Robust Determination of Periodic Correlation of Speech Signals using Empirical Mode Decomposition and Higher-Order Spectra

Md. Khademul Islam Molla[1, 2], Keikichi Hirose[1] and Nobuaki Minematsu[3]
[1]Graduate School of Information Science and Technology, [3]Graduate School of Engineering
The University of Tokyo, Tokyo, JAPAN
[2]Dept. of Computer Science and Engineering, The University of Rajshahi, Rajshahi, BANGLADESH
E-mail: {molla, hirose, nime}@gavo.t.u-tokyo.ac.jp

***Abstract*-This paper presents a new method of periodic/non-periodic (P/nP) classification of noisy speech signals. Empirical mode decomposition (EMD), a newly developed tool to analyze nonlinear and non-stationary signals is used to filter the additive noise with the speech signal. The normalized autocorrelation of the filtered speech signal is computed to enhance the periodicity of the analyzing speech signal if any. It is considered that the voiced speech (with fundamental periodicity) signal is periodically correlated and the unvoiced signal is not. A noise robust P/nP decision rule is formulated based on third-order autocumulants of the autocorrelation function of speech signal. The experimental results show that the use of EMD improves the classification performance and the overall efficiency is noticeable as compared to other existing algorithms.**

Keywords: Empirical mode decomposition, fundamental frequency, higher-order spectral analysis, periodic correlation, speech denoising, voiced/unvoiced speech classification.

I. INTRODUCTION

Reliable classification of short time speech signal into periodic and non-periodic (P/nP) is a crucial preprocessing step in many speech processing applications and is essential in most analysis and synthesis system. The essence of classification is to determine whether the speech production system involves the vibration of the vocal cords [1]. The speech signal originated from the speaker's vocal cord contains a sequence of periodic correlation. Such signal is also called voiced speech signal and unvoiced with absence of periodically correlated sequences. Hence the proposed classification problem can be treated as the discrimination of speech signals into voiced and unvoiced (V/UV) segments. The discrimination problem is an important one and has been worked on extensively during the last three decades [2].

The classification can effectively be performed using a single feature or parameter which is closely associated with the presence of periodicity and non-periodicity of the analyzing speech signal. Many algorithms have been reported for solving the detection problem [3] – [7]. In [3], Gaussian mixture model with cepstrum coefficients features is proposed for robust V/UV classification. A higher order statistics (HOS) based method is proposed in [4] for V/UV detection and pitch estimation simultaneously. The matching pursuit algorithm is used in [5] with Gabor decomposition. The wavelet transform is proposed in pitch and V/UV detection in [6]. A statistical model applied in autocorrelation domain is also reported in [7]. In most of the existing algorithms are not so much noise robust and also the intensive threshold and training data are required for classification. Such requirements are troublesome for the use in application domain.

The proposed method is noise robust and based on the higher-order spectral analysis (HOSA) [8] method for P/nP detection in speech signal without any training requirement. To reduce the effect of noise on speech signal, a data adaptive time domain filtering is proposed using newly developed empirical mode decomposition method [8]. Although speech signal is non-stationary in nature, Fourier based frequency domain filtering assumes that it is piecewise stationary. The speech decomposition is performed by fitting some predefined bases without satisfying its non-stationary nature. Whereas, EMD based approach decompose the speech signal as non-stationary time series and hence better performance in noise filtering.

It is assumed that the speech signal contains periodic correlation if the fundamental frequency is carried with. The proposed method is based on determining the presence of fundamental periodicity contained with the analyzing speech signal. The method is based on higher-order spectral analysis (HOSA) finding the location of the highest energy peak in the bi-spectrum domain [9]. The autocorrelation function (ACF) makes the periodicity more prominent if any. The proposed HOSA model, less affected by additive Gaussian type noises is applied in the autocorrelation domain rather than original time domain of the speech signal. The presence of periodicity detection method is implemented in spectral domain to classify the speech segment into P/nP based on that it contains periodic correlation (fundamental frequency) or not.

