

HMM for Classification of Parkinson's Disease Based on the Raw Gait Data

Abed Khorasani · Mohammad Reza Daliri

Received: 14 January 2014 / Accepted: 22 October 2014 / Published online: 30 October 2014
© Springer Science+Business Media New York 2014

Abstract The central nervous system (CNS) plays an important role in regulation of human gait. Parkinson's disease (PD) is a common neurodegenerative disease that may cause neurophysiologic change in the CNS and as a result change the gait cycle duration (stride interval). This article used the Hidden Markov Model (HMM) with Gaussian Mixtures to separate the patients with PD from healthy subjects. The results showed that the performance of the HMM classifier in classifying the gait data corresponding to 16 healthy and 15 PD subjects is comparable to the results obtained from the least squares support vector machine (LS-SVM) classifier. In this study, the leave-one-out cross-validation method was used to evaluate the performance of each classifier. The HMM method could effectively separate the gait data in terms of stride interval obtained from healthy subjects and PD patients with an accuracy rate of 90.3 %. All in all, the results showed that the proposed method can be used for distinguishing PD patients from healthy subjects based on the gait data classification.

Keywords Gait classification · Hidden markov model · Movement disorders · Parkinson's disease

Introduction

Parkinson's disease (PD) is the most common neurodegenerative movement disorder. This type of disease is caused by both despair of motor control and malfunction of rhythm generation in the basal ganglia which has a strong effect on

voluntary movement control [1]. Typical motor symptoms of PD known as movement disorders are tremor, bradykinesia, rigidity, and postural instability [2]. By progression of this disease, not only postural instability but also gait disturbances can be commonly seen in many cases. Some gait disturbances such as festination, short gait step and freezing gait can make diagnosis of PD easier and so investigation of parameters of gait would be a very useful method for both understanding of the mechanism of motor control and recognizing of the neurological disease progression [3,4].

Many different methods have been proposed in recent years for the diagnosis of PD [5–7]. Furthermore, in order to measure the parameters of the gait in PD and also investigate its characteristics, the computer-based methods have been widely utilized in the previous studies [8–14]. In [15], the coefficient of variation as a criterion for stride-to-stride fluctuations was used for analyzing of the gait data of both healthy control subjects and PD patients. They showed that this coefficient is larger in the gait data of PD patients and also can be used as the degree for disease severity. In addition, in [16] gait data in terms of acceleration signals corresponding to PD and healthy subjects were analyzed and the results showed that fractal dimensions of the body in PD patients is higher than that of healthy subjects. The same study also recommended that the acceleration signal during locomotion in both old and PD subjects alter with a complex pattern [17]. Although in these methods the stride-to-stride variability has been seen and analyzed, representing a powerful model for characterizing the gait variability has been remained an open problem.

In [18], in order to evaluate the gait variability in PD patients, two features were extracted and used for classification of gait data derived from healthy and PD subjects. They showed that by using a nonlinear support vector machine, it is possible to distinguish these gait data with an appropriate rate of accuracy. However, in this method most of temporal information of gait data has been neglected and only two features

This article is part of the Topical Collection on *Education & Training*

A. Khorasani · M. R. Daliri (✉)
Biomedical Engineering Department, Faculty of Electrical Engineering, Iran University of Science and Technology (IUST), Narmak, 16846-13114 Tehran, Iran
e-mail: daliri@iust.ac.ir

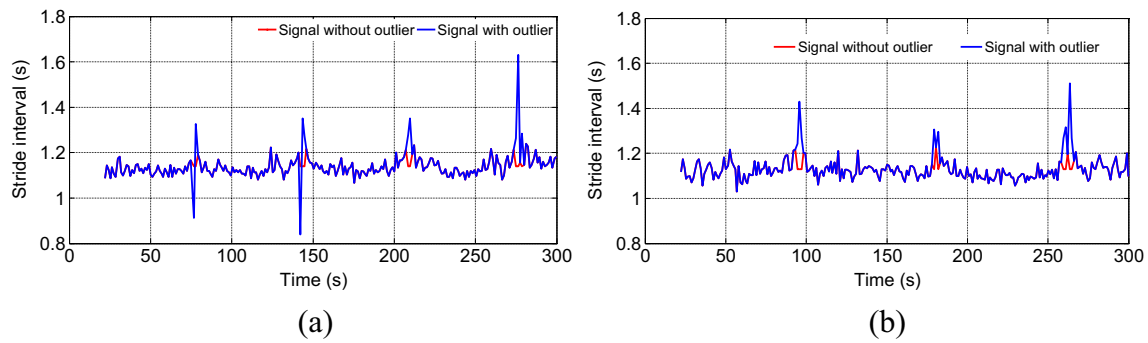


Fig. 1 Typical result of outlier removal from the raw gait data in terms of stride interval: **(a)** of a 74-year-old male healthy subject; **(b)** of a female subject with PD disease. The gait signal with outliers is shown with blue

