Weighted Multi-band Summary Correlogram (MBSC)-based Pitch Estimation and Voice Activity Detection for Noisy Speech

RASHIDA AKHTAR RAKHI[1], HUMAYAN KABIR RANA[2] AND MD. KISLU NOMAN[3]
[1]Dept. of CSE, Varendra University, BANGLADESH (rakhi.pustcse@gmail.com)
[2]Dept. of CSE, Green University of Bangladesh, BANGLADESH (humayan.pustcse@gmail.com)
[3]Dept. of CSE, Pabna University of Science and Technology, BANGLADESH (md.k.noman@gmail.com)

Abstract: - The pitch estimation and Voice activity detection (VAD) is the task of classifying an acoustic signal stream into voiced and unvoiced segments that plays as a crucial preprocessing tool to a wide range of speech applications. In this paper, a weighted multi-band summary correlogram (MBSC)-based pitch estimation algorithm (PEA) as well as voice activity detection (VAD) is proposed. The PEA performs pitch estimation and voiced/unvoiced (V/UV) detection via novel signal processing schemes that are designed to enhance the MBSC's peaks at the most likely pitch period. This technique computes an independent normalized auto-correlation function (NACF) for each channel or frame which is relatively insensitive to phase changes across channels firstly and then filtered these NACFs to remove a significant portion beyond the pitch range 50-500 Hz and then finding an adaptive threshold from filtered NACFs. This threshold acts as a pitch position indicator and a voiced/unvoiced region detector. The accurate pitch period is obtained from the weighted MBSC. The proposed algorithm has the lowest gross pitch error (%GPE) for noisy speech in the evaluation set among the algorithms evaluated. The proposed PDA also achieves the lowest average voicing detection errors.

Keywords:- multi-band summary correlogram, empirical mode decomposition, normalized autocorrelation, voiced/unvoiced speech.

1 Introduction

Speech is a complex wonder and one of the unique characteristics of the human species [1]. Speech can be divided into numerous voiced and unvoiced regions. The classification of speech signals into voiced/unvoiced provides a preliminary acoustic segmentation for speech processing applications, such as speech synthesis, speech enhancement, and speech recognition [2]. The estimation of pitch period of speech signal plays an important role in different speech processing applications including speech enhancement using harmonic model, automatic speech recognition and understanding, analysis and modeling of speech prosody, low-bit-rate speech coding etc. Although many methods of pitch estimation have been proposed, reliable and accurate detection is still challenging task. The speech waveform is weakly periodic and the instantaneous values of pitch are different even within a frame. The presence of noise further complicates the problem and deteriorates the performance of the pitch estimation algorithms (PEAs).

Fundamental frequency (F0) or pitch information of speech is required for many speech applications. Although F0 estimation is a well-researched topic, accurate F0 estimation in noise still poses a challenge. There are many pitch detection algorithms such as the short-time average magnitude difference function (AMDF) [3], short-term autocorrelation function (ACF) [8], direct time domain fundamental frequency estimation (DFE) [4], weighted autocorrelation (WAC) [5], and zero-cross rate with autocorrelation [6] algorithms. Although many pitch detection algorithms have been discovered, few of them have been built in special purpose digital hardware able to work on noisy environment and real time [7]. In most of the existing algorithm to estimate pitch used the Fourier transform and wavelet transform for signal decomposition. Although speech signal is non-stationary in nature, those transformations assume that it is piecewise stationary. The speech decomposition is performed by fitting some predefined bases without satisfying its non-stationary nature. Hence, Fourier transform and wavelet analysis make great contributions, they suffer from many shortcomings in case of nonlinear and non-stationary signals [8]. Therefore, in this thesis work a most recently developed technique EMD(Empirical Mode Decomposition) is used for signal decomposition before pitch estimation which is a new data analysis method for non-linear and non-stationary signals, has made a new path for speech enhancement research and also data adaptive decomposition method that decomposes data into zero mean oscillation components, named intrinsic mode function(IMFs) but its mode mixing is evident between IMFs or a single mode is 'leaked' into two IMFs and splitting into IMFs is also time consuming[9]. To overcome such demerit a new pitch estimation as well as voiced/unvoiced detection technique proposed here based on Multi-band Summery based Correlogram. The MBSC is computed from the filtered normalized autocorrelation function (NACF) from each channel separately so it is insensitive to of phase changes across channels[10]. The MBSC has been weighted and plotted as an image for illustration. The proposed method is very easy to implement and the periodicity of speech is well represented in the multi-band summary correlogram(MBSC).