II. EMD BASED NOISE REDUCTION

Empirical mode decomposition (EMD) represents any time-domain signal into a finite set of AM-FM oscillatory components which are the bases of the decomposition. The key benefit of using EMD is that it is an automatic decomposition and fully data adaptive. The principle of the EMD technique is to decompose a signal $x(n)$ into a sum of band-limited functions, $g_b(n)$, called intrinsic mode functions (IMFs). Each IMF must satisfy two properties: (i) the number of extrema and the number of zero crossings are either equal or differ by one; (ii) the mean

value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. A special "sifting" process is employed to extract all of the IMFs [10, 11]. This sifting process is described as follows.

Firstly, the upper- and lower-envelopes of the signals $x(n)$, as well as their mean value $\mu_1(n)$, are calculated respectively. The first step of the sifting process is to calculate the difference $h_1(n)=x(n)-\mu_1(n)$. However, $h_1(n)$ rarely satisfies the two IMF properties and is not taken as the first IMF of the signals straightway. Therefore, the sifting usually has to be implemented for more times, where the "difference" obtained in the previous sifting is taken as "signals" in present sifting. If after $(k+1)^{\text{th}}$ sifting, corresponding difference $h_{1k}(n)=h_{1(k-1)}(n)-\mu_{1k}(n)$ satisfies the IMF properties, then it can be taken as the first IMF component, denoted by $g_1(n)$, that is, $g_1(n)= h_{1k}(n)$. In practice, to determine whether or not $h_{1k}(n)$ well satisfies the IMF properties, we usually use so-called standard deviation (δ) criterion, that is, to check if the following inequality holds [10]:

$$\delta_k = \sum_{n=0}^{N-1}\left[\frac{\left|h_{1(k-1)}(n)-h_{1k}(n)\right|^2}{h_{1(k-1)}^2(n)}\right] \le 0.2-0.3\,, \qquad (1)$$

where N is the frame length. Next, taking rest data $r_1(n)=x(n)-g_1(n)$ as "new" signals and implementing the sifting process on it, we can obtain the second IMF $g_2(n)$. This procedure should be repeatedly used for B (total number of IMF components) times until the last residue $r_B(n)$ becomes a monotonic function. When the decomposition procedure finished, the signals then can be expressed as

$$x(n)=\sum_{b=1}^{B} g_b(n)+r_B(n)\,, \qquad (2)$$

where $g_1(n), g_2(n),..., g_B(n)$ are all of the IMFs included in the signals and $r_B(n)$ is a negligible residue.

Another way to explain how EMD works is that it extracts out the highest frequency oscillation that remains in the signal. Thus locally, each IMF contains lower frequency oscillation than the one extracted just before. The decomposition is performed in a dyadic nature [12]. Being data adaptive, the basis usually offers a physically meaningful representation of the underlying processes. Unlike Fourier transform, there is no need of considering the signal as a stack of harmonics and, therefore, EMD is ideal for analyzing non-stationary and nonlinear data. Each IMF is considered as a mono-component contribution such that the derivation of instantaneous amplitude and frequency provides a physical significance. The advantage of this time-space filtering is that the resulting band passed signals preserve the full non-stationary in physical space. This filtering method is intuitive and direct, its basis is a posteriori and data adaptive. The completeness of the decomposition is given by the Eq. (2). The original signal can easily be reconstructed by simply adding the bases with negligible error term.

A noisy speech signal and some selected IMF components are shown in Fig 1. It can be observed that higher order IMFs contain lower frequency oscillations than that of lower order IMFs. This is reasonable, since sifting process is based on the idea of subtracting the component with the longest period from the data till an IMF is obtained. Therefore the first IMF will have the highest oscillating components; the components with the highest frequencies. Consequently, the higher the order of the IMF, the lower its frequency content will be.

