Gait Rhythm Fluctuation Analysis for Neurodegenerative Diseases by Empirical Mode Decomposition

Peng Ren*, Shanjiang Tang, Fang Fang, Lizhu Luo, Lei Xu, Maria L. Bringas-Vega, Dezhong Yao, Keith M. Kendrick*, and Pedro A. Valdes-Sosa*

***Abstract*—Previous studies have indicated that gait rhythm fluctuations are useful for characterizing certain pathologies of neurodegenerative diseases such as Huntington's disease (HD), amyotrophic lateral sclerosis (ALS), and Parkinson's disease (PD). However, no previous study has investigated the properties of frequency range distributions of gait rhythms. Therefore, in our study, empirical mode decomposition was implemented for decomposing the time series of gait rhythms into intrinsic mode functions from the high-frequency component to the low-frequency component sequentially. Then, Kendall's coefficient of concordance and the ratio for energy change for different IMFs were calculated, which were denoted as *W* and R_E, respectively. Results revealed that the frequency distributions of gait rhythms in patients with neurodegenerative diseases are less homogeneous than healthy subjects, and the gait rhythms of the patients contain much more high-frequency components. In addition, parameters of *W* and R_E can significantly differentiate among the four groups of subjects (HD, ALS, PD, and healthy subjects) (with the minimum *p*-value of 0.0000493). Finally, five representative classifiers were utilized in order to evaluate the possible capabilities of *W* and R_E to distinguish the patients with neurodegenerative diseases from the healthy subjects. This achieved maximum area under the curve values of 0.949, 0.900, and 0.934 for PD, HD, and ALS detection, respectively. In sum, our study suggests that gait rhythm features extracted in the frequency domain should be given consideration seriously in the future neurodegenerative disease characterization and intervention.**

***Index Terms*—Amyotrophic lateral sclerosis, empirical mode decomposition (EMD), gait rhythm fluctuation, Huntington's disease, Kendall's coefficient of concordance, neurodegenerative disease, Parkinson's disease, ratio for energy change, stance, stride, swing, synthetic minority oversampling technique (SMOTE).**

I. Introduction

WALKING involves the alternating action of the two lower extremities and its pattern is described as a gait cycle, which is defined as the interval of time between any of the repetitive events of walking [1]. The gait cycle consists of two periods: stance and swing [2], [3]. The stance phase begins at the instant that one extremity contacts the ground and continuous only as long as some portion of the foot is in contact with the ground. On the contrary, the swing phase begins as soon as the toe of one extremity leaves the ground and ceases just before heel strike or contact of the same extremity [2], [3]. Winter has proposed several tasks for walking, for example, maintenance and support of head, arm, and trunk; maintaining the balance of the body; and so on [1], [4], [5].

Several neurological disorders are characterized by the progressive degeneration of different mechanisms involved in motor responses and provoke gait's abnormality. Observation and evaluation of patients' walking are crucial to the neurological examination and diagnosis [6]. Parkinson's disease (PD) is a chronic neurodegenerative disorder characterized by static tremor, rigidity, bradykinesia, abnormal gait, and posture. PD is mainly caused by the degeneration of dopaminergic neurons and apoptosis of the nigrostriatal area in the midbrain, leading to a decline of dopamine levels in neostriatum and a decrease of excitatory neurotransmission from the thalamus to the motor cortex. PD is a chronic and slowly progressive disease. The symptoms can be persistent and worsen progressively over years. PD is not considered to be fatal, but people with PD do have a shorter life expectancy [7]. Huntington's disease (HD) is a chronic progressive chorea caused by the inherited degeneration of basal ganglia and cerebral cortex. HD is caused by the mutation of HTT gene, and code a protein called Huntington. This is considered to play an important role in the neuron degeneration of the brain. The HD patients have the symptoms of chorea with involuntary jerking or twitching movements. While the disease progresses, these abnormal movements may become more evident. Patients with HD may find difficulty in walking, speaking, and swallowing [8]. Amyotrophic lateral sclerosis (ALS) is a chronic progressive neurodegenerative disorder that damages the functions of the motor neurons in the central nervous system composed of the spinal cord and the brain. It will cause the disorder of the muscle movement and a series of clinical symptoms, such as muscle weakness, atrophy, fasciculation, and hyperreflexia [9]. The causes of such disease

Manuscript received September 28, 2015; revised January 18, 2016; accepted February 23, 2016. Date of publication March 1, 2016; date of current version December 20, 2016. *Asterisk indicates corresponding author.*

*P. Ren, *Keith M. Kendrick, and *Pedro A. Valdes-Sosa are with the Key Laboratory for NeuroInformation of Ministry of Education, Center for Information in Medicine, University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: pren28@uestc.edu.cn; k.kendrick.uestc@gmail.com; pedro@uestc.edu.cn).

S. Tang is with the School of Computer Science and Technology, Tianjin University.

F. Fang is with the Department of Neurosurgery, Sichuan Provincial People's Hospital and Affiliated Hospital, University of Electronic Science and Technology of China.

L. Luo, L. Xu, M. L. Bringas-Vega, and D. Yao are with the Key Laboratory for NeuroInformation of Ministry of Education, Center for Information in Medicine, University of Electronic Science and Technology of China.

Digital Object Identifier 10.1109/TBME.2016.2536438

are assumed to be genetic, viral, and autoimmune diseases. The respiratory muscular paralysis or pulmonary infection may be the fatal outcome of ALS patients.