used for classification task. In this paper, Hidden Markov Models (HMM) is used for modeling of the raw gait data instead of the extracted features. HMMs have been widely investigated and applied for past several years in the automatic speech recognition applications [19]. Recently, this method have been successfully used for both medical diagnosis and system monitoring applications such as EEG classification, ECG classification, character recognition and fault diagnosis system [20–23]. Especially, this method would be useful for the classification of signals in terms of time sequences. In this article, we used HMM to classify the gait data derived from healthy and PD subjects and compared its performance against a sophisticated nonlinear classifier.

This paper is organized based on the following sections. Section II describes gait data and preprocessing procedures and reviews HMM formulation. Section III presents our results on the gait data classification problem. Finally, section IV concludes this paper.

Methods

Data description

The gait data provided by Hausdorff et al. were used in this study [10,15], and also can be downloaded from web page of physioNet (<http://www.physionet.org>) [24]. These gait data were recorded from 16 healthy subjects (2 men and 14 women) with age of 20–74 years and 15 PD subjects (10 men and 5 women) with age of 44–80 years as they walked at their normal pace in a 77 m long hallway for 5 min. The mean age (standard deviation) of the healthy, PD subjects were 39 (18.5), 67 (10.9) years, respectively. In the healthy group, any mental problems or motion disorders was not reported. The ages of the two groups were not noticeably different. From the recorded force applied to the ground during walking, 5 min of recording, consisting of stride,

swing, and stand times for each leg and double support signals for both groups were derived.

Preprocessing

In the current study, the gait data from the right foot of both PD and healthy subjects in terms of stride interval (time from the contact of a foot to the ground to the following contact of same foot) were used. In order to remove the effect of unwanted artifacts in the start of recording, the first 20s of each gait data samples were removed before analysis of the gait data [10,15]. Based on the gait experiment described in [10], the subjects were requested to walk through a hallway during recording of gait signals and turn around at the end of the hall way and then continue walking. Because of this reason, the extracted gait signals represent large values in these points and so should be removed before processing. For removing these outliers, according to “sigma rule” method, the samples in the gait data with amplitude greater or less than 2 SDs of median value of whole signal were replaced by median value [25]. In Fig. 1 the typical results of removing of outliers from stride interval time sequence of a PD patient and a healthy subject are shown.

Hidden marko model

A hidden Markov model (HMM) is a state machine with two layers including state and observation layers in which a

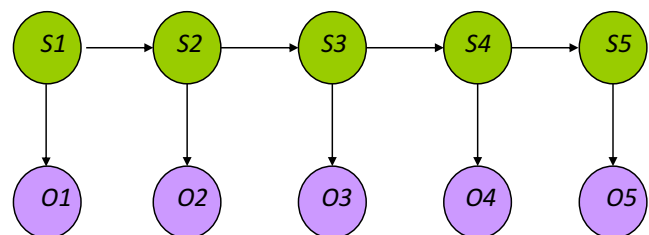
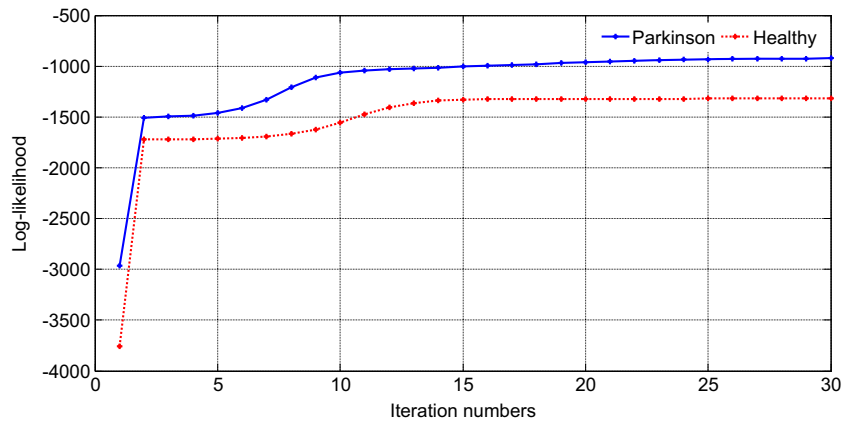


Fig. 2 Representing an HMM. The states in each time are hidden but only outputs can be observed

Fig. 3 Log-likelihood increase during increasing of iterations of the training process for both PD and healthy gait data



Marcovian process controls the selection of the state in each time. In standard HMMs, by using a discrete hidden state at time (t) all of the needed information before this time would be known and so the observation at any time depends only on its current hidden state. In each time, the HMM system is located in one state and transition between these states are defined based on an associated probability. Furthermore, each state is related to the output observation with its associated probability. The graphic representation of the HMM is illustrated in Fig. 2. The states of the Markov chain are hidden, but the outputs from the Markov chain are observable.