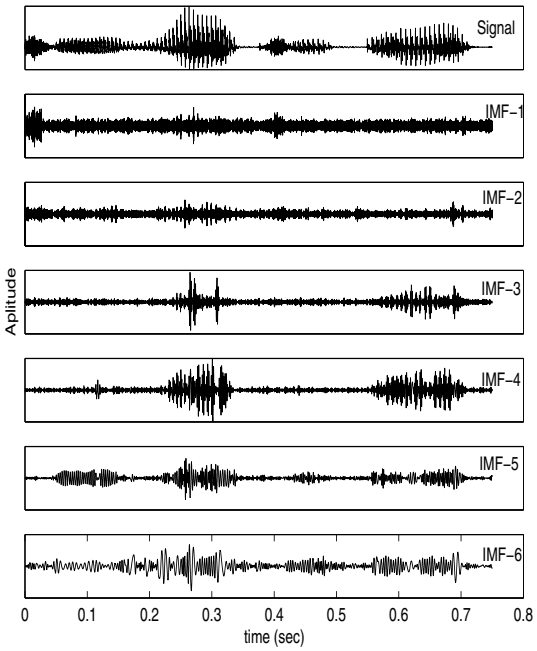

Figure 1. The illustration of EMD. A noisy speech signal at 10 dB SNR and its first 6 IMFs out of 13.

A. Instantaneous Frequency

Instantaneous frequency (IF) represents signal's frequency at any time instance and it is defined as the rate of change of the phase angle at the instant of the "analytic" version of the signal. Every IMF is a real valued signal. The discrete Hilbert transform (HT) denoted by $H_d[.]$ is used to compute the analytic signal for an IMF. Then the analytic version of the b^{th} IMF $g_b(t)$ is defined as:

$$z_b(t)=g_b(t)+jH_d[g_b(t)]=a_b(t)e^{j\theta_b(t)} \qquad (3)$$

where $a_b(t)$ and $\theta_b(t)$ are instantaneous amplitude and phase respectively of the b^{th} IMF. The analytic signal is advantageous in determining the instantaneous quantities such as energy, phase and frequency. The discrete-time IF of b^{th} IMF is then given as the derivative of the phase $\theta_b(t)$ calculated at t i.e.

$$f_b(t)=\frac{d\tilde{\theta}_b(t)}{dt} \qquad (4)$$

where $\tilde{\theta}_b(t)$ represents the unwrapped version of instantaneous phase $\theta_b(t)$. The derivative in Eq. (4), is evaluated at discrete instant of t. It should be noted that

such derivative introduces the abrupt fluctuations of IF and hence nonlinear smoothing is required. Here, the moving average smoothing filtering is used to remove such fluctuations. The filtering scheme improves the effectiveness of computing IF using discrete derivative. The concept of IF is physically meaningful only when applied to mono-component signals. In order to apply the concept of IF to arbitrary signals it is necessary to decompose the signal into a series of mono-component contributions. In the recent approaches [13], EMD technique decomposes a time domain signal into a series of mono-component IMFs. Then the IF derived for each component provides the meaningful physical information.

B. Noise Filtering

Although the IMFs may have frequency overlaps but at any time instant, the instantaneous frequencies represented by each IMF are different. This phenomenon can be well understood in Figure 2 which shows the instantaneous frequencies of the first 6 IMFs. Therefore, EMD is an effective decomposition of non-linear and non-stationary signals in terms of their local frequency characteristics.

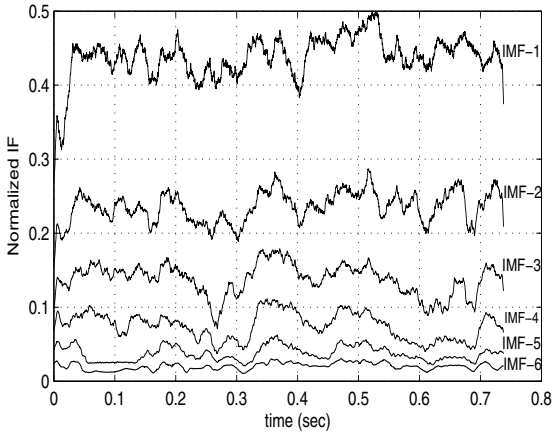

Figure 2. Instantaneous Frequencies of the first 6 IMFs.