2 Background and Preliminary Results

This paper describes a pitch detector that mimics the human perceptual system. The model of human pitch perception has three stages. The inner ear or cochlea encodes the information in the acoustic signal into a multi-channel representation that may be thought of as instantaneous nerve firing probabilities. The second stage of processing produces a correlogram, a two-dimensional image in which each row is the running short-time autocorrelation of the corresponding cochlea channel. Finally, a pitch detector combines the information in all the channels of the correlogram to decide on a single pitch. Humans can perceive multiple pitches but for the purposes of this paper we choose a “best” pitch. Figure 1 three stages of neural processing are used in our model of pitch perception [11].

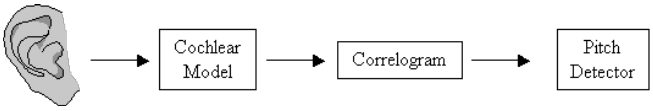

Figure 1: Three stages of neural processing used in the algorithm.

A cochlear model designed by Lyon and described by Slaney [12] to convert a sound waveform into a vector of numbers that represent the information sent to the brain. This system is diagrammed in Figure 2.

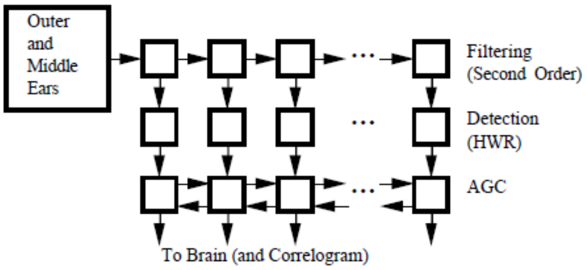

Figure 2: Block diagram of cochlear model.

It is important to remember that the cochlear model used here does not try to accurately model the internal structure of the ear but only to approximate the information contained in the auditory nerve. Other, more accurate models can be substituted to get better results.

A cascade of second order filters is used to model the propagation of sound along the Basilar Membrane (BM). At each point along the cochlea the BM responds best to a broad range of frequencies and it is this movement that is sensed by the Inner Hair Cells. The “best” frequency of the cochlea varies smoothly from high frequencies at the base to low frequencies at the apex.

Inner Hair Cells only respond to movement of the BM in one direction. This is simulated in the cochlear model with an array of Half Wave Rectifiers (HWRs) that detect the output of each second order filter. The HWR nonlinearity serves to convert the motion of the BM at each point along the cochlea into a signal that represents both the envelope and fine time structure.

Finally, four stages of Automatic Gain Control (AGC) allow the cochlear model to compress the dynamic range of the input to a level that can be carried on the auditory nerve. The multi-channel coupled AGC used here simulates the ear’s adaptation to spectral tilt as well as to loudness [13].

2.1 The Correlogram

The correlogram is an animated picture of the sound that shows frequency content along the vertical axis and time structure along the horizontal axis [14]. If a sound is periodic, the autocorrelation functions for all cochlear channels will show a peak at the horizontal position that corresponds to a correlation delay equal to the period of repetition. This is generally equal to the perceived pitch period. Since the peaks in all channels, or rows of the image, occur at the same delay, or horizontal position, they form a vertical line in the image. The duplex theory says that sounds with a perceived pitch, even if they are not periodic, will produce a vertical structure in the correlogram image at the delay equal to the perceptual pitch [15].On the other hand, formants, or narrow resonances in the frequency domain, are displayed as horizontal bands in the correlogram. The correlogram is computed by finding the (short-time, windowed) autocorrelation of the output of each cochlear frequency channel [16].

