 5600 (ii) 9600 to 10200 (iii) 12500 to 13500. The results are shown are shown in Fig. 4.12-4.14. The observed pitch frequencies are 129Hz, 216 Hz, and 205 Hz.

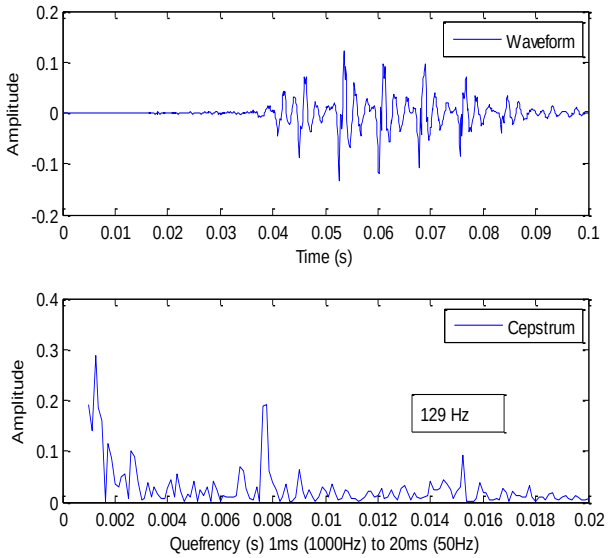


Fig. 4.12: (I): Female Voiced Segment and Cepstrum (Top to Bottom).

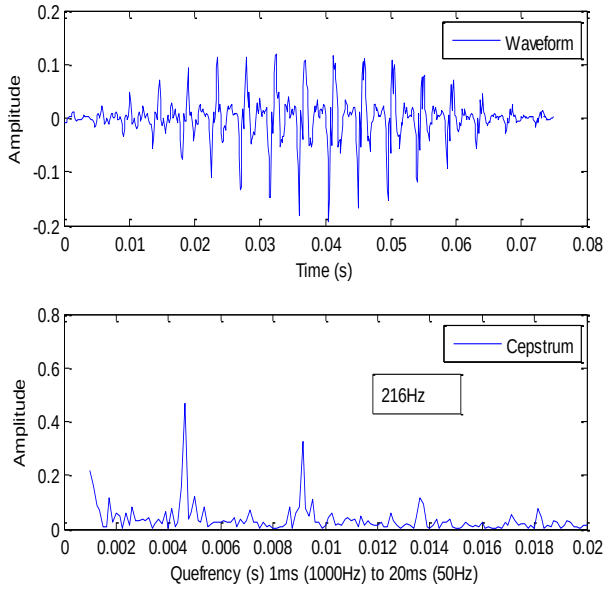


Fig. 4.13: (II): Female Voiced Segment and Cepstrum (Top to Bottom).

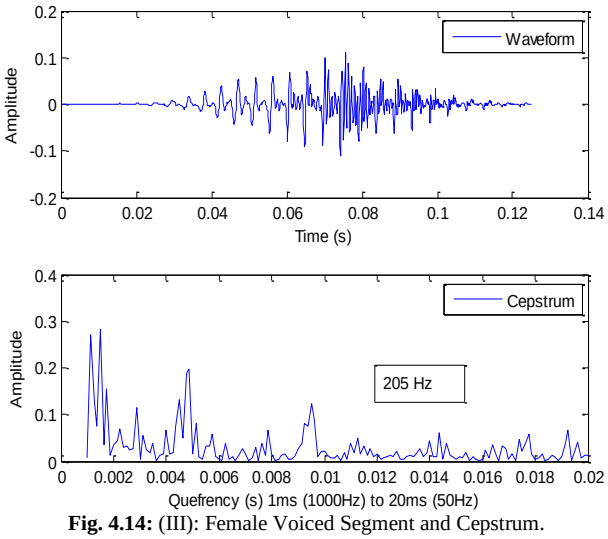


Fig. 4.14: (III): Female Voiced Segment and Cepstrum.

The pitch tract for male voice is shown in Fig. 4.15. The voiced parts are with constant pitch whereas the unvoiced parts have unstable values,