With these powerful characteristics, recent studies have shown that it is possible to successfully identify and remove a significant amount of the noise components from the IMFs of a noisy speech. Although all IMFs contain energy from both the original speech and the noise, the amount of the energy distribution is different. Since speech signals are mainly concentrated in the low and mid frequency bands, the high frequency noise components dominate the first IMFs. For instance, in case of white noise, most of the noise components are centered on the first two IMFs, while the speech signals dominate between 3rd and 6th IMFs, as can be observed in Figure 1. Therefore, EMD makes it possible to some extent separate the high frequency noise from the major speech components. The instantaneous frequency vector is normalized between 0 and 0.50 to align with the Nyquest frequency. Then the IMFs with higher frequencies (>1.5kHz) are discarded. Thus most of the high frequency noise will be filtered out. The rest of the IMFs (including residue) is summed up to reconstruct the speech signal with less noise which is termed here as pre-filtered noisy speech (PFNS) what will be processed to detect the periodicity.

III. DETECTING PERIODIC CORRELATION

The presence of periodic correlation i.e. the fundamental frequency contained inside the speech signal is determined by analyzing the energy distribution in bispectrum domain. The range of fundamental periodicity of the speech signal is considered between 50Hz and 500Hz. The use of HOSA based model is more efficient when the noise robustness is in consideration. The motivation to employ higher order (k>2) cumulants and polyspectra is given by the following:

i) If $z(n)=x(n)+y(n)$, and $x(n)$ and $y(n)$ are mutually independent processes, then $C_{kz}(\boldsymbol{m}_k)=C_{kx}(\boldsymbol{m}_k)+C_{ky}(\boldsymbol{m}_k)$; where $\boldsymbol{m}_k=(m_1,m_2,\ldots,m_{k-1})$
ii) If $x(n)$ is Gaussian, then $C_{kx}(m_k)=0$, for k>2
iii) Hence, if $z(n)=x(n)+w(n)$, where $w(n)$ is Gaussian and independent of $x(n)$, then for k>2, $C_{kz}(\boldsymbol{m}_k)=C_{kx}(\boldsymbol{m}_k)$. Thus we can recover the higher order cumulants of non-Gaussian signal even in presence of Gaussian noise.

To detect the presence of periodicity within the analyzing speech signal higher-order (k=3) sequence of spectral cumulats is employed to reduce the noise effects.

A. HOSA Based Model

The Bispectra of third order cumulants is used here to formulate the robust decision rule for P/nP classification of speech signals. The cumulant spectra are of great importance in the analysis of stochastic signal such as speech [8]. The third order spectra called bispectrum which is defined to be the Fourier transform of the third order cumulant sequence, is employed here. The speech signal is segmented into blocks considering that each block $x(n)$ represents a zero-mean (subtracting mean from the original sequence) sequence of stationary process. The third-order cumulant is defined as:

$$C_3^x(\tau_{1,}\tau_2) = E\{x^*(n)x(n+\tau_1)x(n+\tau_2)\} \qquad (5)$$

Then the third-order polyspectrum (bispectrum) as a function of two frequencies is represented as

$$S_3^x(f_1,f_2) = \sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty} C_3^x(\tau_1,\tau_2)e^{-j2\pi f_1\tau_1}e^{-j2\pi f_2\tau_2} \qquad (6)$$

For a real-valued process, symmetry properties of cumulants carry over to symmetry properties of polyspectra [15]. The symmetry properties of the bispectrum are given by [16]:

$$\begin{aligned} &S_3^x(f_1,f_2) = S_3^x(f_2,f_1) = S_3^x(f_1,-f_1-f_2) \\ &= S_3^x(f_1-f_2,f_2) = S_3^x(f_1,-f_2) \end{aligned} \qquad (7)$$

Hence the nonredundant region of support for the bispectrum is the triangle with vertices (0,0), (1/3,1/3) and (1/2,0). The most of the energies contained in the signal are concentrated around the energy peak in the bispectrum.