2.2 A Pitch Detector

Pitch detector consists of four steps. A preprocessing step modifies the correlogram to enhance the peaks. The value at each time lag in the enhanced correlogram is then summed across all frequencies. Peak locations at this stage give estimates of all the possible periodicities in the correlogram. The third step is to combine evidence at the subharmonics of each pitch to make the pitch estimate more robust. Finally, the largest peak is picked, being careful to avoid octave errors, and a numerical value of the pitch is determined based on the location of the peak [17].

3 The Proposed Method

Pitch is that auditory attribute of sound according to which sounds can be ordered on a scale from low to high. A pitch detector performs both pitch estimation and voiced/unvoiced (V/UV) detection. In pitch estimation, the rate of vocal-fold vibration is estimated, while in V/UV detection, voiced or quasi-periodic speech frames are distinguished from the rest of the signal.

This proposed approach accurately models how humans perceive pitch. It correctly identifies the pitch of complex harmonic and inharmonic stimuli, and is robust in the face of noise and phase changes. This perceptual pitch detector combines a cochlear model with a bank of autocorrelators. By performing an independent auto-correlation for each channel, the pitch detector is relatively in-sensitive to phase changes across channels. The information in the correlogram is filtered, nonlinearly enhanced, and summed across channels. Peaks are identified and a pitch is then proposed that is consistent with the peaks.

3.1 Adaptive Thresholding

The correlogram will be weighted by a threshold that differentiate voiced/unvoiced regions and find weighted correlogram. The threshold is defined as,

$$\mu = \frac{1.25 \times SD\ (MACF)}{Mean\ (MACF)} \qquad (1)$$

Where, the autocorrelation function of input speech, $y(n)$, at time lag τ can be expressed as

$$R_{xy}\ (\tau) = \frac{1}{N} \sum_{n=0}^{N-|\tau|-1} y(n) y(n+|\tau|) \qquad (2)$$

$$NACF = \frac{R_{xy}\ (\tau)}{Max\ (R_{xy}\ (\tau))} \qquad (3)$$

$$MACF = Max\ (NACF) \qquad (4)$$

SD(MACF) = Standard deviation of MACF and Mean(MACF)=Mean of MACF

The higher value of NACF from threshold exhibits voiced regions and lower values exhibits unvoiced regions of speech signal.

3.2 Weighted Multi-band Summary Correlogram-based Pitch estimation and Voice Activity Detection (VAD) for noisy speech

A multi-band summary correlogram (MBSC)-based pitch detection algorithm (PDA) is proposed. The PDA performs pitch estimation and voiced/unvoiced (V/UV) detection via novel signal processing schemes that are designed to enhance the MBSC's peaks at the most likely pitch period. The steps of the proposed algorithm are as follows:

a) Input speech, y(n), is divided into frames, each of duration 25.6 ms having 512 samples for 20 kHz sampling frequency with 16-bit resolution of the given speech. Each frame is overlapping with 100 samples
b) Find the normalized autocorrelation function (NACF) of each frame.
c) Filter the NACF to remove a significant portion beyond the pitch range 50-500 Hz and find first filtered NACF.
d) Find a threshold from maximum values of first filtered NACF using Eq. (4.1) that act as boundary for voiced and unvoiced regions of speech signal.
e) Compare first filtered NACF with the threshold. Keep the same value of first filtered NACF if it is greater than threshold, otherwise zero to obtain second filtered NACF.
f) Multiply second filtered NACF with first filtered NACF to obtain 2D weighted autocorrelation (WAC).
g) Plot the 2D WAC as an image called weighted correlogram.
h) Determine the range of pitch by [finding the maximum value of the concentrated energy in the weighted correlogram (WAC).
i) Find the maximum value of second filtered NACF from the WAC for each frame and its corresponding time index.
j) Corresponding time index or autocorrelation lag of each frame, if greater than 1 indicates the pitch period, otherwise the frame is unvoiced i.e. has no period.

4 Experimental Results

Keele pitch extraction reference database [18] is used to test the performance of the proposed pitch estimation method. Both male (M2-M3) and female (F2-F3) mature speakers' recorded speech are used here. The speech signal was sampled at 20 kHz with 16-bit resolution segmented into frames of length 25.6ms with 5ms overlapping.