One of the accurate methods to derive gait rhythm fluctuations such as the time series of stride time, swing time, and stance time is based on the signals of vertical ground reaction force (VGRF) recorded from ultrathin force-sensitive switches placed inside each subject's shoes [10]. Because these derived gait rhythm fluctuations are nonlinear and nonstationary time series, which have no specific sampling frequencies as well, most of the previous studies implemented the variables calculated from the time domain to characterize the gait patterns of the patients with neurodegenerative diseases. Hausdorff *et al.* revealed that the "noisy" variations in the time series of stride time display a hidden and unexpected fractal-like property [10]. Wu and Krishnan employed the swing-interval turns count and the averaged stride interval as the parameters to differentiate ALS patients from the healthy subjects [11]. Frenkel-Toledo *et al.* found that the swing time variability may be used as a speed-independent biomarker of rhythmicity and gait steadiness. They also suggested that the increased gait variability in PD is not simply a consequence of bradykinesia [12]. Ren *et al.* implemented phase synchronization and conditional entropy to analyze the relationship between gait rhythms of both feet for the patients with neurodegenerative diseases and suggested that it can be applied to neurodegenerative disease characterization and therapeutic intervention [13]. In addition, some previous studies attempted to use time–frequency (TF) analysis to investigate gait rhythms. Khandoker *et al.* used a wavelet-based scheme to compute the relative risk of falls for subjects with locomotion disorders [14]. Baratin *et al.* first implemented the procedure of data resampling and then extracted features based on wavelet analysis for the detection of neurological abnormalities [15]. Although TF analysis has the ability to reveal frequency information of gait rhythms, no previous studies have investigated their frequency range distributions in patients with different neurodegenerative diseases (in our case, the gait rhythms of PD, HD, ALS, and healthy subjects were used for analysis). Furthermore, so far, there is no biomarker for gait rhythms, which can effectively differentiate among these four groups of people, as mentioned above, through exploration of the frequency-domain information of their gait rhythms. Finally, the commonly used TF method, wavelet analysis, must select a set of basic signal components as the priori assumptions and then calculates the parameters for each of these signals such that their aggregate will compose the original signal. Therefore, in our study, empirical mode decomposition (EMD), a type of data-driven and self-adaptive signal decomposition approach, was implemented for the purpose of neurodegenerative disease characterization based on the features represented in the frequency domain.

II. Materials and Methods

A. Gait Rhythm Records

In our study, we utilized the gait rhythm dataset provided by Hausdorff *et al.* in PhysioBank for analysis (http://www.physionet.org/physiobank/database/gaitndd/). This is a large and extensively used archive of physiological data for public research founded by the Harvard-MIT division of health sciences and technology [16]. This dataset contains 64 subjects in total, which include 15 PD patients (age: 66.8 $\pm$ 10.9 (SD) years, 10 men and 5 women), 13 ALS patients (age: 54.9 $\pm$ 13.4(SD) years, 10 men and 3 women), 20 HD patients (age: 47.7 $\pm$ 12.2 (SD) years, 6 men and 14 women), and 16 healthy subjects (age: 39.3 $\pm$ 18.5 (SD) years, 2 men and 14 women), respectively. In the experiment, each subject was required to walk at his/her self-paced speed along a 77-m-long hallway for 5 min. In order to retrieve the VGRF signals, each subject's shoes had ultrathin force-sensitive switches placed inside. Sampling frequency of VGRF signals was 300 Hz and the recorded data were stored in a lightweight ankle-worn recorder. Next, the measures of footfall contact times for each stride were derived from these VGRF signals. It should be noted that the first 20 s of gait rhythm data were removed before analysis for the sake of minimizing start-up effects during recording. According to the previous studies of Hausdorff *et al.*, a median filter was applied in order to remove data points that were three standard deviations greater than or less than the mean value. These outliers were largely due to the turns at the end of the hallway.

In our study, we employed five types of time series of gait rhythm fluctuations for analysis: stride time (the time period from initial contact of one foot to the subsequent contact of the same foot), swing time (amount of time one foot is in the air during one stride time), stance time (amount of time one foot is on the ground during one stride time), percentage swing time (100 $\times$ swing time/stride time), and percentage stance time (100 $\times$ stance time/stride time). It is worth noting that these five types of gait rhythms do not have the sampling frequency since they were derived from the VGRF signals. In addition, only the patients at the advanced stages of diseases were selected for disease group comparisons in line with previous studies (PD: Hoehn and Yahr score $\geq$3, nine subjects in total; HD: total functional capacity score $\leq$5, nine subjects in total; ALS: the number of months since diagnosis $\geq$9, nine subjects in total) [17], [18]. The reasons for this are as follows: first, because three groups of patients have varying degrees of impairment, it is easy to obtain an unreliable conclusion if we compare one entire patient group with the other. Second previous studies found that there is no strong correlation relationship between the ages of the patients and the severities of the diseases. Third, at advanced stages of neurodegenerative disease, gait patterns of the patients are more seriously disrupted by this than by physiological aging.

B. Empirical Mode Decomposition

EMD, a type of TF analysis approach, is a procedure for decomposing a complicated set of data into a finite and often-small number of intrinsic mode functions (IMFs), which are defined as functions having the same number of extrema and zero crossings, with its envelopes being symmetric with regard to zero [19]. The EMD method is data-driven, locally adaptive, multiscale, and robust. It is, therefore, very suitable for processing nonstationary, nonlinear, and time-varying data [20]. The EMD method behaves like bandpass filtering. Hence, EMD could represent the information of gait rhythms in the frequency domain to some extent [21]. The procedures of extracting an IMF are as follows:

1) Identify the local maxima and minima.

2) Interpolate between the maxima by a cubic spline curve, which obtains the upper envelopes, denoted as $E_u(t)$.
3) Interpolate between the minima by a cubic spline curve, which obtains the lower envelopes, denoted as $E_l(t)$.
4) Calculate the mean value of the upper and lower envelopes from steps 2 and 3:

$$m_{11}(t) = \frac{1}{2}\left(E_u(t) + E_l(t)\right). \tag{1}$$

5) Achieve the candidate h_{11} for the first IMF component:

$$h_{11}(t) = x(t) - m_{11}(t). \tag{2}$$

6) In the most cases, the first candidate h_{11} does not satisfy the IMF conditions, and iterate steps 1–5:

$$h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t) \tag{3}$$

until the threshold for standard deviation SD for the two consecutive siftings is satisfied:

$$SD = \sum_{t=0}^{T} \frac{\left|h_{1(k-1)}(t) - h_{1k}(t)\right|^2}{h_{1(k-1)}^2(t)}. \tag{4}$$

7) After *k* iterations, the first IMF is obtained:

$$c_1 = h_{1k}. \tag{5}$$

8) The first residue can be taken from the difference of the signal *x(t)* and c_1:

$$r_1 = x(t) - c_1. \tag{6}$$

9) Iterate steps 1–6 on the residual r_n in order to obtain all the IMFs :

$$r_1 - c_2 = r_2 \tag{7}$$

$$r_{n-1} - c_n = r_n \tag{8}$$

until the c_n or r_n is smaller than the predefined threshold, or the residual signal is either a constant, a monotonic slope, or a function with only one extrema. The original signal *x(t)* can be expressed as

$$x(t) = \sum_{i=1}^{n} c_i + r_n. \tag{9}$$

Through the EMD method, the instantaneous amplitude of each IMF component can be expressed by means of a Hilbert transform. The equation for the Hilbert transform is given below:

$$y(t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{c(\tau)}{t-\tau} d\tau \tag{10}$$

where *P* is the Cauchy principal value. The analytic signal *z(t)* can be expressed as follows:

$$z(t) = c(t) + iy(t). \tag{11}$$

The instantaneous amplitude *a(t)* can be expressed as

$$a(t) = \sqrt{c^2(t) + y^2(t)}. \tag{12}$$

Because the last IMF may represent the overall signal trend rather than a true IMF, in our study, only the first five IMFs were implemented (all the time series of gait rhythms were decomposed into six IMFs in total).

