ON SELECTING RELEVANT INTRINSIC MODE FUNCTIONS IN EMPIRICAL MODE DECOMPOSITION: AN ENERGY-BASED APPROACH

Douglas Baptista de Souza[†], *Jocelyn Chanussot*[†,§], *Anne-Catherine Favre*[‡]

[†]GIPSA-Lab, [‡]LTHE, Grenoble INP, Domaine Universitaire, Saint Martin d'Hères - FRANCE
[§] Faculty of Electrical and Computer Engineering, University of Iceland

ABSTRACT

Although the empirical mode decomposition is a powerful tool for analyzing complicated datasets, many irrelevant intrinsic mode functions may appear in the decomposition. In this paper, we develop an energy-based method to detect relevant intrinsic mode functions. The new method can be seen as a generalization of techniques that are based on correlation. An experimental study is carried out in different datasets for assessing the performance of the proposed technique.

Index Terms— empirical mode decomposition, intrinsic mode function, correlation, energy, mutual information

1. INTRODUCTION

Traditionally, data analysis has been performed by Fourier-based methods, which requires linearity and stationarity. Even though many processes can be approximated by linear systems, getting rid of the nonlinearity and nonstationarity is often impossible. This problem has been recognized long ago, and the analysis of nonlinear and nonstationary signals has been commonly carried out by techniques such as wavelet transform [1][2], multitaper spectrogram [3][4] or singular spectrum analysis [5][6]. These methods, although widely used, face some problems regarding locality and adaptivity. These conditions should be satisfied for capturing the local variations of nonlinear and nonstationary data. An alternative that fulfills the locality and adaptivity requirements is given by the Empirical Mode Decomposition (EMD) [7], a data-adaptive method that has been extensively used in the past ten years in a multitude of applications [8][9][10].

The EMD is a powerful algorithm to decompose signals into zero-mean amplitude and frequency modulated components called intrinsic mode functions (IMFs), which represent the local oscillations within the signal [11]. The EMD has proved to outperform traditional methods in the analysis of nonlinear and nonstationary data [12][7].

Despite of its success, the EMD is a numerical procedure that is susceptible to errors. These erros may appear as extra, undesirable IMFs [11]. Hence, for obtaining a more reliable representation of the signal when using EMD, some techniques have been proposed to identify relevant IMFs. In [13], the authors assigned the statistical significance of the IMFs by means of numerical experiments. This idea, however, is applicable only to white noise signals. In a different setting, the authors in [14] used information analysis to select IMFs that retain most of the information content of the signal. A simpler idea, on the other hand, has been adopted by [15] and [16], which considered that relevant IMFs would have a good correlation with the original signal, while irrelevant ones would have a poor correlation. Thus, a threshold τ can be set to discriminate between relevant and irrelevant IMFs. In this paper, we propose a new method which does not assume the specification of a single threshold for selecting relevant IMFs. The new algorithm is based on the energies of the IMFs and generalizes the correlation-based procedures of [15] and [16].

This paper is organized as follows. In section 2, we briefly explain the EMD technique. In section 3, we generalize the correlation criterion for selecting relevant IMFs to an energy-based framework, from which we develop a novel algorithm that is explained in section 4. The experimental study and the conclusions are shown in section 5 and 6, respectively.

2. EMD BASICS

The EMD method decomposes the signal into a superposition of oscillatory, signal-dependent functions: the IMFs. The IMFs are computed iteratively and must satisfy two conditions regarding the number of zero crossing versus extrema and the mean values of local envelopes [17]. The algorithm to extract the IMFs from a given time series $x(t)$ is called *sifting process* [7], which can be described as follows [11]:

i) identify all the local extrema in the signal $x(t)$,
ii) interpolate all the local maxima and minima by a cubic spline to produce an upper envelope $e_{\text{up}}(t)$ and a lower $e_{\text{lo}}(t)$ envelope, respectively,
iii) compute the mean $\rho(t) = [e_{\text{up}}(t) - e_{\text{lo}}(t)]/2$,
iv) compute the detail $m(t) = x(t) - \rho(t)$,
v) repeat steps i to iv until the detail $m(t)$ can be considered to be a zero-mean signal according to some stopping criterion [7][11]. If so, $m(t)$ is called an IMF and the procedure continues by iterating on the residual $\rho(t)$.