The experiments are conducted by adding white Gaussian noise to the signal. Each 25.6 msec analysis frame is weighted by a 512-point rectangular window $w(n)$ ($w(n)$=1 for 0≤n≤511 and $w(n)$=0, elsewhere). The frame shift was set to be 10 msec as used to generate the reference pitch values given in the database.

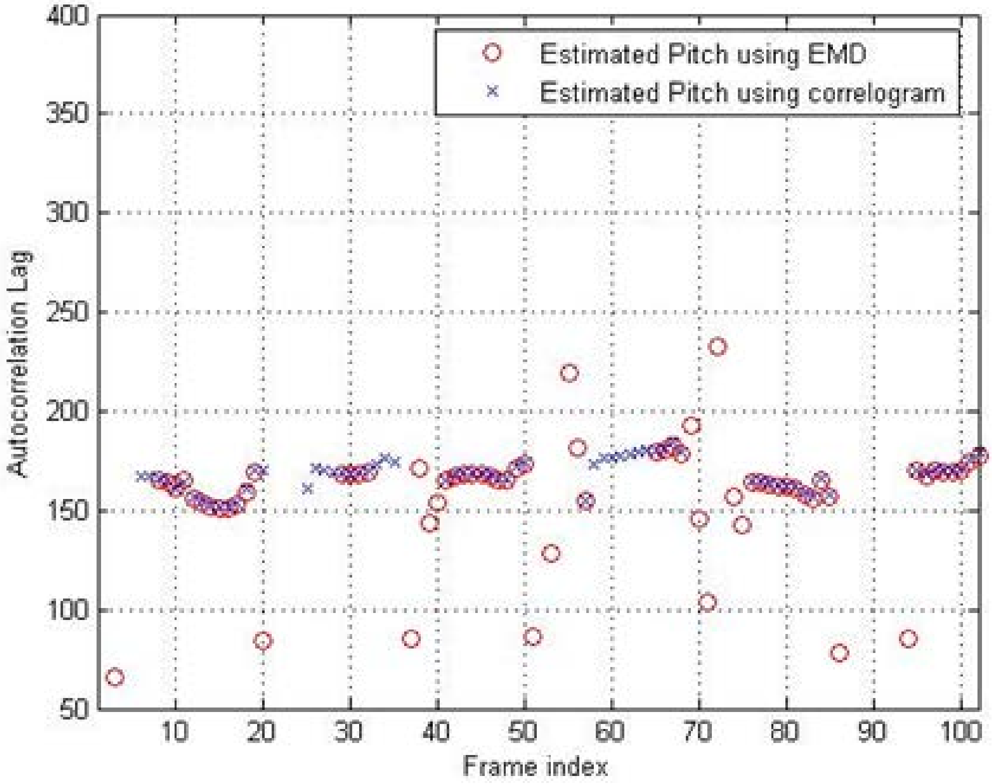

Figure (a): Male Speaker

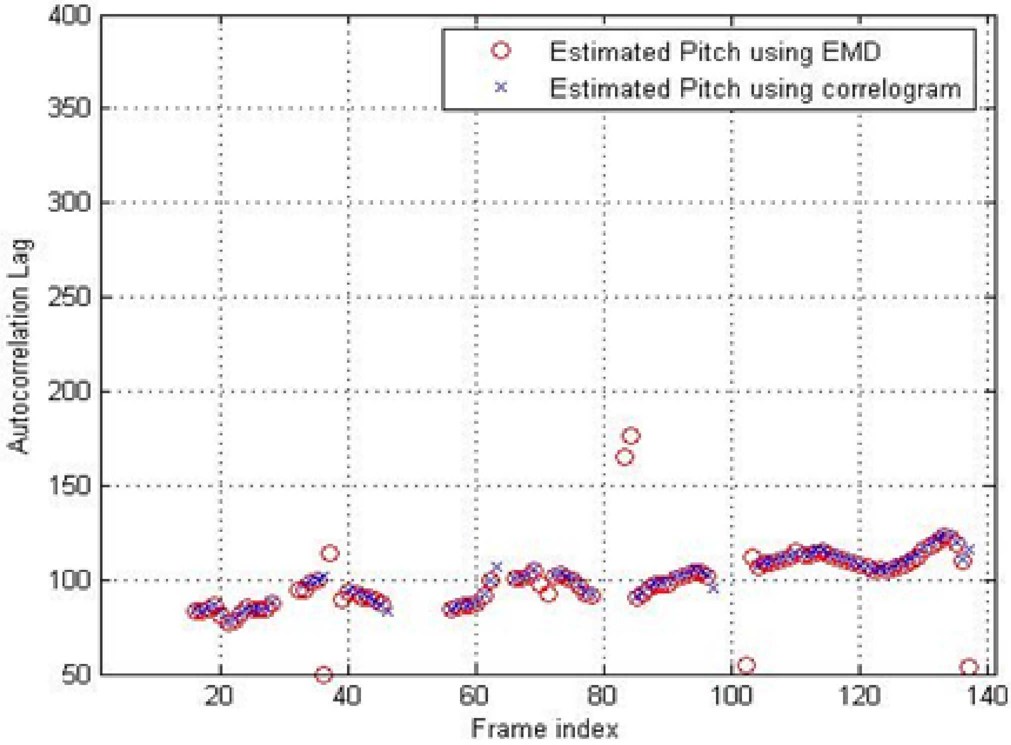

Figure (b): Female Speaker

Figure 3: comparison of pitch estimation between EMD and MBSC for (a) male and (b) female speaker at SNR=20 dB and SNR=10 dB