C. Kendall's Coefficient of Concordance

Kendall's coefficient of concordance, known as Kendall's *W*, is a coefficient used to measure the association between two pairs of ranked data. It is a descriptive statistical approach, which represents the degree of relationship between two or more variables [22]. Especially, Kendall's *W* allows a researcher to evaluate the degree of agreement between *m* sets of ranks for *n* subjects or objects according to a particular characteristic [23]. The range of Kendall's *W* is from 0 (no agreement) to 1 (complete agreement). For example, each of a number of "judges" (*m*) ranks a given set of *n* objects, i. e., from the best to the worst. (The "judges" can be different variables, characters or any other physical entities, in reality.) If the test statistical value of *W* is 1, it means that all the "judges" have unanimous agreement in their rankings. On the contrary, if the test statistical value of *W* is 0, it indicates that there is no overall trend of agreement among the different "judges," and their responses are extremely diverse. It should be noted that Kendall's *W* does not make any assumptions about the probability distribution properties of the population. In the biomedical signal processing area, for *f*MRI studies, Kendall's *W* is one of the significant approaches to measure the regional homogeneity of the time series of neighboring voxels of interests in order to differentiate normal and abnormal brain functions [24]. In addition, Kendall's *W* is also widely used in time-domain measures of inter-channel electroencephalogram correlations [25]. Suppose the number of a "judge" is *j* and the object *i* is given by him the rank $r_{i,j}$. Then, the total rank given to object *i* is

$$R_i = \sum_{j=1}^{m} r_{i,j}. \tag{13}$$

And *S* is defined as

$$S = \sum_{i=1}^{n} \left(R_i - \overline{R}_i\right)^2 = \sum_{i=1}^{n} R_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} R_i\right)^2. \tag{14}$$

And the definition of Kendall's *W* is given by

$$W = \frac{S}{\frac{1}{12}m^2\left(n^3 - n\right)}. \tag{15}$$

The value of *W* is from 0 to 1 and all the values of $r_{i,j}$ should be ordinal before calculating the Kendall's *W*. In our study, the first five IMFs were implemented for analysis. Hence, *n* is given by 5 and *m* is equal to the length of gait rhythms. In addition, it should be noted that in our study, the smaller the value of *W*, the less difference in the instantaneous amplitudes of the IMFs are observed.

D. Ratio for Energy Change

EMD behaves spontaneously as a filter bank and decomposes the time series into IMFs from the "high-frequency" to "low-frequency" components sequentially [26]. Hence, it has

been shown that analyzing the inherent relations between IMFs contributes to exploring a variety of biomedical problems such as epileptic spike detection [27], [28]. In our study, the instantaneous amplitudes of the first five IMFs were computed in order to derive the ratio for energy change of different IMFs, which could manifest the hidden relationship between various "frequency bands" of gait rhythm to some extent. We regard the first and the second IMFs as the relatively "high-frequency" components, whereas the fourth and the fifth IMFs as the relatively "low-frequency" components of gait rhythms. The ratio for energy change is defined as follows:

$$E_h = \sum_{t=1}^{N} a_1^2(t) + \sum_{t=1}^{N} a_2^2(t) \tag{16}$$

$$E_l = \sum_{t=1}^{N} a_4^2(t) + \sum_{t=1}^{N} a_5^2(t) \tag{17}$$

$$R_E = \frac{E_h - E_l}{E_l} \tag{18}$$

where a_1, a_2, a_4, and a_5 denote the instantaneous amplitudes of the first, second, fourth, and fifth IMFs, respectively. N is the length of the gait rhythm. E_h denotes the whole energies of the first and the second IMFs. E_l denotes the whole energies of the fourth and the fifth IMFs. Because gait is composed of the movements of both feet, the average R_E (of both feet) was calculated as a newly developed parameter for further analysis.

E. Pattern Classification

In our study, a machine learning technique was implemented in order to evaluate the abilities of W and R_E to distinguish the patients with HD, ALS, and PD from the healthy subjects. Because this dataset contains numbers of patients with HD, ALS, or PD, which are not equal to that of healthy subjects, the unequal prior probabilities of the two class datasets might lead the classifiers to be more sensitive in detecting the majority rather than the minority class. Oversampling and undersampling in data analysis are the two frequently used approaches to adjust the class distribution of unbalanced datasets. The oversampling method has the advantage of not only being beneficial for correcting for a bias in the original dataset but also preventing loss of information. Therefore, in most classification problems related to skew class distributions, the oversampling method is selected as the best way. In our study, the synthetic minority oversampling technique (SMOTE), a mature approach to address unbalanced dataset problems, is implemented, which will not exaggerate the receiver operating characteristic curve of the features or cause overfitting problems [29]–[31].

We discussed above how to handle class imbalance problems by applying SMOTE. Next, we used five representative classifiers to test the proposed algorithm comprehensively: Naïve Bayes (NB), support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), and simple logistic regression (SLR), which are all implemented using the Waikato Environment for Knowledge Analysis software [32]. The RF is a learning ensemble algorithm, which consists of a bagging of decision trees with a randomized selection of features at each split. Decision tree is one of most popular methods used for data exploration. One type of decision tree is called classification and regression tree, which divides feature space into sets of disjoint rectangular regions by means of greedy and recursive partitioning [33]. RF has many advantages. For example, it can produce a highly accurate classifier. It is also effective for missing data estimation and large database running. The NB classifier is a generative model that determines the class of a given data object by the maximum *a posteriori* estimation [34]. The NB can be trained very efficiently in a supervised learning setting. However, the parameter estimation requires a large number of labeled data. SVM is a maximum margin-based algorithm for efficiently training learning machines in kernel-induced feature spaces, and at the same time, it generalizes well to unseen data and is insensitive to the class imbalance problem [35]. MLP is a type of feed-forward artificial neural network model, aiming at learning linear functions that project data objects from the feature space to the target space. Each layer of MLP has a number of perceptrons and different layers are fully connected. MLP is often trained by the back propagation algorithm, which tunes parameters of MLP by minimizing the least-mean-squares error [36] between the output and the target. SLR is a classifier built on linear logistical models that learns the best function that fits the data. By training regression functions for different classes, an unknown data object can be classifier to the class to which the data object has the shortest distance in the feature space.