The algorithm above stops when the slowly-varying residual function has no more oscillations. This last residual is then

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

represented by $\rho_I(t)$. By applying the procedure above, we obtain a collection of I IMFs $\{m_i(t), i = 1, ..., I\}$ plus $\rho_I(t)$ as a representation of the the time series $x(t)$, i.e.:

$$x(t) = \sum_{i=1}^{I} m_i(t) + \rho_I(t), \quad (1)$$

where the IMFs ranging from $m_1(t)$ to $m_I(t)$ represent local oscillations going from the shortest period $(m_1(t))$ to the longest one $(m_I(t))$ [18]. Finally, the IMFs have two important properties that are invoked several times in this paper:

I each IMF is a zero-mean function by construction [17],

II for all practical purposes, the consecutive IMFs can be considered to be locally orthogonal to each other [7].

3. GENERALIZING CORRELATION-BASED METHODS TO AN ENERGY-BASED APPROACH

For discriminating between relevant and irrelevant IMFs, the following threshold has been proposed by [15]:

$$\tau = \frac{\max(\boldsymbol{\rho})}{10}, \quad \text{where} \quad \boldsymbol{\rho} = [\rho_{m_1,x}, ..., \rho_{m_I,x}]. \quad (2)$$

In (2), $\rho_{m_i,x}$ stands for the Person's correlation coefficient between the i^{th} IMF and the original signal $x(t)$, i.e.:

$$\rho_{m_i,x} = \frac{\mathbb{C}\text{ov}\{m_i(t), x(t)\}}{\sigma_{m_i}\sigma_x}, \quad \text{for } i = 1, ..., I. \quad (3)$$

where σ_{m_i} and σ_x are, respectively, the standard deviations of the i^{th} IMF $(m^{(i)}(t))$ and the original signal $(x(t))$. Given (2) and (3), the criterion for selecting relevant IMFs is then expressed as follows: If $\rho_{m_i x} > \tau$ keep the i^{th} IMF. Otherwise, eliminate the i^{th} IMF and add it to the residual. In [16], a modified threshold for selecting IMFs has been proposed, while keeping the same idea of the criterion above. The new threshold τ' is more stringent and works better for noisy signals. We can express τ' in terms of τ as follows: $\tau' = 10\tau/(100\tau - 3)$. The idea behind [15] and [16] is similar, in the sense that the significance of the IMFs is assessed by setting up a threshold for the correlation. In this section, we show that this equivalent to thresholding the most energetic IMFs. In the next section, we propose a more general procedure for selecting relevant IMFs.

If we consider the signal $x(t)$ under the EMD notation (see (1)), it can be shown that $\mathbb{C}\text{ov}\{m_i(t), x(t)\}$ can be expressed as follows (due to the properties of covariance[1]):

$$\mathbb{C}\text{ov}\{m_i(t), x(t)\} = \mathbb{C}\text{ov}\left\{m_i(t), \left[\sum_{i=1}^{I} m^{(i)}(t) + \rho^I(t)\right]\right\} = \sigma^2_{m_i} + \sum_{i\neq j}^{I} \mathbb{C}\text{ov}\Big\{m_i(t), m_j(t)\Big\} + \mathbb{C}\text{ov}\{m_i(t), \rho_I(t)\}. \quad (4)$$

[1] If $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$ are real-valued random variables, then $\mathbb{C}\text{ov}\left\{\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i\right\} = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{C}\text{ov}\{x_i, y_j\}$ and $\mathbb{C}\text{ov}\{x_i, x_i\} = \sigma^2_{m_i}$.