SNR(dB)		-15	-5	0	10	20
M1	MBSC	28.89	9.87	4.07	2.03	2.03
	EMD	37.78	12.88	7.05	4.53	4.45
	WAC	61.42	25.20	15.23	7.57	6.17
M2	MBSC	37.79	6.67	1.89	1.23	0.17
	EMD	43.07	9.11	3.88	1.62	0.91
	WAC	69.20	24.29	12.07	3.38	1.2
F1	MBSC	51.45	12.87	5.57	1.19	1.03
	EMD	61.57	18.09	8.71	1.93	1.32
	WAC	68.56	23.05	11.58	3.97	1.93
F2	MBSC	47.46	12.23	2.67	1.27	1.00
	EMD	57.34	16.40	4.87	1.48	1.34
	WAC	67.32	21.78	7.85	1.64	1.62

Table 1: performance comparison of different pitch estimation algorithms

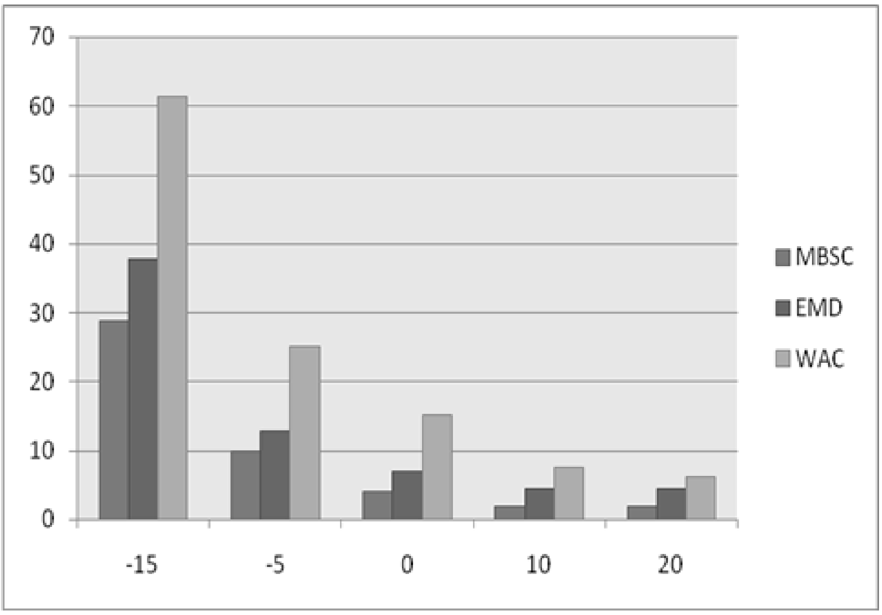

Figure 4: Performance of proposed method (%GPE)