Principal component analysis (PCA) is a feature dimension reduction technique that aims to project a data object to a new feature space with fewer dimensions while minimizing the reconstruction error. The projection function is identified by the eigenvalue decomposition method. The eigenvectors with top eigenvalues will be selected as the projection function. Principal components (PCs) are a new set of variables, which are linear combinations of the observed ones. Because of the decreasing variance property, much of the variance tends to be concentrated in the first few PCs. In order to take proper number of PCs for dimension reduction, the cumulative percentage of total variation for the selection of PCs needs to be calculated. Each eigenvalue can represent the variance of each PC. Hence, if λ_i represents the ith eigenvalue for the ith PC, then the cumulative percentage of total variation for the first r PCs is calculated as follows:

$$\frac{\sum_{k=1}^{r} \lambda_k}{\sum_{k=1}^{N} \lambda_k} = \frac{1}{N} \sum_{k=1}^{r} \lambda_k \tag{19}$$

where N denotes the total number of eigenvalues. Generally, we choose the first r PCs, whose cumulative percentage of total variation is just above 95%. In addition, to avoid overfitting problem, the tenfold cross-validation method was employed for assessing the generalization ability for an independent dataset. In the experiment, the area under the curve (AUC) is used to measure the performance for classification and regression tasks. The greater potential of classification always has a larger AUC value [37].

III. Results

Fig. 1 shows the first five IMFs of the time series of stride time for one healthy subject. It is obvious that these five IMFs were displayed from the relatively high- to low-frequency com-

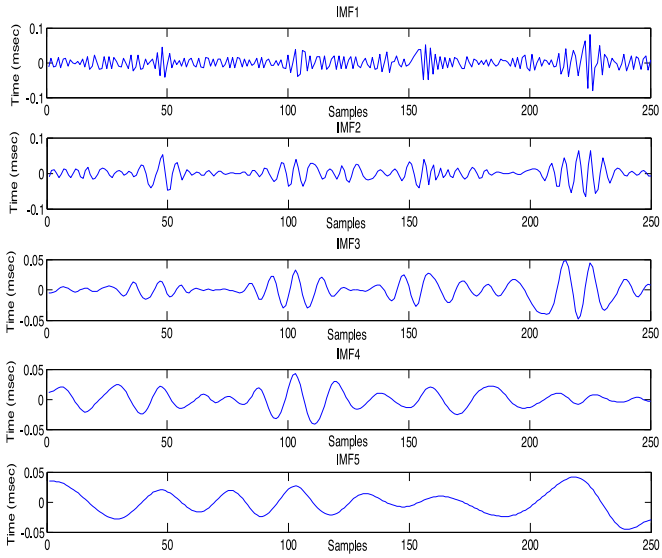


Fig. 1. EMD of a healthy subject record. The first five IMFs obtained by EMD are from the top to the bottom.

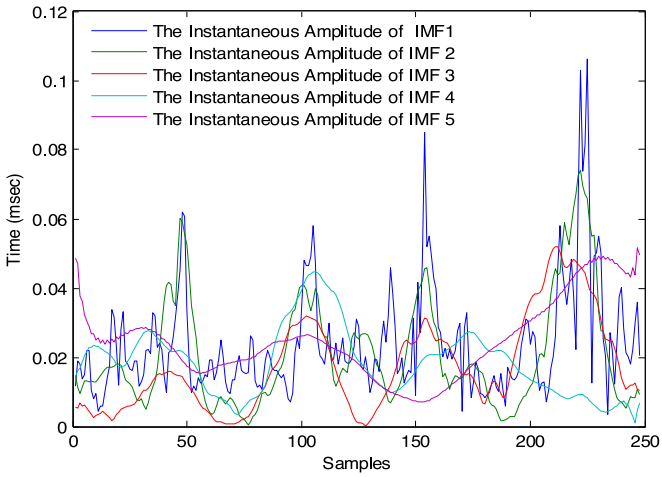


Fig. 2. Instantaneous amplitudes of the first five IMFs.

ponents sequentially. Fig. 2 shows the instantaneous amplitudes of the IMFs displayed in Fig. 1. Table I shows the means and the standard deviations of the values of W and R_E for the five types of gait rhythms for all the HD, ALS, and PD patients, and healthy subjects in the dataset. From Table I, it is obvious that for the parameter of W, the healthy subjects have almost the minimum means and standard deviations, which implies that the frequency range distributions of gait rhythms of the healthy subjects are relatively homogeneous. In our study, the smaller the value of W, the less difference in instantaneous amplitudes of the IMFs was observed. In addition, from Table I, it is also apparent that, compared with the healthy subjects, the patients with neurodegenerative diseases have larger values of R_E, which indicates that the gait rhythms of the patients contain much more high-frequency components. In our study, the Kruskal–Wallis test was first implemented to test for statistical differences among the four different groups and the results were shown in Table I as well. It is apparent that all the p values of the derived parameters are less than 0.05. In addition, the Wilcoxon ranked sum tests of W and R_E were employed and the results were shown in Table II (ALS versus healthy subjects, HD versus healthy subjects, PD versus healthy subjects, advanced ALS versus advanced HD, advanced PD versus advanced HD, and advanced ALS versus advanced PD). No assumptions are made about the underlying distributions of the data being compared for these nonparametric tests. In Tables I and II, if the p value is less than or equal to 0.05, we reject the null hypothesis.

In our study, because of the limited number of subjects, if all the extracted features were implemented for classification directly (which contain five values of W and R_E, respectively), it might lead to the problem of overfitting. Hence, PCA was utilized. For the classification of PD and healthy subjects, the first four PCs were selected. For the classification of ALS and healthy subjects, the first five PCs were selected. For the classification of HD and healthy subjects, the first four PCs were selected. Table III shows the values of AUC for the classifications of PD and healthy subjects, ALS and healthy subjects, and HD and healthy subjects, respectively.