The rationale behind [15] and [16] can be simply expressed as follows: the i^{th} IMF will be relevant if $\rho_{m_i,x} > \widetilde{\tau}$, where $\widetilde{\tau}$ is the an arbitrary threshold. By considering (3) and (4), we can express the condition for $\rho_{m_i,x}$ as follows:

$$\frac{\sigma^2_{m_i} + \sum_{i\neq j}^{I} \mathbb{C}\text{ov}\{m_i(t), m_j(t)\} + \mathbb{C}\text{ov}\{m_i(t), \rho_I(t)\}}{\sigma_{m_i}\sigma_x} > \widetilde{\tau}. \quad (5)$$

If we let $x'(t) = x(t) - \mu_x$, where μ_x is the mean of the underlying signal, we have $\sigma_{x'} = \sigma_x$, as the standard deviation is invariant to changes in location. Also, due to the property **I** of the IMFs, we have $x'(t) \approx x(t)$, and the only difference will be a slightly shifted residual (which can be considered to be nearly constant in comparison to other IMFs). By considering $x'(t)$ and $\sigma_{x'}$, (5) can be rearranged as:

$$\sum_{i\neq j}^{I} \mathbb{C}\text{ov}\{m_i(t), m_j(t)\} + \mathbb{C}\text{ov}\{m_i(t), \rho_I(t)\} > \sigma_{m_i}(\widetilde{\tau}\sigma_{x'} + \sigma_{m_i}). \quad (6)$$

The computation of (6) is carried out by using the sample covariance $(\widehat{\gamma}_{i,j})$ and the sample variance $(\widehat{\sigma}^2_{m_i})$:

$$\widehat{\gamma}_{i,j} = \frac{1}{N-1}\sum_{t=1}^{N}\Big[m_i(t) - \widehat{\mu}_i\Big]\Big[m_j(t) - \widehat{\mu}_j\Big] = \frac{1}{N-1}\sum_{t=1}^{N} m_i(t)m_j(t),$$
$$\widehat{\sigma}^2_{m_i} = \frac{1}{N-1}\sum_{t=1}^{N}\Big[m_i(t) - \widehat{\mu}_i\Big]^2 = \frac{1}{N-1}\sum_{t=1}^{N}\Big[m_i(t)\Big]^2, \quad (7)$$

where N is the length of the signal and $\widehat{\mu}_i = \widehat{\mu}_j = 0$ due to the property **I** of the IMFs. Note that, if we define the energies the i^{th} IMF and $x'(t)$ as $E_i = \sum_{t=1}^{N}[m_i(t)]^2$ and $E_{x'} = \sum_{t=1}^{N}[x'(t)]^2$, we have $\widehat{\sigma}^2_{m_i} = E_i$ and $\widehat{\sigma}^2_{x'} = E_{x'}$. Thus, the condition in (6) can be given as follows:

$$\sum_{i\neq j}^{I} \widehat{\gamma}_{i,j} + \widehat{\gamma}_{i,\rho} > \sqrt{E_i}(\widetilde{\tau}\sqrt{E_{x'}} - \sqrt{E_i}). \quad (8)$$

Due to the property **II** of the IMFs, the term $\sum_{i\neq j}^{I} \widehat{\gamma}_{i,j}$ in (8) should be close to zero. Moreover, since the residual $\rho_I(t)$ is nearly a constant in comparison to the IMFs, the value of $\widehat{\gamma}_{i,\rho}$ should also be close to zero. Hence, in practice, we could replace the left-hand side of (8) by a given variable ϵ, where the more efficient the sifting process is, the more $\epsilon \to 0$. Thus, the condition in (6) could be expressed simply as:

$$\sqrt{E_i}(\widetilde{\tau}\sqrt{E_{x'}} - \sqrt{E_i}) < \epsilon, \quad (9)$$

where, the greater the value of the threshold $\widetilde{\tau}$, the more energetic a given IMF should be in order to attend the condition in (9). Note that, at the end, the threshold $\widetilde{\tau}$ only controls the *fraction of the energy* corresponding to the whole data $(\sqrt{E_{x'}})$, in which $\sqrt{E_i}$ should be greater to in order to make the i^{th} IMF relevant. In other words, correlation-based methods such as [15] and [16] can be seen as energy-based ones.