An HMM for continuous data processing is represented by three matrix, $\lambda=(\pi, A, B)$, consisting of a vector of initial probabilities π , a matrix of transition probabilities A, and a vector of probabilistic output functions B. Each element of mentioned matrixes is represented based on the following equations:

$$\pi_i = \Pr(s_i \text{ at } t = 0) \tag{1}$$

$$a_{ij} = \Pr(s_i \text{ at } t + 1 | s_j \text{ at } t) \tag{2}$$

$$b_i(x) = \Pr(O_t = x | s_i \text{ at } t) \tag{3}$$

where x is an observed signal value and s is the state of the HMM model. In the current study, the output function B is a Gaussian density function which is shown in the following equation:

$$n(O, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}[\frac{O-\mu}{\sigma}]^2} \tag{4}$$

where O, μ and σ represents sequence of observation, average and standard deviation, respectively. In order to utilize an HMM model in the real- world applications, three main problems should be solved:

Problem one Given the observation sequence $O = (o_1, o_2, \dots, o_T)$ and the model $\lambda=(\pi, A, B)$, what is the best way for computing $P(O|\lambda)$, the probability of the observation sequence. This problem is known as Evaluating.

Fig. 4 The relationship between Recognition rates and number of Gaussian mixtures for PD and healthy subjects during training phase

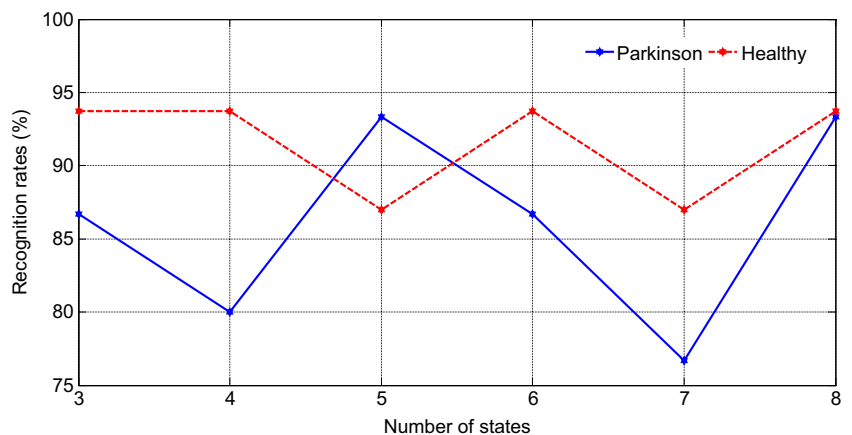
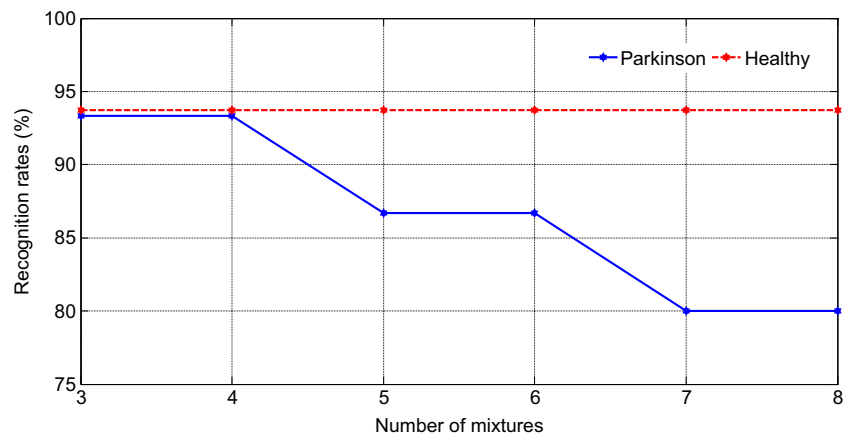


Fig. 5 The relationship between Recognition rates and number of states for PD and healthy subject during training phase



Problem two Given the observation sequence $O = (o_1, o_2, \dots, o_T)$ and the model $\lambda = (\pi, A, B)$, what is the best way for finding corresponding sequence of states $S = (s_1, s_2, \dots, s_T)$. This problem is known as Decoding.

Problem three what is the best way for adjustment of HMM model parameter (λ) to maximize $P(O|\lambda)$. This problem is known as Training.