IV. Discussion

Gait is a rhythmic and semiperiodic behavior, which has been demonstrated in a number of previous studies [7], [10]–[12]. In addition, all of these studies have shown that the fractal structures of the gait rhythms of PD, HD, and ALS are significantly different from those of healthy subjects [7]. Hence, in our study, we assume that their properties extracted in the frequency domain should also be altered due to the disease disturbance.

In our study, we proposed a new approach to investigate the frequency range distributions of gait rhythms for PD, HD, and ALS patients, compared to healthy subjects, to provide new potential biomarkers for neurodegenerative disease characterization. Gait rhythm fluctuations comprise nonlinear, multiscale, and nonstationary signals in essence, so it is very appropriate to use EMD to analyze them.

Although EMD is a self-adaptive and data-driven method, its frequency-domain properties can be compared with the spectrum analysis to some extent. If we assume that the sampling frequency of one signal is 1 Hz (normalized frequency), the mean frequency of the fifth IMF is around 0.025 Hz. Hence, the fifth IMF can represent the low-frequency components of the signal. Although the EMD might decompose the signal into more numbers of IMFs if changing their predefined convergence threshold into smaller values, the later IMFs might not truly reflect the frequency components of the signal. Therefore, it is important to choose the suitable number of IMFs for research. Until now, there are no standard criteria for IMF selection and this is one of the deficiencies for EMD. In our study, we implemented the commonly used convergence threshold for EMD analysis, which is relatively reliable and robust.

Kendall's W was implemented to measure the homogeneity of the instantaneous amplitudes of IMFs for the four groups of subjects. In Table I, it is obvious that compared with the PD, HD, and ALS patients, the healthy subjects have almost the minimum mean values, which implies that the instantons amplitudes of IMFs of the healthy subjects have inconspicuous difference in the observations of the whole time series. In other words, all the frequency ranges of the gait rhythms of the healthy

TABLE I
MEANS, STANDARD DEVIATIONS, AND P VALUES OF KRUSKAL–WALLIS TESTS OF THE PARAMETERS W AND R_E FOR HD, PD, ALS, AND HEALTHY SUBJECTS

		HD	PD	ALS	Healthy Subjects	P Value
Stride Time (s)	W	5.465e-04 ± 8.835e-04	1.693e-04 ± 4.614e-04	9.662e-04 ± 0.003	9.155e-06 ± 4.085e-06	0.000
	R_E	5.489 ± 4.597	3.891 ± 4.033	3.096 ± 3.677	0.910 ± 0.905	0.000
Swing Time (s)	W	2.191e-04 ± 3.492e-04	2.695e-05 ± 2.395e-05	3.116e-05 ± 2.414e-05	4.236e-06 ± 2.433e-06	0.000
	R_E	6.982 ± 3.088	5.041 ± 3.430	4.840 ± 1.879	3.147 ± 1.453	0.001
% Swing Time	W	0.740 ± 1.018	0.124 ± 0.008	0.111 ± 0.103	0.024 ± 0.016	0.000
	R_E	7.485 ± 3.690	4.735 ± 2.441	3.955 ± 2.224	3.974 ± 1.324	0.002
Stance Time (s)	W	4.245e-04 ± 6.709e-04	8.558e-05 ± 2.361e-04	0.002 ± 0.006	5.571e-06 ± 3.074e-06	0.000
	R_E	4.776 ± 3.781	2.341 ± 2.726	5.423 ± 9.013	0.756 ± 0.689	0.001
% Stance Time	W	0.740 ± 1.018	0.124 ± 0.008	0.111 ± 0.103	0.024 ± 0.016	0.000
	R_E	7.485 ± 3.690	4.735 ± 2.441	3.955 ± 2.224	3.974 ± 1.324	0.002

TABLE II
P VALUES OF WILCOXON RANKED SUM TESTS OF THE PARAMETERS W AND R_E FOR ALS AND HEALTHY SUBJECTS, HD AND HEALTHY SUBJECTS, PD AND HEALTHY SUBJECTS, HD AND PD, HD AND ALS, AND ALS AND PD

		ALS versus Healthy Subjects	PD versus Healthy Subjects	HD versus Healthy Subjects	HD versus PD	HD versus ALS	ALS versus PD
Stride Time (s)	W	0.000	0.000	0.000	0.003	0.031	0.489
	R_E	0.020	0.007	0.000	0.161	0.050	0.310
Swing Time (s)	W	0.000	0.000	0.000	0.006	0.004	0.730
	R_E	0.008	0.101	0.000	0.113	0.077	0.730
% Swing Time	W	0.001	0.000	0.000	0.000	0.000	0.387
	R_E	0.948	0.545	0.001	0.006	0.004	0.730
Stance Time (s)	W	0.000	0.000	0.000	0.050	0.161	0.931
	R_E	0.020	0.012	0.000	0.014	0.011	0.730
% Stance Time	W	0.000	0.000	0.000	0.000	0.000	0.387
	R_E	0.948	0.545	0.001	0.006	0.004	0.730

TABLE III
VALUE OF AUC FOR THE CLASSIFICATIONS OF HD AND HEALTHY SUBJECTS, PD AND HEALTHY SUBJECTS, AND ALS AND HEALTHY SUBJECTS BASED ON THE DERIVED PARAMETERS OF W AND R_{E1}

	HD versus Healthy Subjects	PD versus Healthy Subjects	ALS versus Healthy Subjects
RF	0.885	0.865	0.900
SL	0.843	0.949	0.859
MLP	0.878	0.910	0.934
NB	0.898	0.875	0.891
SVM	0.900	0.906	0.906
Average	0.881	0.901	0.898

subjects have the highest degree of uniformity and monotonic distributions among the four groups.

In order to further investigate the dominant frequency range, which alters the homogeneity of the frequency distributions of the gait rhythms of the patients with neurodegenerative diseases, the ratio for energy change, denoted as R_E, was implemented. The results revealed that the high-frequency components of the gait rhythms of the PD, HD, and ALS patients are much stronger than the healthy subjects. This might be due to the fact that the neuromuscular system is disturbed by the neurodegenerative diseases, leading to the gait variability change. Previous studies have shown that in healthy subjects, the stride-to-stride fluctuations are relatively small and the variations of many parameters (such as the stride time) are on the border of just a few percent. However, for the patients with neurodegenerative diseases, the fractal dynamics of gait are significantly changed, which adds the so-called noisy variations to the gait rhythms [12]. In addition, analyzing the energy change between various frequency bands of the physiological time series is very meaningful. For example, in heart rate variability signal, both sympathetic and parasympathetic systems embody their activities in the frequency range from 0.05 to 0.15 Hz, whereas the activities of sympathetic system are only reflected in the frequency range from 0.15 to 0.4 Hz. In our study, we found that the increase of the high-frequency component of gait rhythms might suggest abnormality in the neuromuscular system, which is consistent with the viewpoint of previous studies that for some measures of gait dynamics, less variability is better (i.e., fractal index, standard deviation) [10].