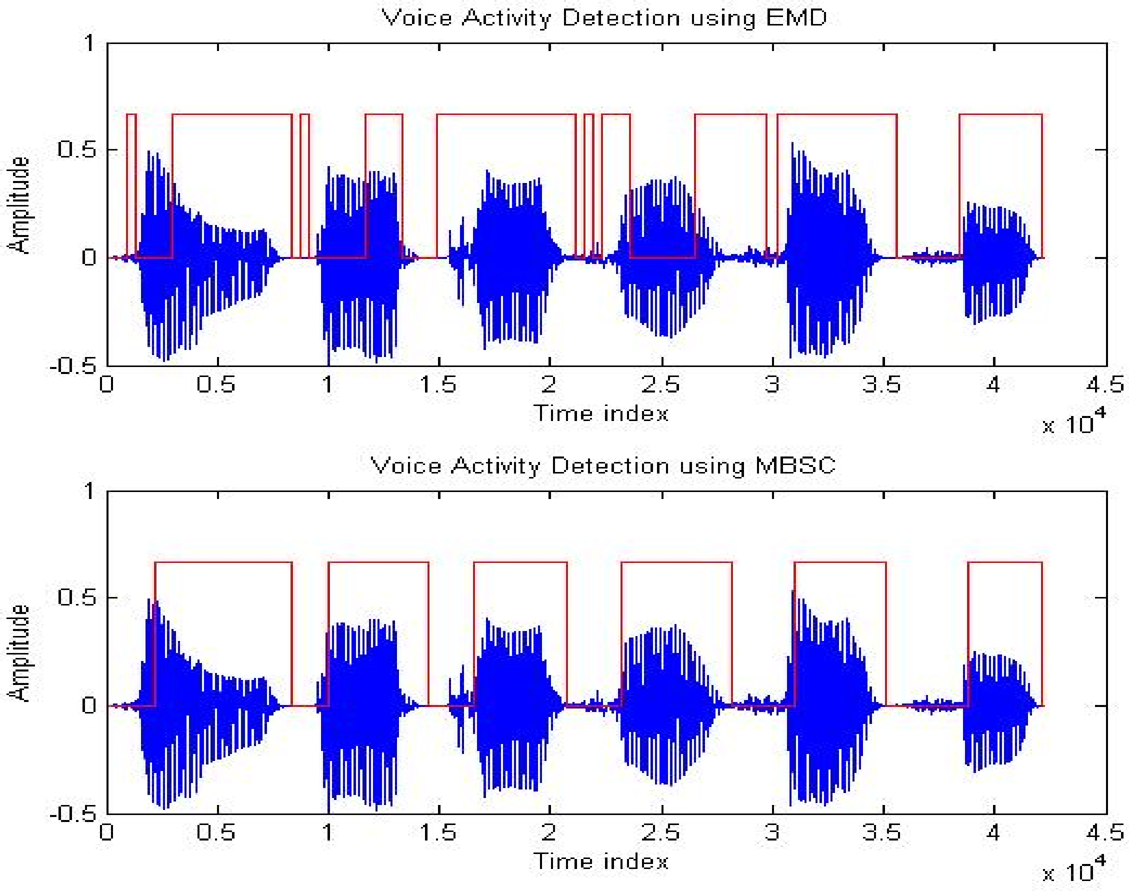

Figure 5: The voice activity detection (VAD) comparison between EMD and proposed MBSC for a recorded speech uttered by a female speaker at SNR =10 dB

Gross Pitch Error (GPE): If the estimated pitch for a frame deviates from the reference by >20%, we recognize the error as a gross pitch error (GPE). The index GPE is often expressed in percentage denoted as %GPE. The true pitch values are obtained from the original database. The performance comparison of different pitch estimation algorithms (PEAs) is presented using the modified Keele database which contains 'clearly voiced' reference frames as used in other works [11]. Results obtained using the three PEAs namely, the proposed one using MBSC, conventional EMD method [7] and WAC [6] are presented in Table 1.

The proposed method also proves its superiority in voiced/unvoiced detection shown in figure 5.

The result of the proposed method is noticeable in pitch estimation as well as voiced/unvoiced detection that produce lower error (%GPE) about less than 30% in pitch calculation in noise-free environment as well as in noisy condition while the EMD produce about 35% (GPE) [7] and 60% in WAC [5] shown in figure 4.