4. A NEW CRITERION FOR SELECTING RELEVANT IMFS BASED ON THEIR ENERGIES

As $\widetilde{\tau}$ is only a *fraction* of the total energy, we propose not to consider a single value for $\widetilde{\tau}$ for selecting or not a given IMF, but rather to consider as many values of $\widetilde{\tau}$ as possible, ranging, for instance, from 0.01 to 0.99. We represent this range in the vector $\boldsymbol{\tau} = [\widetilde{\tau}_1, ..., \widetilde{\tau}_f]$, where f is the length of the vector. For $\boldsymbol{\tau}$, the spacing ($\Delta_\tau = \widetilde{\tau}_l - \widetilde{\tau}_{l-1}$), the initial value ($\widetilde{\tau}_1$), and the final one ($\widetilde{\tau}_f$) are set as $\Delta_\tau = \widetilde{\tau}_1 = 0.01$ and $\widetilde{\tau}_f = 0.99$, as default[2]. Then, for each $\widetilde{\tau}_l \in \boldsymbol{\tau}$, we gather the indices of the IMFs whose energies attend (9), and put it in an array $\mathbf{m}_l$, *called a mode candidate*. By repeating the procedure for all $\widetilde{\tau}_l \in \boldsymbol{\tau}$, we obtain a collection $\mathcal{M} = [\mathbf{m}_1, ..., \mathbf{m}_f]$ of mode candidates. Then we analyze all entries of $\mathcal{M}$ to separate and count the mode candidates that are equal, i.e., the mode candidates that refer to the same group of IMFs. Finally, we define the most significant IMFs as those contained in the mode candidate that appeared most often in $\mathcal{M}$. The proposed algorithm is described step by step as follows:

i) for $\boldsymbol{\tau} = [\widetilde{\tau}_1, ..., \widetilde{\tau}_f]$, set $\widetilde{\tau}_1 = \Delta_\tau = 0.01$ and $\widetilde{\tau}_f = 0.99$,

ii) apply the EMD to the signal $x(t)$ to obtain a collection of IMFs $(m_1(t), ..., m_I(t))$ plus the residual $(\rho_I(t))$,

iii) compute an array $\mathbf{e} = [E_1, ..., E_I, E_\rho]$ formed by the energies of the IMFs and the residual, compute $\epsilon = \widehat{\gamma}_{i,j} + \widehat{\gamma}_{i,\rho}$ by means of (7). Set $l = 1$,

iv) for $\widetilde{\tau} = \widetilde{\tau}_l$, where $\widetilde{\tau}_l \in \boldsymbol{\tau}$, check which $E_i \in \mathbf{e}$ attend the condition: $\sqrt{E_i}(\widetilde{\tau}\sqrt{E_{x'}} - \sqrt{E_i}) < \epsilon$,

v) collect the indices of all $E_i \in \mathbf{e}$ that attended the condition in step vi, and create an array $\mathbf{m}_l$,

vi) increase l to $l + 1$ and go back to step iv. If $\widetilde{\tau}_l = \widetilde{\tau}_f$, group all $\mathbf{m}_l, ..., \mathbf{m}_f$ in $\mathcal{M}$ and go to the next step,

vii) Count the repeated modes in $\mathcal{M}$. The IMFs contained in the most frequent mode are the most relevant IMFs.

Different from other techniques, our approach focuses on a criterion for selecting a *group* of relevant IMFs, not a given IMF individually. Having presented a new approach for selecting relevant IMFs, we present in the next section the results of applying the proposed algorithm to different data sets.