B. Implementation

The proposed HOSA based model is implemented on the normalized autocorrelation (NACF) of the speech signal rather than in time domain. The noisy speech pre-

filtered to reduce the noise effect hence obtaining the PFNS signal $\varphi(n)$, $0 \le n \le N\text{-}1$ is used in periodicity detection.

The NACF produces better results in P/nP detection than the simple autocorrelation function as the peaks are more prominent and the less affected by the rapid variation in the signal amplitude. In this paper, the NACF of the PFNS signal $\varphi(n)$, $0 \le n \le N\text{-}1$ is used to detect the presence of periodicity and is computed as

$$NACF\ (k) = \frac{1}{\sqrt{\lambda_0 \lambda_k}} \sum_{n=1}^{N-1} \varphi(n)\varphi(n+k) \quad , \qquad (8)$$

where

$$\lambda_k = \sum_{n=k}^{k+N-K} \varphi^2(n) \quad , 0 \le k \le K\text{-}1. \qquad (9)$$

Even with the aforementioned pre-processing, the periodicity determination with NACF may give erroneous results under strong noisy condition due to the presence of spurious peaks obscuring the actual prominent peaks and also due to the inherent shortcoming introduced with the NACF. The NACF functions of noisy voiced speech signal and its pre-filtered one are shown in Figure 3. It is observed that the PFNS produces more prominent peaks in the NACF domain.

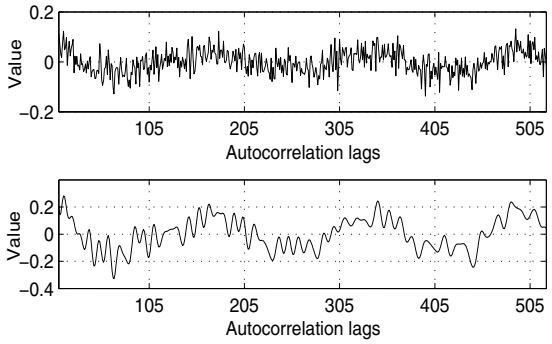

Figure 3. NACF of noisy speech signal (upper) and of PFNS signal (lower one).

The bispectrums of different signals are shown in Figure 4. In the bispectrum (Figure 4(a)) of the periodic speech signal (with 0dB SNR) before noise filtering, the energy distribution is spread over a wide range or frequencies, whereas for NACF signal shown in Figure 4(b), it is relatively conversed within smaller range of frequencies. In the normalized autocorrelation of the noisy speech signal, the fundamental periodicity becomes more prominent and hence the energy concentration is within the range of fundamental frequency (50Hz – 500Hz, as illustrated in Figure 4(c)) of the speech signals. In the bispectrum of non-periodic signal, the energy distribution is more scattered and the energy concentration occurs at high frequency regions as shown in Figure 4(d). The signal power is mostly localized around the frequencies with highest power. The decision rule is that when the energy peak is located within low frequency region (considerably within the range of fundamental periodicity), the speech segment is considered as periodic i.e. voiced and non-periodic otherwise.