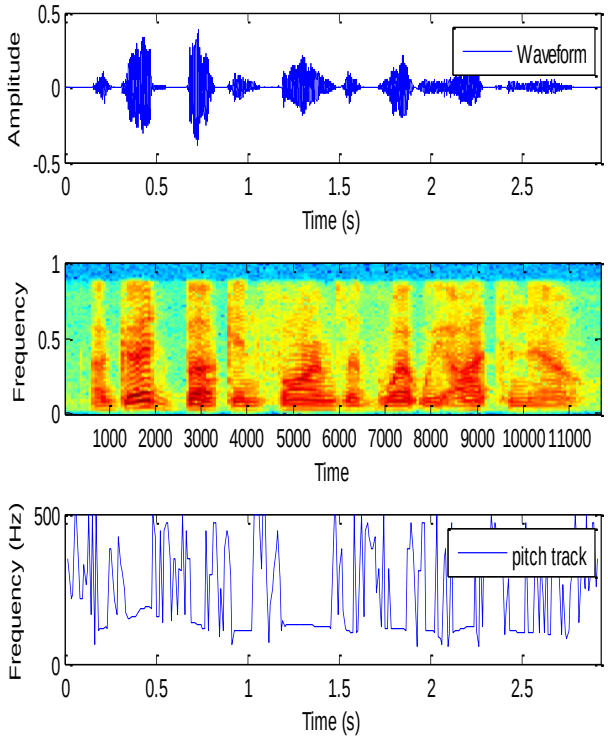


Fig. 4.15: Male Speech, Spectrogram and Pitch Tract (Top to Bottom).

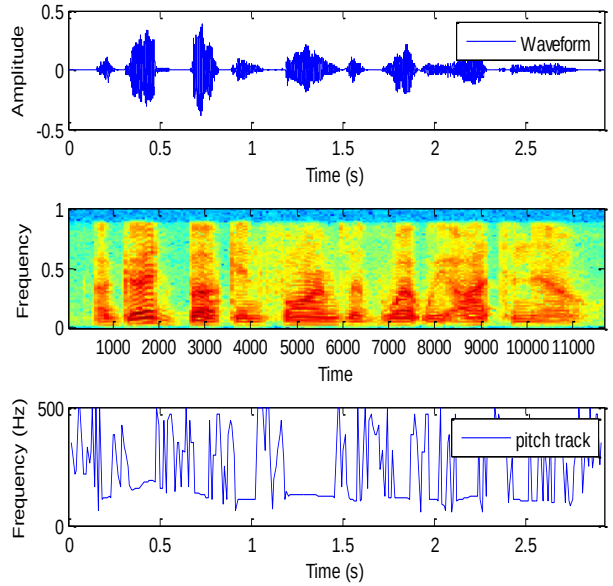


Fig. 4.16: Female Speech, Spectrogram and Pitch Track (Top to Bottom).

It is clear that the voiced part with constant pitch values whereas the unvoiced parts having different pitch values.

c) Formats Estimation and tracking

The formats represent the resonance frequencies of the vocal tract system. The male speech signal and tracking of the resonant frequencies are shown in Figure 4.17. The LPC method is used to extract the resonant values.

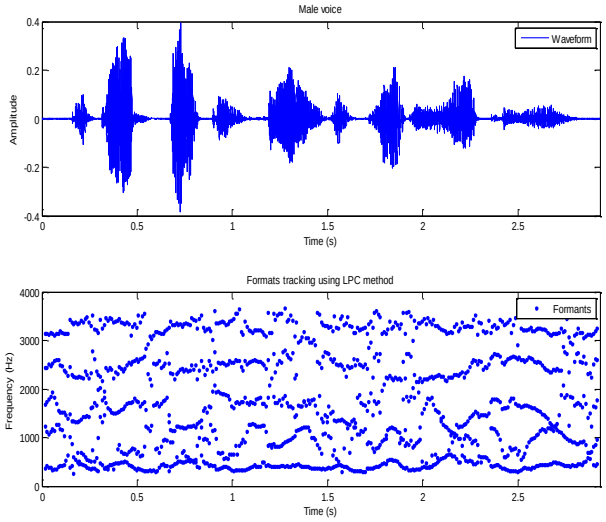


Fig. 4.17: Male Speech, Formant Tracking (Top to Bottom).

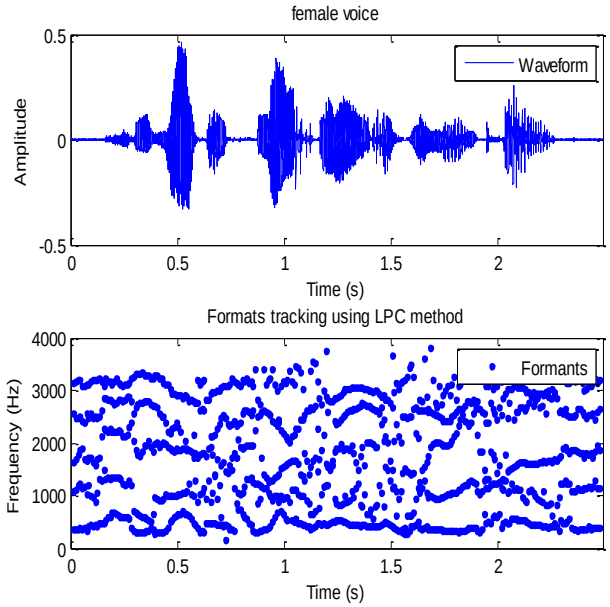


Fig. 4.18: Female Speech and Formant Tracking (Top to Bottom).