5. Conclusion

Voice activity detection (VAD) is the task of classifying an acoustic signal stream into speech and non-speech segments. The distinction between voiced and unvoiced sounds is that the excitation is quasi-periodic for voiced sounds, and white noise for unvoiced sounds. Pitch is one of the most important features of speech. In this proposed algorithm, pitch refers to the fundamental frequency contour of successive frames. Pitch of the segmented speech is estimated by searching the peak of the weighted NACF. The weighted NACF is plotted as weighted MBSC. The region for searching the pitch peak is set to be from 50 Hz to 500 Hz, which corresponds to the region of the fundamental frequencies of most males (50-250 HZ) and females (120-500 Hz).

From the above results that recent [7] EMD algorithm showed poor results in accuracy of pitch estimation and pattern recognition and also error prone and much time consuming. Error is occurred in the estimated pitch of this algorithm due to incorrect detection of voiced/unvoiced region and time complexity arises due to the decomposition of the signal empirically into IMFs. The proposed Weighted Multi-band Summary Correlogram (MBSC) method proves superiority in pitch estimation as well as voiced/unvoiced detection.

References:

[1] Zhaohua, W., and Huang, N. E., "A Study of the Characteristics of White Noise using Empirical Mode Decomposition Method", in Proc. Roy. Soc. Lond. A(460), pp. 1597-1611, 2004.

[2] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.

[3] K. Kasi, "Yet another algorithm for pitch tracking," Masters Thesis, Old Dominion University, Norfolk, VA, 2002.

[4] G.Fant, Acoustic theory of speech production, Mouton and Co., Gravenhage, The Netherland, 1960.

[5] T. Shimamura, H. Kobayashi, "Weighted Autocorrelation for Pitch Extraction of Noisy Speech", IEEE Trans. on Speech and Audio Processing, vol. 9, n. 7, pp. 727–730, October 2001.

[6] M. K. I. Molla, K. Hirose, N. Minematsu and M. K. Hasan, " Pitch Estimation of Noisy Speech Signals using Empirical Mode Decomposition", Interspeech 2007

[7] Md. Khademul Islam Molla, Keikichi Hirose and Nobuaki Minematsu , "Robust Voiced/Unvoiced Speech Classification using Empirical Mode Decomposition and Periodic Correlation Mode", Interspeech 2008.

[8] Hasan, M. K., Hussain, S., Setu, M. T. H. and Nazrul, M. N. I., "Signal reshaping using dominant harmonic for pitch estimation of noisy speech", Signal Processing, 86(5):1010-1018, 2005.

[9] S. Nakamura, K. Yamamoto, K. Takeda, S. Kuroiwa, N. Kitaoka, T. Yamada, M. Mizumachi, T. Nishiura, M. Fujimoto, A. Sasou and T. Endo, "Data Collection and Evaluation of AURORA-2 Japanese Corpus", IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 619-623, 2003.

[10] D. Talkin, A robust algorithm for pitch tracking (RAPT), Speech Coding and Synthesis, Elsevier Science, pp. 495-518, 1995.

[11] Whitham, G. B., "Linear and nonlinear waves", New York, Wiley, 1975

[12] Liu, S. C., "An approach to time-varying spectral analysis", J. EM. Div. ASCE 98, 245-253, 1973.

[13] Bitzer, Joerg, Simmer, K. U., and Kammeyer, K. D., "Multi-microphone noise reduction techniques as front-end devices for speech recognition", Speech Communication, vol. 34, pp. 3-12, 2001.

[14] Bitzer, Joerg, Simmer, K. U., and Kammeyer, K. D., "Mult-microphone noise reduction techniques as front-end devices for speech recognition", Speech Communication, vol. 34, pp. 3-12, 2001.

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

[16] Shah, J. K. et. al., "Robust voiced/unvoiced classification using novel features and Gaussian mixture model", in Proc. Of ICASSP04, 2004.

[17] Ahmadi, S., and Spanias, A. S., "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," IEEE Trans. Speech Audio Processing, vol. 7 No. 3, pp. 333-338, 1999.

[18] ftp://ftp.cs.keele.ac.uk/pub/pitch/

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper