

Control Limits of Time Series Data using Hilbert-Huang Transform : Dealing with Nested Periods

Jung-Yul Suh · Sae Jae Lee[†]

School of Industrial Engineering, Kumoh National Insitute of Technology

힐버트-황 변환을 이용한 시계열 데이터 관리한계 : 중첩주기의 사례

서정열 · 이세재[†]

금오공과대학교 산업공학부

Real-life time series characteristic data has significant amount of non-stationary components, especially periodic components in nature. Extracting such components has required many ad-hoc techniques with external parameters set by users in a case-by-case manner. In this study, we used Empirical Mode Decomposition Method from Hilbert-Huang Transform to extract them in a systematic manner with least number of ad-hoc parameters set by users. After the periodic components are removed, the remaining time-series data can be analyzed with traditional methods such as ARIMA model. Then we suggest a different way of setting control chart limits for characteristic data with periodic components in addition to ARIMA components.

Keywords : Hilbert-Huang Transform, ARIMA, Control Limits, Periodic Data, Time Series Model

1. Introduction

Real-life time-series characteristic data tend to contain many levels of dynamics operating on different time-scales. That is, time-series can be an aggregation of components, whose rates of change can range from “very fast” to “very slow.” Slow changing ones tend to create a significant level of long-distance autocorrelation. For example in time-series data showing multiple periods of varying lengths, we need more than the autoregressive integrated moving average (ARIMA henceforth) model to help address the additional periodicity problem. Assuming no further complications than

periodicity, in this paper, we discuss how to identify, estimate and extract these periodic components with varying time-scales, and subtract them from the original time-series data. The resulting data can then be modeled with ARIMA (in our case, ARMA). We show how to identify the presence of different time-scales in a time-series.

Once we have a time-series model, a combination of ARIMA model and nested periodic model, then we proceed to determine the control limits of the time-series, i.e. upper control limit (UCL henceforth) and lower control limit (LCL henceforth). While it is straightforward to estimate the control limits for pure ARIMA model, this new combined model requires more qualified approach. It depends on how we interpret the nature of nested time-scale components. That is, would they represent acceptable change or unwanted distortion which needs to be monitored? Based on this inter-

Received 11 August 2014; Finally Revised 10 November 2014;
Accepted 10 November 2014

[†] Corresponding Author : saejalee@kumoh.ac.kr

pretation, different control limits need to be estimated. In this paper, we take the latter approach.

In section 2, survey of previous research on periodic time series analysis is presented including Hilbert-Huang Transform (HHT henceforth). In section 3 we explore how HHT will be applied to extract periodic components. Main advantage of HHT will be discussed. In section 4, an estimation of periods will be presented in detail. In section 5 a new method to set up control boundaries for time-series containing periodic components will be shown. In section 6 the result of applying the new method to real life time-series data will be presented. In section 7, the main contribution is presented along with some suggested directions of future research.

2. Background and Previous Research

Stochastic methods of analyzing time-series data have been extensively used. Most assume that the data set represents stochastically stationary processes in nature. Various control charts were used to monitor if data value deviates away from the baseline of time-series data [12, 14]

As for non-stationary data, there is a notion of seasonality which refers to a periodic trend or fluctuation which frequently appears in time series. They could be well-defined and precise or merely semi-regular. It frequently shows up in economic data as a regular seasonal variation, hence the name ‘seasonality’. However, such periodic fluctuations are common in data from various fields. The typical method to identify seasonality is as follows : (a) Run sequence plot (b) Seasonal subseries plot (c) Multiple box plots (d) Auto-correlation plot (e) Seasonal index measures. These techniques can be useful but they assume that periods of seasonal components are already known to us beforehand. It is not an effective way to find out periods we are not aware of in advance. Besides, it is not systematic and tends to depend on visual inspection by humans.

Fast Fourier Transform (FFT henceforth) is the most common method used for analyzing periodicity. However, FFT is not good at detecting dynamically changing frequency patterns. It assumes that the time-series data maintains the same periodic pattern over the entire time span of interest. To overcome this shortcoming, windowed FFT was developed [8]. Unfortunately, the artifact introduced by the choice of window span creates unwanted distortions in the analysis of periodic patterns.

In order to dissolve this problem, a new approach, Wavelet analysis, was developed [1, 2]. It can show real-time change of frequency patterns. Wavelet assumes that such periodic components can change over time as can be observed in human voice spectrogram. So Wavelet analysis provides, 3-dimensional plot, made up of time, frequency and amplitude. Wavelet resolved many difficulties arising from windowed FFT or other variation of FFT to capture slow change in major period (frequency) components, that is, a spectral spectrum Wavelet algorithm is more efficient than FFT algorithm. While there is some limitation on the frequency resolution in certain frequency band, it poses little problem because real-world data tend to share the same characteristic. With the dynamic frequency tracking, it is possible to detect the characteristics of time-series over time. It can analyze time-series data to detect subtle anomaly in many applications [4, 17].

Least Squares Spectral Analysis (LSSA henceforth) also is a method to address a problem of FFT [16]. FFT generates many spurious frequencies which are not part of any periodic component. It is necessary to identify dominant periodic frequencies if we want to study periodicity of time-series. By using least-square method, it iteratively examines frequency to tell if it is important. Obviously, it is time-consuming, and techniques were devised to narrow down candidate frequencies [11, 13]. While it makes the algorithm faster, the method of choosing candidates tends to be ad-hoc.

Another one, Recurrence period density entropy (RPDE) method, has its origin in dynamical systems theory. Specifically, it is based on a delayed embedding theorem, called Takens’ Theorem [15]. The computation proceeds as follows:

For a time-series $\{x_n\}$, create a time-delayed vector $X_n = [x_n, x_{n+\tau}, \dots, x_{n+(m-1)\tau}]$ where m and τ is called an embedding dimension, and an embedding delay (To find the right pair for m and τ requires search over all possible combinations, which can be computationally expensive.) It counts how many time steps it takes for X_n to return within the predefined boundary of its original starting point. Such time steps T is recorded and a frequency distribution (histogram) can be created according to value of T , and the entropy E of the distribution is calculated, where $E = 0$ if $\{x_n\}$ is periodic, and $E = 1$ if it is uniformly random. This method can measure degree of repetitiveness or periodicity. It is robust and does not require restrictive assumptions such as linearity or normal distribution. It is shown to be able to detect subtle abnormality in biological signals [10]. However, as

has been pointed out earlier, critical weakness is the difficulty in finding the right value of m and τ .

- (1) Find all local maxima for entire x_t , and compute their envelope curve using cubic-spline lines.
- (2) Do the same for local minima.
- (3) Find the mean curve $m_t^{(1)}$ from the two envelope curves.
- (4) Define $h_t^{(1)} = x_t - m_t^{(1)}$
- (5) Do the same with $h_t^{(1)}$ in place of x_t . Then, $h_t^{(11)} = h_t^{(1)} - m_t^{(11)}$
- (6) For $k \geq 2$, do (5) with $h_t^{(1k)}$ in place of $h_t^{(1)}$. Then, $h_t^{(1k)} = h_t^{(1(k-1))} - m_t^{(1k)}$
- (7) Repeat (6) until $h_t^{(1(k-1))}$ and $h_t^{(1k)}$ are close enough according to a predetermined convergence criteria.
- (8) Define $c_t^{(1)} = h_t^{(k)}$ and $r_t^{(1)} = x_t - c_t^{(1)}$
- (9) For $k \geq 2$, take $r_t^{(k)}$ in place of x_t and repeat (1)~(8). Then $r_t^{(k)} = r_t^{(k-1)} - c_t^{(k)}$
- (10) When $c_t^{(k)}$ becomes a monotone function which can generate no more IMF, the process ends.

$$x_t = \sum_{k=1}^n c_t^{(k)} + r_t^{(n)} \text{ and } \{c_t^{(1)}, \dots, c_t^{(n)}, r_t^{(n)}\}$$

make up intrinsic mode functions(IMF,) each representing a different time-scale.

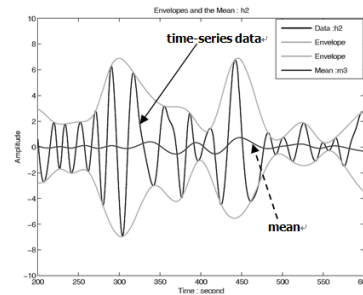
<Figure 1> Empirical Mode Decomposition

All these methods require some ad-hoc techniques to make it work. Essentially, it goes down to setting up appropriate filters (e.g FFT, windowed FFT, and Wavelet) or frequency bands (e.g. LSSA) whose parameter values are determined in ad-hoc fashion.

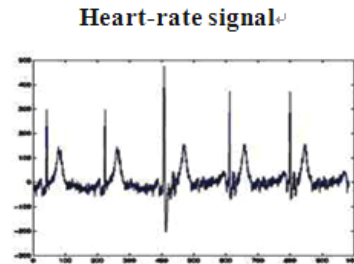
The Hilbert-Huang transform [3, 5, 6, 7, 9] is a method which can analyze time series data and split it into multiple oscillatory components, called Intrinsic Mode Functions (IMF henceforth) The result is superior to any comparable techniques, especially for non-stationary and nonlinear time series. Oscillations in time series frequently exhibit intra-wave frequency modulation, which does not conform to typical sinusoidal shape. This is typical of oscillatory behavior from highly non-linear systems. Fourier transform tends to destroy essential feature of such pattern, while HHT can capture it rather effectively. On the other hand, unlike Fourier analysis, HHT was not derived from pre-existing mathema-

tical theory. It is more of an empirically conceived algorithm, whose effectiveness has been shown by applying it to time series data from various fields.

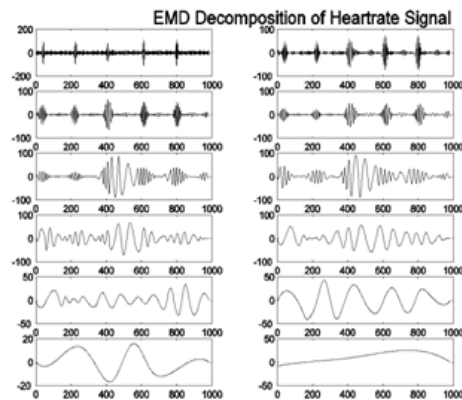
It makes use of iterative procedure in which the component (IMF) of highest frequency was computed and subtracted out and proceeds to isolate the components (IMF) of lower and lower frequency. The procedure is called Empirical Mode Decomposition (EMD henceforth) method. The rough description of HHT is in <Figure 1>. The result of EMD is shown in <Figure 2>. The left side of <Figure 2> shows how steps (1) ~ (4) are done. EMD takes the heart rate curve as shown in <Figure 2>(b) and produces IMF's in <Figure 2>(c). x_t is now decomposed into 12 empirical modes.



(a) Empirical Mode Decomposition : Envelop Curve and Mean Curve



(b) Time-Series Data of Heart Pulse



(c) EMD Decomposition of Heart Pulse Data

<Figure 2> EMD Illustration and Example

3. Determining Optimal Periodic Component

Since we are not interested in dynamic change of frequency distribution in this paper, it is assumed that the periodic component does not change over time. We only use EMD to extract dominant periodic component in a clean and efficient manner, while we forgo other more powerful features of HHT such as analysis via the chart of instantaneous amplitude and frequency. This is, in a way, a grossly limited use of HHT. However, it is still quite effective for our purpose.

We will combine LSSA, FFT, and HHT to extract a periodic component underlying the time series. There are two criteria in selecting these methods : a) the process should be efficient (not time-consuming) b) the introduction of rather subjective external parameters should be kept to a minimum.

HHT can separate time-series data into components of different time-scales efficiently. The concept of time-scale components is more general than that of periodic components, because it can address non-periodic oscillatory components as well. One of the most appealing features of HHT is that it does not need many parameter values set externally in an ad hoc fashion, except for two cases: the convergence threshold for sifting operation and terminating further component extraction. The real-life time-series often do not have components of exact periods. However, it could be also the case that periodic components are obscured due to random fluctuations in the time-series data. In the latter case, HHT produces an oscillatory component which may not give good indication of what its period is by merely counting peaks, valleys or zero crossings. A typical case is the daily temperature changes. They are expected to have a period of 24 hours, considering earth's rotation. However, daily highs and lows are usually different from day to day, due to weather changes. The pattern of rise and fall during the day also differs from day to day. In this case, we need to subject the result from HHT to further process in order to find what the dominant period is. If there is a discernible periodic pattern, we want to extract it.

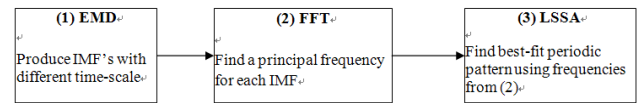
By incorporating features of LSSA, FFT, and HHT discussed above, we can formulate a four step process which can produce best-fit periodic component for a given time series.

- 1) Perform HHT on the time series data.
- 2) For each time-scale components HHT generated, perform

FFT respectively. Find the most dominant period (one with maximal amplitude) in FFT chart.

- 3) Make a list of periods consisting of dominant periods from 2) and their companion periods, the definition of which is to be addressed later. Perform multi-linear regression on the original time series with a linear equation made from the period list from 3).

This procedure is summarized in <Figure 3>.



<Figure 3> Extracting Periodic Pattern

4. Two-Stage Estimation Process for Analysis

The model suggested here assumes that a time-series $\{x_t\}$ is defined as follows :

$$x_t = \hat{x}_t + y_t,$$

where \hat{x}_t is a deterministic periodic function of t and y_t is a stochastic time-series of ARMA (autoregressive moving average) model.

We propose that the estimation process be divided into two stages for a time series $\{x_t\}$:

- Stage 1) Extracting a periodic component
- a) Perform HHT/FFT on $\{x_t\}$ and select periods with prominent amplitude.
 - a.1) Apply EMD on $\{x_t\}$ and find its IMF's $c_t^{(1)}, c_t^{(2)}, \dots, c_t^{(i)}, r_t^{(1)}$.
 - a.2) Apply FFT on each $c_t^{(i)}$ and find frequency \hat{f}_i with maximal peak amplitude in its FFT chart.
 - a.3) Find companion frequencies of \hat{f}_i , which are naturally associated frequency multiples of $1/6, 1/4, 1/3, 1/2, 1, 2, 3, 4, 6$ of it. In nature these ratios cover most of the cases. Companion frequencies of a maximal peak amplitude frequency commonly show up together in FFT chart. They account for significant portion of periodic component, too. We

also add some low frequencies in the list (from 1 to 10) to account for strong long-term non-linear non-oscillatory trend. Let $F = \{f_1, f_2, \dots, f_p\}$ be the set made up of all such frequencies.

- b) Perform multiple regression on $\{x_t\}$ with the following equation : $\hat{x}_t = c + \sum_{i=1}^p (a_i \cos 2\pi f_i t + b_i \sin 2\pi f_i t)$ Let $y_t = x_t - \hat{x}_t$ With a periodic component \hat{x}_t removed from the original time series value x_t , resulting new time series $\{y_t\}$ is now more amenable to ARMA model estimation method.'

Stage 2) Check $\{y_t\}$:

$\{y_t\}$ is now a time-series of ARMA. Data from such a time-series form a normal distribution. Then conventional 3σ control limits can be used to locate out-of-control points. One can argue whether or not $\{y_t\}$ always turns out to be ARMA time-series. However, in our model, we assume that it would be the case.

5. Determining Control Limits

Since we decided to regard $\{y_t\}$ as ARMA model time series, we get estimated mean from $\{y_t\}$ as the baseline of our control limits. We define $UCL = \mu_y + 3s_y$, $LCL = \mu_y - 3s_y$ where s_y is the sample standard deviation of $\{y_t\}$ (Note : $x_t = y_t + \hat{x}_t$). To set up the control limits for $\{x_t\}$, we will add contributions from \hat{x}_t to the control limits for $\{y_t\}$. The question is how to do it. For \hat{x}_t , we need to have some way to measure its deviation away from its mean. Since \hat{x}_t is periodic, its values would not make up a normal distribution. They may be distributed more evenly than values from normal distribution which tends to concentrate more around its mean. Therefore, the better measure of deviation could be average deviation $a_{\hat{x}}$, which tends to discount the contribution of extreme values compared with standard deviation. It can be defined as $a_{\hat{x}} = \frac{1}{T} \sum_{t=1}^T |\hat{x}_t - \bar{\hat{x}}|$, where $\bar{\hat{x}}$ is the mean of $\{\hat{x}_t\}$ (Note : $a_{\hat{x}} \leq s_{\hat{x}}$, which can be proved by a simple application of Cauchy-Schwartz Inequality).

For linear component, we won't measure its deviation in any way. Linear component would be directly incorporated into control limits of x_t . In our case, $\bar{\hat{x}}$ and μ_y are close to 0. Then,

$$UCL = Aa_{\hat{x}} + Ds_y, \quad LCL = -(Aa_{\hat{x}} + Ds_y)$$

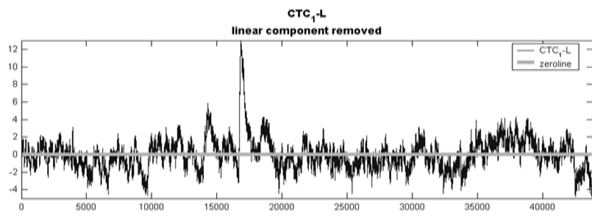
(A and D are positive real numbers. Most of the time $D = 3$. As for A, it may vary depending on the situation. In our case, let us assume $A = 2$, since it covers most of the normal cases.)

Finally, there is some room for debate regarding the nature of periodic components. In our study, we regard periodic variation as something undesirable, hence we included its average deviation $a_{\hat{x}}$ in setting up the control limits of (original) time series $\{x_t\}$. However, in some cases, we could regard periodic variation as acceptable change. If that is the case, we only need to set up the control limits for $\{y_t\}$ to detect event outside control limits.

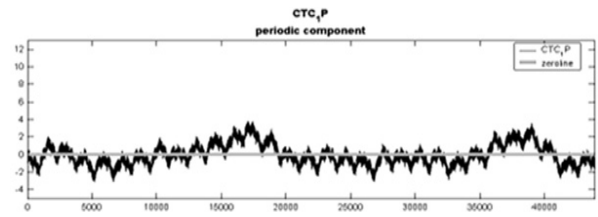
6. Results and Discussion

We now investigate actual data, that are temperature readings measured from 9 different locations of a glass furnace. Each location produces time series data of temperature readings sampled at one minute interval spanning an entire month. There are two features of the times series we need to pay attention to. First, the temperature measurement is subject to steadily widening bias. It is measured by a thermometer embedded inside one of heat-resisting bricks which make up the wall of the furnace. The brick is subject to steady erosion by molten glass (silicon) which comes in contact with the brick. As time goes by, the brick becomes thinner and the thermometer is exposed to more heat. Naturally, even if the temperature of the furnace stays constant, the reading would become higher. Time-series data set we use is the one after we factored out this trend from original time-series, which happens to be linear. Second, the temperature reading is also subject to periodic changes. Two sources of periodic pattern are already known. The first one arises from the design of the furnace. It reverses the flow every 20 minutes, resulting in corresponding temperature change with the same period. The other is due to nature. Temperature goes up during the day and down at night. It affects the temperature inside glass furnace, which also changes in the same direction. After linear component is removed, the time-series data would look like the one (CTC1-L.) in <Figure 4(a)>.

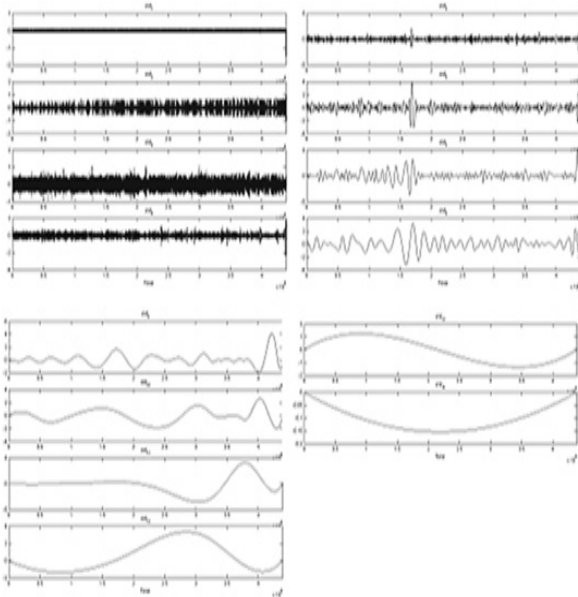
Applying EMD to the time-series, we have 14 IMF's as shown in <Figure 4>(b). EMD neatly decomposes the time-



(a) The Time-Series Data to be Analyzed



<Figure 5> Best-Fit Periodic Component



(b) their IMF's after EMD; upper left : IMF01~IMF04, upper right : IMF05~IMF08, lower left : IMF09~IMF12, lower right : IMF13, IMF14

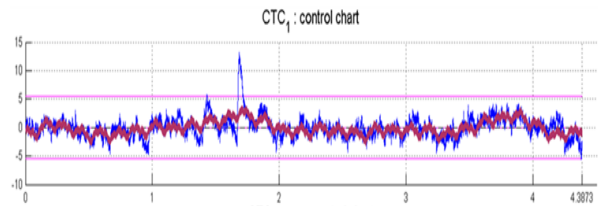
<Figure 4> Extracting Periodic Pattern

series CTC1-L into essential time-scale components without introducing many ad hoc filtering and other rather subjective mechanism.

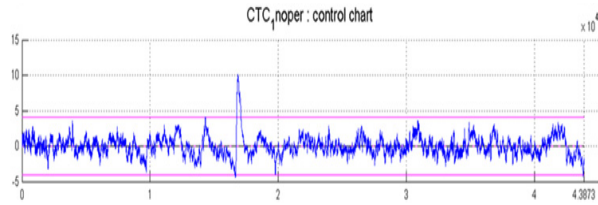
Now to find a dominant frequency (period) of each IMF, we apply FFT to it. The resulting FFT is more localized and simpler than the FFT of entire time-series. We select the frequency (period) with maximum amplitude.

After frequency list is generated, we run regression on linear sum of sinusoids made from them. The best fit (CTC1P) is shown in <Figure 5>. The remainder (CTC1-LP) is in the bottom, which is subject to ARMA analysis. Average deviation of CTC1P and standard deviation of CTC1-LP are calculated and used to set up CTC1's control limits.

Average deviation (adCTC1P) of periodic component CTC1P is 0.9557 Standard deviation (stdCTC1-LP) of ARMA component CTC1-LP is 1.37. $D_{sy} = 4.11$ and $D_{sy} + A a_{\hat{x}} = 6.0214$ (Note : $D = 3, A = 2$). The result can be shown in <Figure 6>.



(a) original time-series CTC1-L and its control limits



(b) ARMA component CTC1-LP and its control limits

<Figure 6> Control Limits

7. Conclusion

In this research, we investigated ways to handle linear and periodic components of a time series, which we often encounter in time series data obtained from the real world. First, we extract linear trend and periodic components from original time series $\{x_t\}$. Periods are extracted using HHT and FFT, obtaining ARMA time series $\{t_t\}$. Using HHT improved the way to find multiple prominent periods in time-series data. Ad hoc deployment of various filters and threshold are no longer necessary. Then we set up the new kind of control limits incorporating average deviation of periodic component $a_{\hat{x}}$.

This study might be examined with some further investigations using more cases of time series. The proper values for UCL and LCL are those which need further empirical investigations. Applying this method to real-world time-series data, detailed discussion on the results would take up too much additional space. It will be described in some future publication on this topic.

Here we only used Empirical Mode Decomposition method of HHT for finding meaningful prominent periods. The

other part, applying Hilbert Transform on IMF's, is another analytical tool we can use. The degree of periodicity of each IMF could be also investigated. This measure can serve as an effective way to characterize time-series. All these further efforts would help better manage and control facilities to be managed.

Acknowledgement

This study has been partially supported by a Research Fund of Kumoh National Institute of Technology, Korea.

References

- [1] Chui, C.K., *An introduction to Wavelets*. San Diego : Academic Press, 1992.
- [2] Daubechies, I., *Ten lectures on Wavelets*, Society for Industrial and Applied Mathematics, 1992.
- [3] Flandrin, P., Rilling, G., and Goncalves, P., Empirical mode decomposition as a filter bank. *IEEE Signal Proc Lett.*, 2003, Vol. 11, p 112-114.
- [4] Huang, C., Sachin T., and Shin, Y.J., *Wavelet-based Real Time Detection of Network Traffic Anomalies*, Securecomm and Workshops, 2006, p 1-7.
- [5] Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N., Tung, C., and Liu, H., The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London*, 1998, A454, p 903-995.
- [6] Huang, N. and Attoh-Okine, N.O., *The Hilbert-Huang transform in engineering*, Taylor and Francis, 2005.
- [7] Huang, N.E. and Shen, S.P., *Hilbert-Huang transform and its applications*, London : World Scientific, 2005.
- [8] Jacobsen, E. and Lyons, R., The sliding DFT. *Signal Processing Magazine*, 2003, Vol. 20, No. 2, p 74-80.
- [9] Lee, S.J. and Suh, J.Y., Evaluating Efficacy of Hilbert-Huang Transform in Analyzing Manufacturing Time Series Data with Periodic Components. *J. Soc. Korea Ind. Syst. Eng.*, 2012, Vol. 35, No. 2, p 106-112.
- [10] Little, M., McSharry, P., Roberts, S., Costello, D., and Moroz, I., Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *Biomed Eng Online*, 2007, Vol. 6, No. 1, p 23.
- [11] Lomb, N.R., Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 1976, Vol. 39, p 447-462.
- [12] Reynolds, Jr., Marion R., and Lu, C., Control charts for monitoring processes with autocorrelated data. *Non-linear Analysis, Theory, Methods and Applications*, 1997, Vol. 30, No. 7, p 4059-4067.
- [13] Scargle, J.D., Studies in astronomical time series analysis II : Statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal*, 1982, Vol. 263, p 835-853.
- [14] Sheu, S.H., Huang, C.J., and Hsu, T.S., Extended maximum generally weighted moving average control chart for monitoring process mean and variability. *Computers and Industrial Engineering*, 2012, Vol. 62, p 216-225.
- [15] Stark, J., Broomhead, D., Davies, M., and Huke, J., Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis-Theory Methods and Applications*, 1997, Vol. 30, No. 8, p 5303-5314.
- [16] Vanicek, P., Approximate Spectral Analysis by Least-squares Fit. *Astrophysics and Space Science*, 1969, Vol. 4, p 387-391.
- [17] Wei, L. and Ghorbani, A.A., Network Anomaly Detection Based on Wavelet Analysis. *EURASIP Journal on Advances in Signal Processing*, 2009, p 16.