5. EXPERIMENTAL STUDY

In this section, we compare the developed method with the one proposed in [16]. For evaluating how efficiently the chosen IMFs capture key characteristics of the signal, we propose to measure how much of the information content of the original signal is retained in the IMFs selected by each approach. To do so, we compute the *mutual information* between the selected IMFs and the original signal. For two discrete random variables X and Y, the mutual information is given as:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left[\frac{p(x,y)}{p(x)p(y)} \right],$$

where $p(x)$, $p(y)$ and $p(x,y)$ are, respectively, the marginal pdfs and the joint pdf of X and Y evaluated at x and y. In this work, we have estimated the mutual information by using the method described in [19]. The experimental study was performed in three different types of signals: i) a signal composed of two sinusoids embedded in white noise, ii) a signal composed of two sinusoidal FM components and one Gaussian wavepacket, and iii) a real world signal.

5.1. First case: two sinusoids embedded in white noise

In this case, we have tested $x(t) = \sin(3\pi t) + \sin(5\pi t) +$ WGN(0,1), which is one of the signals tested in [16]. Sinusoids embedded in background noise are commonly used for evaluating the performance of the EMD, as without the noise we would expect to have each component of the signal represented by a single IMF. In Table 1, we show the selected IMFs and the corresponding mutual information with the original signal, for the correlation-based approach of [16] and the new one. Note that the new method is more stringent and selects IMFs giving the highest values of mutual information. In Fig. 1, we show the IMFs, the original signal, and the one obtained by summing up the relevant IMFs selected by the new method. Notice the resemblance between the original signal and the one reconstructed with the chosen IMFs.

Table 1. For the first example, mutual information (MI) between the IMFs selected as relevant and the original signal

New approach						
IMFs	1	2	8	res.		
MI	**0.186**	**0.073**	**0.166**	**0.258**		
Correlation-based approach						
IMFs	1	2	3	7	8	res.
MI	0.186	0.073	0.049	0.026	0.166	0.258

5.2. Second case: two sinusoidal FM components and one Gaussian wavepacket

Originally present in [11], this signal is composed of three components: two sinusoidal FM waveforms and a Gaussian logon. These three components overlap significantly in time and frequency, which forbids the components to be decomposed by any non-adaptive filtering method [11]. Such behavior can be seen in Fig. 2, where all the IMFs and the time-frequency (TF) representations (estimated by multitaper spectrogram [4]) of the first three IMFs are shown. Notice that the three TF signatures that are overlapped for the original signal are successfully separated in the first three IMFs. In Table 2, we show that the new method and the one proposed in [16] selected the first three IMFs as the most relevant ones.

[2]One could use even a finer resolution (i.e. $\Delta_\tau < 0.01$), but we have observed that it does not affect the final results at all.

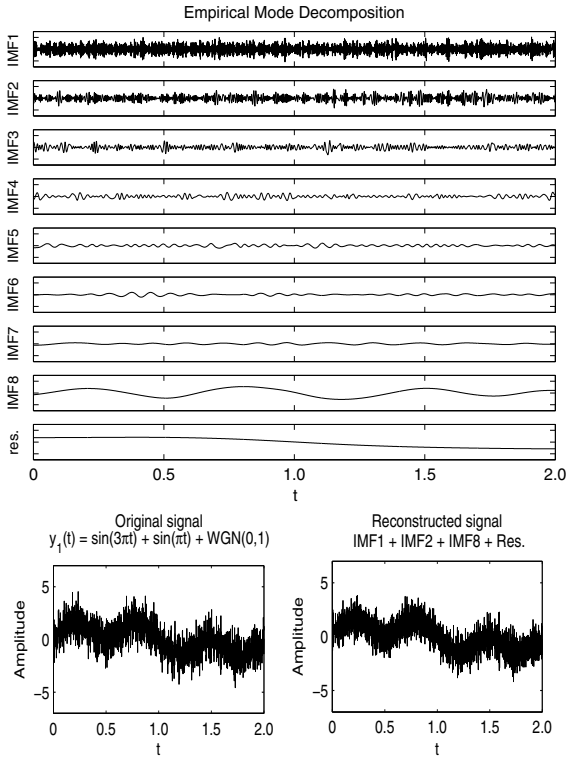

Fig. 1. For the first test signal, the IMFs returned by the EMD together with the original and reconstructed signals.