Similarly the resonant values of female voice are shown in Figure 4.18. Even though the LPC method has some limitations, it is widely used in speech processing applications as it is robust for noise signals.

Table 4.1 and 4.2 shows the results of the auto-correlation and cepstrum algorithms with respect to clean and noisy speech. The clean and noisy speech data base selected from Noizus [9].

Table 4.1: Pitch Comparisons for the Clean and Noisy Speech (Train Noise, 0db), (Female Speech)

Voiced segment	Auto-correlation		Cepstrum	
	Clean/Hz	Noisy/Hz	Clean/Hz	Noisy/Hz
4800-5600	421	444	129	129
9600-10200	210	210	216	216
12500-13500	200	205	205	500

Table 4.2: Pitch Comparisons for the Clean and Noisy Speech (Train Noise, 0db), (Male Speech)

Voiced	segment	Auto	correlation	Cepstrum	
Clean	Noisy	Clean	Noisy		
3600	4600	186	181	186	500
7000-	7200	111	160	500	500
12500	13500 103	470	98	101	

The simulations are conducted for clean and noisy speech (0dB train noise). It is observed from Tables 4.1 and 4.2; the pitch is increased in noisy speech compared to clean speech

5. Conclusion

Pitch is one of the most important characteristic needed in many speech processing applications. In this paper pitch is estimated using auto-correlation and cepstrum methods. Male and Female speech signal simulations are executed using Matlab programming. The simulations are conducted at different voiced segments. The methods are shown to be efficient in pitch estimation.
The representation or modeling of speech in terms of formants is useful in several areas of speech processing, coding, recognition, synthesis and enhancement. However estimating formants is more difficult than simply searching for peaks in an amplitude spectrum.

References

[1] Krishna Kohhatkar, Mahesh Kolte, Jyothi Lele, “Implementation of pitch detection algorithms for pathological voices”, International Conference on Inventive Computational Technologies”, Vol. 3, 26 August 2016, Coimbatore, India.

[2] MNA Aadit, SG Kirtana, MT Mahin, “Pitch and formant estimation of bangle speech signal using auto-correlation, cesptrum and LPC algorithm”, 19th Itnernational Conference on Computer and Information Technology, 18-20 Dec. 2016, Dhaka, Bangladesh.

[3] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform." Proceedings of the IEEE, vol. 66, no. 1, pp. 51¬84, 1978.

[4] A. H. Nuttall, "Some Windows With Very Good Side lobe Behavior," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 1, pp. 84¬91, 1981.

[5] M. M. Sondhi, "New Methods of Pitch Extraction," IEEE Trans. on Audio and Electroacoustic, vol. 16, no. 2, pp. 262¬266, 1968.

[6] Kirill Sakhnov, Boris Simak, “Pitch detection algorithms and voiced/unvoiced classification for noisy speech”, 16th International Conference on System, Signals and Image Processing”, 28 Dec. 2009, Chalkida, Greece.

[7] Y. M. Cheng and D. O'Shaughnessy, "Automatic and Reliable Estimation of Glottal Closure Instant and Period," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 37, no. 12, pp. 1805¬1815, 1989.

[8] Cansu, Seda, Yard, “A study on pitch detection algorithms”, 23rd Conference on Signal Processing and Communications Applications”, 16-19 May 2015, Malatya, Turkey.

[9] NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms

[10] P. R. Cook, "An Automatic Pitch Detection and MIDI Control System for Brass Instruments," Stanford Center for Computer Research in Music and Acoustics.

[11] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 27, no. 4, pp. 309¬319, 1979.

[12] J. F. Deem, W. H. Manning, J. V. Knack and J. S. Matesich, "The Automatic Extraction of Pitch Perturbation Using Microcomputers: Some Methodological Considerations," Journal of Speech and Hearing Research, vol. 32, pp. 689-697, 1989.

[13] H. Chamberlin, "Musical Applications of Microprocessors". New Jersey: Hayden Book Company, 1980.

[14] J. M. Cioffi, "Limited Precision Effects in Adaptive Filtering," IEEE Transactions on Circuits and Systems, vol. 34, no. 7, pp. 821¬833, 1987.

[15] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 22, no. 5, pp. 353¬362, 1974.

[16] R. L. Miller and E. S. Weibel, "Measurements of the Fundamental Period of Speech Using a Delay Line," Journal of the Acoustical Society of America, vol. 28, Abstract, 1956.

[17] A. Moorer, "The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 22, no. 5, pp. 330¬338, 1974.