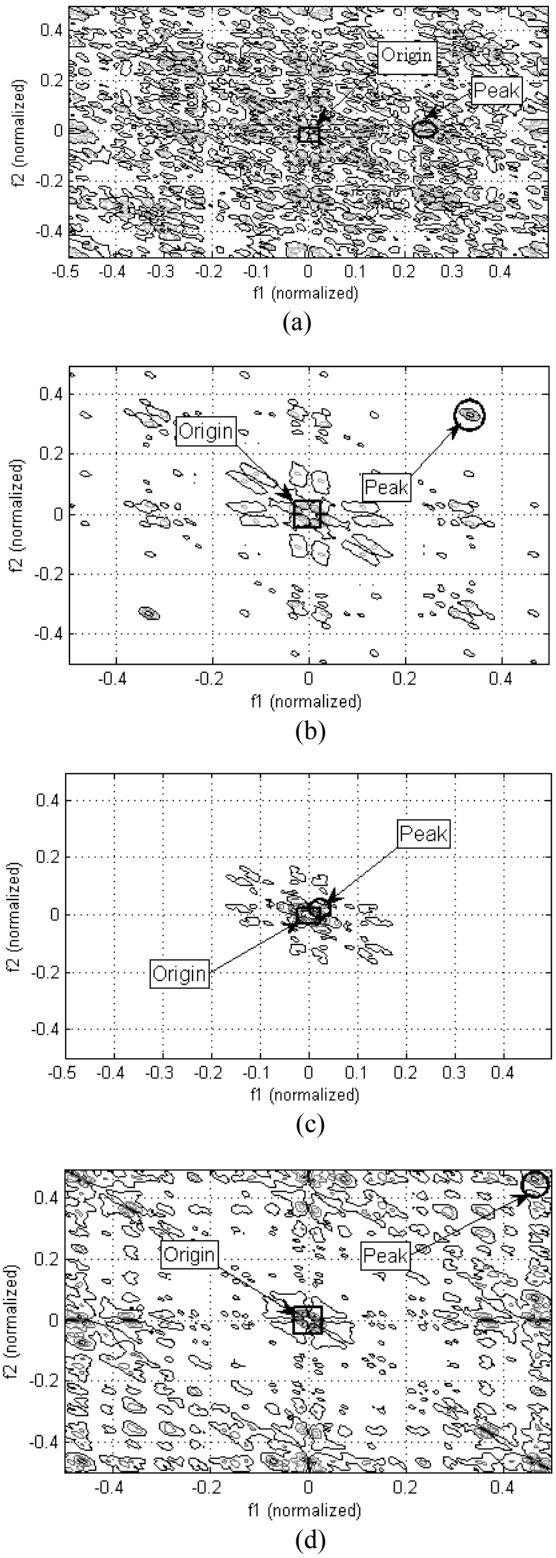

Figure 4. Bispectrums of different stages of speech signal; (a) for periodic noisy speech (0dB SNR), (b) for NACF of the periodic noisy speech (0dB SNR), (c) for prefiltered NACF of that noisy speech and (d) for prefiltered NACF of non-periodic noisy speech (0dB SNR).

IV. Experimental Results

The performance of the proposed method is evaluated by using speech data taken from TIMIT database. The speech material used in this experiment is re-sampled to 20 kHz and segmented into frames of length 30ms with 10ms shifting. 16 bit resolution. Approximately 2010

frames including male and female speech are used. Each frame is accurately labeled for voiced and unvoiced. The error rates are compared for two criterions – with EMD based noise filtering (nfEMD) and without noise filtering (WnF) for different noise levels. The white Gaussian noise is added to obtain different levels of segmental SNR (SSNR). Periodic – to – non-periodic (P-nP) and non-periodic – to – periodic (nP-P) error rates denote the accuracy in correctly classifying P/nP speech frames. A nP-P error occurs when an non-periodic frame is classified erroneously as periodic, and a P-nP error occurs a periodic segment is detected as non-periodic. The overall error rate is obtained by summing up the two error factors. The performance of the proposed method for different SSNRs is shown in Table I.

TABLE I
PERFORMANCE OF THE PROPOSED METHOD WITH *nfEMD* AND *WnF* AS A FUNCTION OF DIFFERENT SSNR

SSNR (dB)	nfEMD (%)			WnF (%)		
	P-nP	nP-P	Overall	P-nP	nP-P	Overall
Clean	0.71	0.41	1.12	1.05	0.84	1.89
10	1.12	0.85	1.97	1.51	1.32	2.83
0	3.22	1.97	5.19	4.97	3.17	8.14
-5	4.93	2.79	7.72	7.19	6.23	13.42
-10	6.11	4.37	10.48	9.14	7.20	16.34