Table 2. For the second example, mutual information (MI) between the IMFs selected as relevant and the original signal

New approach			
Selected IMFs	1	2	3
MI	**0.745**	**0.945**	**0.148**
Correlation-based approach			
Selected IMFs	1	2	3
MI	0.745	0.945	0.148

5.3. Third case: real world signals

As the EMD has been used for analyzing environmental datasets, we have chosen as a the third example the annual maximum daily precipitation time series obtained from the Canadian Regional Climate Model (CRCM), which has been used for climate change analysis [20]. Testing this kind of time series is complicated, as we do not have any *a priori* information about its functional form. In Table 3, we show the results of applying both approaches tested in this section. Notice that the correlation-based method fails at selecting relevant IMFs, as it has accepts all of them. The new technique, on the other hand, selected the IMFs that happen to give the highest mutual information with the original signal.

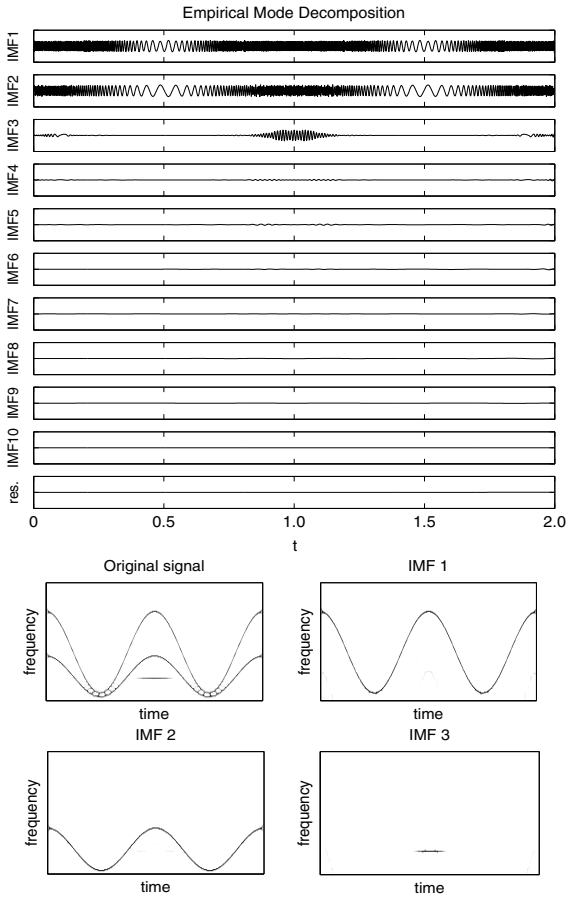

Fig. 2. For the second test signal, the IMFs returned by the EMD together with the TF representations of the whole signals and the first three IMFs.

Table 3. For the third example, mutual information (MI) between the IMFs selected as relevant and the original signal

New approach							
IMFs	1	2	res.				
MI	**0.375**	**0.201**	**0.123**				
Correlation approach							
IMFs	1	2	3	4	5	6	res.
MI	0.375	0.201	0.119	0.103	0.117	0.074	0.123

6. CONCLUSIONS

In this paper, we have generalized some approaches that use correlation to evaluate the relevance of the IMFs to energy-based approaches. With this generalized framework, we have proposed a new algorithm for selecting relevant IMFs, which does not require the specification of an external threshold. By comparison with the traditional technique that uses correlation, we have observed that the new method is more stringent, while being able to select those IMFs capturing the relevant information from the signal.

7. REFERENCES

[1] V. Venkatachalam and J.L. Aravena, "Nonstationary signal enhancement using the wavelet transform," in *System Theory, 1996., Proceedings of the Twenty-Eighth Southeastern Symposium on*, 1996, pp. 98–102.

[2] R.D. Nowak and R.G. Baraniuk, "Wavelet-based transformations for nonlinear signal processing," *Signal Processing, IEEE Transactions on*, vol. 47, no. 7, pp. 1852–1865, 1999.

[3] D. J. Thomson, *Nonlinear and Nonstationary Signal Processing*, pp. 317–394, Cambridge Univ. Press, 2001.

[4] M. Bayram and R. Baraniuk, *Nonlinear and Nonstationary Signal Processing*, chapter Multiple Window Time-Varying Spectrum Estimation, pp. 292–316, Cambridge Univ. Press, 2001.

[5] R. Vautard and M. Ghil, "Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series," *Physica D: Nonlinear Phenomena*, vol. 35, no. 3, pp. 395 – 424, 1989.

[6] S.H. Iranmanesh, A. Miranian, and M. Abdollahzade, "Signal processing using singular spectrum analysis for nonlinear system identification," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, 2012, pp. 722–727.

[7] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995.

[8] K.T Coughlin and K.K Tung, "11-year solar cycle in the stratosphere extracted by the empirical mode decomposition method," *Advances in Space Research*, vol. 34, no. 2, pp. 323–329, 2004.

[9] E. P. Souza et al., "Assessment of cardiovascular autonomic control by the empirical mode decomposition," in *4th Int. Workshop on Biosignal Interpretation*, 2002, pp. 123–126.

[10] D. Boutana et al., "On the selection of intrinsic mode function in emd method: Application on heart sound signal," in *Applied Sciences in Biomedical and Communication Technologies (ISABEL), 2010 3rd International Symposium on*, 2010, pp. 1–5.

[11] G. Rilling, P. Flandrin, and P. Gonçalvès, "On empirical mode decomposition and its algorithms," in *Proceedings of the 6th IEEE/EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP '03), Grado, Italy*, 2003.

[12] Q. Wang and S.-T. Quek, *Comparison of Hilbert-Huang, Wavelet, and Fourier Transforms for Selected Applications*, pp. 213–244, CRC Press, 2005.

[13] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 460, no. 2046, pp. 1597–1611, 2004.

[14] M. Zareen and Y. Hujun, "Information preserving empirical mode decomposition for filtering field potentials," in *Intelligent Data Engineering and Automated Learning - IDEAL 2009*, vol. 5788 of *Lecture Notes in Computer Science*, pp. 226–233. Springer Berlin Heidelberg, 2009.

[15] Z.K. Peng, Peter W. Tse, and F.L. Chu, "A comparison study of improved hilbert-huang transform and wavelet transform: Application to fault diagnosis for rolling bearing," *Mechanical Systems and Signal Processing*, vol. 19, no. 5, pp. 974 – 988, 2005.

[16] A. Ayenu-Prah and N. Attoh-Okine, "A criterion for selecting relevant intrinsic mode functions in empirical mode decomposition," *Advances in Adaptive Data Analysis, Theory and Applications*, vol. 2, no. 1, pp. 1–24, 2010.

[17] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *Signal Processing Letters, IEEE*, vol. 11, no. 2, pp. 112–114, feb. 2004.

[18] A. Moghtaderi, P. Flandrin, and P. Borgnat, "Trend filtering via empirical mode decompositions," *Computational Statistics and Data Analysis*, vol. 58, pp. 114–126, feb 2013.

[19] Y. I. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Phys Rev E*, vol. 52, no. 3, pp. 2318–2321, 1995.

[20] B. Music and D. Caya, "Evaluation of the hydrological cycle over the mississippi river basin as simulated by the canadian regional climate model (CRCM)," *J. Hydrometeor*, vol. 5, no. 8, pp. 968–988, 2007.