It is evident that in Tables I and II, the increases in W and R_E are common and nonspecific changes, which can be observed in walking by neurodegenerative disease patients. These two newly derived parameters both represent the frequency information of gait rhythms to some extent. For HD, the most common symptom is chorea, that is, dance-like or jerky movement of the arms and legs. Previous studies also showed that HD patients exhibit significantly more gait variability than PD and ALS patients. In our study, we discovered that the HD patients generally have the largest value of W (the larger the value of W, the weaker the homogeneity of the frequency distributions of the gait rhythms) and R_E (the larger the value of R_E, the more high-frequency components the gait rhythms have), which are easily explained by the physical features of HD as mentioned above. PD is a type of neurodegenerative disease producing akinesia (an absence of spontaneous movement) whereas HD is one that produces hyperkinesia (rapid involuntary movements). Hence, the values

of W and R_E for PD are generally smaller than HD. It should be noted that the pathologies of PD and HD are both related to the disorder of the basal ganglia and their gait patterns are very similar as well. However, in Table II, it is obvious that these two types of neurodegenerative diseases can be easily differentiated by the parameters of W and R_E, which suggests that the extracted features of gait rhythms in the frequency domain is also very useful for the diagnosis of neurodegenerative disease.

Considering almost of the previous studies kept the same comparison strategies as Hausdorff *et al.*, in our work, we take the same approach (only the patients with advanced ALS, advanced HD and advanced PD were implemented for disease group comparisons). In order to further investigate whether our developed parameters are robust or not, we further analyzed all the patients without any selection. The results are very similar with the ones shown in Table II. For the comparison of HD and PD, there are eight out of ten variables, whose p values are less than 0.05. For the comparison of HD and ALS, there are six out of ten variables, whose p values are less than 0.05.

In this dataset, it also contains a separate file which includes the clinical information for each subject such as a measure of disease duration or severity. For the PD patients, the Hohn and Yahr score is provided (a higher score indicates more advanced disease). For the HD patients, the total functional capacity score is provided (a lower score indicates more advanced functional impairment). For the ALS patients, the time in months since the diagnosis of the disease is given. In order to further investigate whether our developed parameters are useful in the characterization of pathological alteration related to the disease, we also calculated the correlation coefficients of the values of W and R_E with the severities of PD, HD, and ALS, respectively. The maximum absolute values of correlation coefficients for PD, HD, and ALS are 0.631, 0.781, and 0.595, respectively, which indicate the possibility of using W and R_E for computer-aided diagnosis. Compared with the absolute values of correlation coefficients derived from HD, the calculated ones for PD and ALS are not high enough, which might be due to the following reasons. First, the Hohn and Yahr scale only uses five stages to evaluate the symptoms of PD progress, which may not reflect the severity of PD patients completely (more and more clinicians have already implemented the Unified Parkinson's Disease Rating Scale (UPDRS) to evaluate the disease severity of the PD patients. The UPDRS can assess the limitation of PD daily activities in more details). Second, for ALS patients, only the number of months since the diagnosis of the disease is provided, which might not truly reflect the severity of ALS disease because the start of ALS may be so subtle that the symptoms are overlooked, or every patient does not have the same speed of progression of the disease. In our study, these two newly developed parameters can reflect the physiological information in the frequency domain, which have totally different biological meanings from the other previously derived parameters. Hence, in the future, it is possible to combine all these developed parameters together (such as stride time, fractal index etc.) and employ a regression technique to accurately evaluate neurodegenerative disease progression.

In our study, there are some limitations needed to be clarified and explained. As mentioned in the previous literature, in this dataset, the ages of the groups were not matched very well. However, many previous studies neglected this mismatch and investigated the gait rhythms of patients with neurodegenerative diseases based on this dataset due to the following three reasons [7], [14], [38]. First, it has been widely accepted that the impact of a neurodegenerative disease on the alterations of gait patterns is more pronounced than those due to physiological aging. Hausdorff *et al.* once showed that a young man with neurodegenerative disease possesses obviously abnormal gait patterns if compared with a much older healthy subject [7]. Second, this is the only public data archive which contains the records of gait rhythms measured from most kinds of neurodegenerative diseases. Most of the other archives only have datasets of PD patients. Third, the worldwide prevalences of ALS and HD are very low (ALS: one to three cases per 100 000 individuals; HD: five to ten cases per 100 000 individuals). Hence, it is very difficult to collect such a valuable dataset [39]–[41]. It is worth noting that in the dataset, the age of HD patients was not significantly different from healthy subjects and their corresponding p values shown in Table II are all less than 0.05, with the smallest one of 0.0000493. In order to eliminate the possible influence of age on the group comparisons, six healthy subjects and four PD patients were removed, resulting in no significant age difference between ALS and healthy subjects ($p > 0.05$), and PD and healthy subjects ($p > 0.05$). The results show that for these two group comparisons, there are six out of ten (ALS versus healthy subjects) and five out of ten (PD versus healthy subjects) parameters, whose p values are less than 0.05. This could demonstrate the effectiveness and validity of our developed approach to some extent.

V. Conclusion

In summary, we have presented a new insight on gait rhythm fluctuation analysis in humans, which is based on investigating their frequency range distributions. In addition, the parameters (W and R_E) were developed in order to characterize the clinical symptoms of the patients with HD, ALS, and PD, which could potentially facilitate diagnosis of these neurodegenerative diseases. Overall, the results using this novel approach are strong and have high predictive sensitivity. Therefore, our research suggests that the gait rhythm features extracted from the frequency domain should be given serious consideration in future study.

Acknowledgment

The authors would like to thank the dataset provider—Harvard-MIT Division of Health Sciences and Technology.