Comparisons with existing methods: The performance of the proposed method is evaluated under noisy conditions for a wide range of segmental SNRs. The overall classification accuracy with EMD based filtering method is always better than the existing reported algorithms [2]-[5]. Although none of those methods demonstrate the performances for SNR less than 0, the comparisons with the proposed algorithm is described here. Using cepstrum-based modified algorithm [2], the overall error is reported as 6.16% for 0dB SSNR. Gaussian mixture model (GMM) with cepstral features is proposed in [3] with 8% error for 15dB SNR. In [4], higher order statistics (HOS) based method is employed in P/nP classification for low SNR (up to 0dB) but no quantitative result is reported. Gabor atomic decomposition method is proposed in [5] with 16% error rate for 5dB SNR speech. Based on the above mentioned performances of the existing algorithms, the proposed method proves its superiority in P/nP classification of noisy speech signals

V. CONCLUSIONS

An efficient P/nP speech classification algorithm for improving robustness in noisy environment is presented in this paper. The voiced speech signal is considered as a time series with periodically correlated process, and, the unvoiced signal does not contain any periodic correlation (PC). EMD based data adaptive and time domain noise filtering method is proposed to increase the robustness of detecting periodic correlation. A higher-order analysis is employed here to formulate the robust P/nP classification rule. As a result, it leads to clear improvements in periodic/non-periodic speech discrimination without any training data especially when the SNR drops. The use of the proposed method in robust pitch (fundamental frequency) detection is the future target.

REFERENCES

[1] B. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced – Unvoiced – Silence Classification with Applications to Speech Recognition", *IEEE Transaction on Acoustics, Speech and Signal Processing*, Vol: 24, No. 3, pp: 201-212, 1976.

[2] S. Ahmadi, and A. S. Spanias, "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.

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

[4] A. Alkulaibi, J. J. Soraghan, and T. S. Durrani, "Fast HOS based simultaneous voiced/unvoiced detection and pitch estimation using 3-level binary speech signals", *Proceedings of 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, pp. 194-197, 1996.

[5] Lobo, and P. Loizou, "Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition", *Proceedings of ICASSP*, pp. 820-823, 2003.

[6] L. Janer, J. J. Bonet and E. L. Solano, "Pitch detection and voiced/unvoiced detection algorithm based on Wavelet transform", in the *Proceedings of ICSLP*, 1996

[7] K. Giridharan, B. Y. Smolenski and R. E. Yantorno, "Statistical And Model Based Approach To Unvoiced Speech Detection", in the *Proceedings of ISPACS*, 2004.

[8] C. L. Nikias and A. P. Petropulu, "Higher-Order Spectra Analysis: A nonlinear signal processing framework", PTR Prentice Hall, 1993.

[9] J. M. Gorriz, J. Ramirez, J. C. Segura and S. Hornillo, "Voice Activity Detection Using Higher Order Statistics", In *Proc. of IWANN 2005*, LNCS 3512, pp: 837-844, 2005.

[10] N. E. Huang et. al., "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis", *Proc. Roy. Soc. London* ***A*****,** Vol. 454, pp. 903-995, 1998.

[11] B. Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method", in the *Proc. Roy. Soc. Lond. A* (460), pp: 1597-1611, 2004.

[12] P. Flandrin, G. Rilling and P. Gonqalves, "Empirical mode decomposition as a filter bank", *IEEE signal processing letters*, Vol. 11, No. 2, pp: 112-114, Feb, 2004.

[13] G. Rilling, P. Flandrin and P. Gonqalves, "On empirical mode decomposition and its algorithms", in the *Proceedings of IEEE-EURASIP Workshop on nonlinear signal and image processing (NSIP)*, 2003.

[14] A. Hanssen and L. L. Scharf, "A theory of polyspectra for nonstationary stochastic processes", *IEEE Transaction on Signal Processing*, Vol. 15, No. 5, pp: 1243-1252,

[15] O. Rioul and M. Vetterli, "Wavelet and Signal Processing", *IEEE Signal Processing Magazine*, pp: 14-38, Oct., 1991.