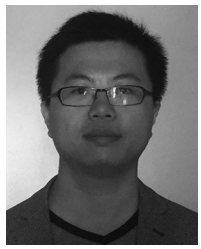
References

[1] D. A. Winter, *Biomechanics and Motor Control of Human Movement*. Hoboken, NJ, USA: Wiley, 2009, pp. 9–12.
[2] D. Levine and J. Richards, *Whittle's Gait Analysis*. London, U.K.: Churchill Living Stone, 2012.
[3] J. Perry and J. Burnfield, *Gait Analysis: Normal and Pathological Function*. Thorofare, NJ, USA: Slack Inc., 2010.
[4] S. Kumar, *Perspectives in Rehabilitation Ergomonics*. London, U.K.: CRC Press, 1997.
[5] D. Purves and G. J. Augustine, *Neuroscience*. Sunderland, MA, USA: Sinauer Assoc., 2011.
[6] J. B. Dingwell *et al.*, "Local dynamic stability versus kinematic variability of continuous overground and treadmill walking," *J. Biomechan. Eng.*, vol. 123, no. 1, pp. 27–32, Jan. 2001.

[7] J. M. Hausdorff *et al.*, "Gait variability and basal ganglia disorders: Stride-to-stride variations of gait cycle timing in Parkinson's disease and Huntington's disease," *Movement Disorders*, vol. 13, no. 3, pp. 428–437, Mar. 1998.
[8] C. A. Ross and S. J. Tabrizi, "Huntington's disease: From molecular pathogenesis to clinical treatment," *Lancet Neurol.*, vol. 10, no. 1, pp. 83–98, Jan. 2011.
[9] M. C. Kiernan *et al.*, "Amyotrophic lateral sclerosis," *The Lancet*, vol. 377, no. 4, pp. 645–646, Apr. 2011.
[10] J. M. Hausdorff, "Gait variability: Methods, modeling and meaning," *J. Neuroeng. Rehabil.*, vol. 2, no. 1, pp. 1–9, Jan. 2005.
[11] W. F. Wu and S. Krishnan, "Computer-aided analysis of gait rhythm fluctuations in amyotrophic lateral sclerosis," *Medical Biol. Eng. Comput.*, vol. 47, no. 11, pp. 1165–1171, Nov. 2009.
[12] S. Frenkel-Toledo *et al.*, "Effect of gait speed on gait rhythmicity in Parkinson's disease: Variability of stride time and swing time respond differently," *J. Neuroeng. Rehabil.*, vol. 2, no. 1, pp. 1–7, Jan. 2005.
[13] P. Ren *et al.*, "Analysis of gait rhythm fluctuations for neurodegenerative diseases by phase synchronization and conditional entropy," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 2, pp. 291–299, Feb. 2016.
[14] A. H. Khandoker *et al.*, "Wavelet-based feature extraction for support vector machines for screening balance impairments in the elderly," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 587–597, Dec. 2007.
[15] E. Baratin *et al.*, "Wavelet-based characterization of gait signal for neurological abnormalities," *Gait Posture*, vol. 41, no. 2, pp. 634–639, Feb. 2015.
[16] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–E220, Jun. 2000.
[17] C. G. Goetz *et al.*, "Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations," *Movement Disorders*, vol. 19, no. 9, pp. 1020–1028, Sep. 2004.
[18] K. Marder *et al.*, "Rate of functional decline in Huntington's disease," *Neurology*, vol. 54, no. 2, p. 452, Feb. 2000.
[19] N. E. Huang *et al.*, "The empirical mode decomposition of the Hilbert spectrum for nonlinear and nonstationary time series analysis," *Proc. Royal Soc. London A*, vol. 454, pp. 903–995, 1998.
[20] H. L. Liang *et al.*, "Empirical mode decomposition: A method for analyzing neural data," *Neurocomputing*, vols. 65/66, pp. 801–807, Mar. 2005.
[21] N. E. Huang *et al.*, "A new view of nonlinear water waves: The Hilbert spectrum," *Ann. Rev. Fluid Mech.*, vol. 31, pp. 417–457, Jan. 1999.
[22] J. G. Proakis and D. K. Manolakis, *Digital Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2006.
[23] A. P. Field, "Kendall's coefficient of concordance," in *Encyclopedia of Statistics in Behavioral Science*. New York, NY, USA: Wiley, 2005.
[24] R. Baumgartner *et al.*, "Assessment of cluster homogeneity in fMRI data using Kendall's coefficient of concordance," *Magn. Resonance Imaging*, vol. 17, no. 10, pp. 1525–1535, Oct. 1999.
[25] J. Bourien *et al.*, "Mining reproducible activation patterns in epileptic intracerebral EEG signals: application to interictal activity," *IEEE Trans. Biomed. Eng*, vol. 51, no. 2, pp. 304–315, Feb. 2004.
[26] L. Stankovic *et al.*, *Time-Frequency Signal Analysis with Applications*. Norwood, MA, USA: Artech House, 2013.
[27] J. C. Echeverria *et al.*, "Application of empirical mode decomposition to heart rate variability analysis," *Med. Biol. Eng. Comput.*, vol. 39, no. 4, pp. 471–479, Apr. 2001.
[28] Z. Chen *et al.*, "Spike extraction of epileptic waves in EEG based on EMD," *J. Biomed. Eng.*, vol. 24, no. 5, pp. 973–977, May 2007.
[29] N. V. Chawla *et al.*, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jan. 2002.
[30] F. Amir and J. Shahram, "An expert system for detection of breast cancer using data preprocessing and bayesian network," *Int. J. Adv. Sci. Technol.*, vol. 34, pp. 65–70, Sep. 2011.
[31] K. E. Mahsereci and I. Turgay, "Effective automated prediction of vertebral column pathologies based on logistic model tree with SMOTE preprocessing," *J. Med. Syst.*, vol. 38, no. 5, pp. 50–50, May 2014.
[32] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA, USA: Morgan Kaufmann, 2005.
[33] A. Cutler *et al.*, "Random forests," *Mach. Learning*, vol. 45, no. 1, pp. 157–176, Jan. 2001.
[34] A. K. Jain *et al.*, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.
[35] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.
[36] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences," *Atmospheric Environ.*, vol. 32, pp. 2627–2636, Oct. 1998.
[37] D. Faraggi and B. Reiser, "Estimation of the area under the ROC curve," *Statist. Med.*, vol. 21, no. 20, pp. 3093–3106, 2002.
[38] H. Zheng *et al.*, "Machine learning and statistical approaches to support the discrimination of neuro-degenerative diseases based on gait analysis." *Stud. Comput. Intell.*, vol. 189, pp. 57–70, 2009.
[39] F. Geser *et al.*, "Amyotrophic lateral sclerosis, frontotemporal dementia and beyond: The TDP-43 diseases," *J. Neurol.*, vol. 256, no. 8, pp. 1205–1214, Aug. 2009.
[40] J. M. Hausdorff, "Gait dynamics in Parkinson's disease: Common and distinct behavior among stride length, gait variability, and fractal-like scaling," *Chaos Interdisciplinary J. Nonlinear*, vol. 19, no. 2, p. 14, Feb. 2009.
[41] J. M. Hausdorff, "Gait dynamics, fractals and falls: Finding meaning in the stride-to-stride fluctuations of human walking," *Human Movement Sci.*, vol. 26, no. 4, pp. 555–589, Apr. 2007.

Peng Ren received the B.S. degree from the Department of Biomedical Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2007, and the Ph.D. degree from the Department of Biomedical Engineering, Florida International University, Miami, FL, USA, in 2013.

He is currently an Assistant Professor in the Department of Biomedical Engineering and the Key Laboratory for NeuroInformation Ministry of Education, Center for Information in Medicine, University of Electronic Science and Technology of China. His research interests include biomedical signal and image processing, biomedical data mining, rehabilitation engineering, and neuroscience.



Shanjiang Tang received the B.S. and M.S. degrees from Tianjin University (TJU), Tianjin, China, in 2008 and 2011, respectively, and the Ph.D. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2015.

He is currently an Assistant Professor in the School of Computer Science and Technology, TJU. He worked with the IBM China Research Lab in the area of performance analysis of multicore-oriented Java multithreaded program as an intern for four months in 2009. His research interests include parallel computing, cloud computing, big data analysis, and computational biology.

Dr. Tang received the "Golden Prize" in the 31th ACM/ICPC Asia Tournament of National College Students in 2006 and the "Talents Science Award" from TJU in 2007.

Fang Fang received the M.B.B.S. degree from Weifang Medical University, Shandong, China, in 2014. She is currently working toward the M.S. degree in Medical School, University of Electronic Science and Technology of China, Chengdu, China.

Her research interests include neurosurgery clinical studies and localization of epileptic foci using multimodality neuroimaging.

Ms. Fang was awarded the outstanding graduate student in 2015.

Lizhu Luo received the B.S. degree and the M.S. degree in psychology from the Department of Psychology, Southwest University, Chongqing, China, in 2010 and 2012, respectively. She is currently working toward the Ph.D. degree in the Department of Biomedical Engineering, University of Electronic Science and Technology of China, Chengdu, China.

Her research interests include cognitive neuroscience and neuroinformation processing.

Lei Xu received the B.S. degree in psychology from the Department of Psychology, Henan Normal University, Xinxiang, China, in 2009, and the M.S. degree in cognitive neuroscience from the Department of Psychology, Southwest University, Chongqing, China, in 2013. She is currently working toward the Ph.D. degree in biomedical engineering at the University of Science and Technology of China, Chengdu, China.

Her research interests include cognitive neuroscience and neuroinformation processing.

Maria L. Bringas-Vega received the B.S. degree in psychology from the University of Havana, Havana City, Cuba, in 1974, and the Ph.D. degree in psychophysiology from the National Center for Scientific Research of Cuba, La Habana, Cuba, in 1991.

She was a Senior Researcher and Head of the Neuropsychology Department, International Center for Neurological Restoration Havana, Havana, Cuba, from 1995 to 2014. She is currently a Senior Professor of neuropsychology at the Cuban Neuroscience Center and a Professor of Neuroinformatics at the University of Electronic Sciences and Technology of China, Chengdu, China, where she has been working in application of neuroinformatics methods to neurorestoration.

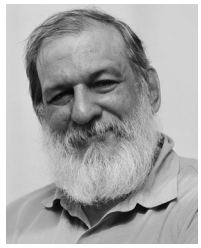
Dezhong Yao received the Ph.D. degrees in applied geophysics from the Chengdu Institute of Geology, Chengdu, China, in 1991, and in biomedical science and engineering from Aalborg University, Aalborg, Danmark, in 2005.

He has been a Faculty Member with the University of Electronic Sciences and Technology of China, Chengdu, since 1993, where he became a full Professor in 1995 and the Dean of the School of Life Science and Technology in 2001. He was a Visiting Scholar at the University of Illinois at Chicago, IL, USA, from September 1997 to August 1998, and a Visiting Professor at McMaster University, Hamilton, ON, Canada, from November 2000 to May 2001, and at Aalborg University from November 2003 to February 2004. He is the author or coauthor of more 100 papers and five books. His main contributions are EEG zero reference (reference electrode standardization technique-REST), scale-free brainwave music, and EEG-fMRI information fusion method. His current interests include brain–computer interface and multimodal imaging-based epilepsy study.

Keith M. Kendrick received the B.A. and Ph.D. degrees in psychology from the University of Durham, Durham, U.K., in 1976 and 1979, respectively.

He has held research positions in the University of Durham Institute of Zoology, London, the University of Cambridge, and at the Babraham Institute in Cambridge, where he was the Head of Cognitive and Systems Neuroscience prior to moving to his current post in the School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, in September 2011 as a 1000 Talent Professor. He has been a Gresham Professor since 2002, delivering 30 public lectures on the biomedical sciences (all available on internet). He has published more than 200 refereed papers, including eight in nature and science and received numerous grants from both U.K.- and U.S.-based funding agencies.

Dr. Kendrick has been a Fellow of the Society of Biology in the U.K. since 1996.



Pedro A. Valdes-Sosa received the M.D. degree and the B.S. degree in mathematics from the University of Havana, Havana City, Cuba, in 1972 and 1973, respectively, and the Ph.D. degree in neuroscience from the National Center for Scientific Research of Cuba, La Habana, Cuba, in 1974. He received the Doctor in Science (higher scientific degree) in 2011.

He is the Cofounder and has been the General Vice-Director for Research of the Cuban Neuroscience Center since 1990. He is also a Distinguished Professor of neuroinformatics at the University of Science and Technology of China, Chengdu, China. He was named a Senior Researcher in 1980 at the National Center for Scientific Research of Cuba. He is also a Second Degree Medical Specialist in neurophysiology (1981) and a Senior Professor of the Higher Institute of Medical Sciences of Havana (2005). He has published more than 240 papers.

Dr. Valdes-Sosa has been a member of the Cuban Academy of Sciences since 2000.