The significance of solving the third problem is obvious because the model parameters must first be estimated before the models can be used for classification purposes. For solving this problem, Baum-Welch algorithm has been introduced in the literature [19]. Furthermore, Forward backward recursive and Viterbi algorithms can be used for solving first and second problems, respectively [19].

In the following, the summary of HMM method in both training and test phases has been presented. For further information about the Hidden Markov Model refer to [19].

1. The number of states (N) and also the structure of transition matrix are identified.
2. The sequence of observation containing all of the training samples which belong to a specific class are calculated and then these samples are divided into the same segments equal to the number of state using K-means algorithm.
3. The initial values of average and standard deviation corresponding to Gaussian model are calculated using K-means algorithm,
4. Two separated HMM model for classification of PD and Healthy subjects are trained using the Baum-Welch algorithm.
5. The parameters of each model corresponding to each class are saved and used for the test phase.
6. In the test phase, Forward-backward algorithm is used and by comparison the calculated log-likelihood probability of test data the class of each data is identified.

Results

As it was told in previous section, the structure of HMM model is described by three matrix $\lambda = (\pi, A, B)$ and so determining the parameters of this model such as the number of states and the number of Gaussian mixture is very important for training of the HMM model. In addition to these parameters, iteration numbers can be regarded as an important factor in training phase that should be considered for training of the model. In order to determine these parameters, the effect of each one on the performance of classifier has been investigated separately when all of the subjects are used for both the training and test phase. As it was told previously, the Baum-Welch algorithm is used to optimize the parameters of HMM model. By using this algorithm through iteration a maximum of the likelihood would be obtained. In Fig. 3 the relationship between number of iterations and log-likelihood for both PD and healthy subjects is shown. As can be seen in this figure, the log-likelihood maximization begins during about 20 iterations for both PD and healthy gait data. Therefore, an HMM classifier with twenty iterations can be used for classification.

In order to investigate the effect of the number of states on the performance of classifier, the relationship between performance of classifier and the parameter of number of states has been investigated. Fig. 4 shows the recognition rates for PD and healthy subjects for different number of states. These results show that the best classification performance is obtained when the number of state is 5.

Table 1 HMM structure

Evaluating problem	Forward backward
Decoding problem	Viterbi algorithm
Training problem	Baum-Welch
Number of iteration	20
Number of mixtures	4
Number of states	5

Table 2 result of recognition of Parkinson and healthy subjects

Classifier	Type	No. of subjects	Detected as PD	Detected as healthy	X=Sensitivity Y=Specificity	Overall accuracy
HMM	PD	15	14	1	X=93.33 %	90.32 %
	Healthy	16	2	14	Y=87.50 %	
LS-SVM	PD	15	13	2	X=86.67 %	90.32 %
	Healthy	16	1	15	Y=93.75 %	

The same procedures have been used to investigate the effect of number of Gaussian mixtures on the performance of HMM classifier. Figure. 5 shows the recognition rates for PD and healthy subjects for different number of Gaussian mixtures. These results show that, the best classification performance is obtained when the number of mixture is 4.

By investigation the effect of aforementioned parameters on the performance of classifier, the number of states, the number of Gaussian mixture and maximum number of iterations are chosen 5, 4 and 20, respectively. The final structure of HMM model with selected parameters for the final evaluation of performance of HMM model in gait data classification is shown in Table 1.

In Table 2 the result of classification of PD and healthy subjects based on the stride-interval gait data using HMM method has been shown. The leave-one-out (LOO) cross-validation method was used for evaluation of performance of the proposed classifier [26]. In this method, one gait pattern is used for the test of validation and the remaining patterns are used for the training of the classifier. According to Table 2, one PD and two healthy subjects were classified incorrectly by HMM. Two statistical parameters including sensitivity and specificity are used for evaluation of the accuracy of classifier in detection of PD and healthy subjects separately. These parameters are defined as follows:

Sensitivity: the percentage of PD patients who are correctly recognized as PD ones.

Specificity: the percentage of healthy subjects who are correctly recognized as healthy ones.

In Table 2 the performance of HMM classifier is compared to LS-SVM method based on the obtained result from [18]. The comparison of results show that the HMM method resulted in higher sensitivity rates. However, higher specificity rates are obtained by using LS-SVM algorithm. The overall accuracy rate of both methods is similar with accuracy of 90.3 %.

Conclusions

Signal processing methods can help engineers to achieve diagnostic information through analysis of walking patterns in the patients suffering from neurological diseases. This information can be used to separate persons with specific neurodegenerative diseases from healthy ones. In the current

study, we have tried to introduce a method to distinguish the PD persons from healthy ones. The current study is consisted of three main steps; raw gait data obtaining, preprocessing and classification using HMM. The gait rhythms of both PD and healthy subjects in terms of stride interval are used as the input of HMM classifier. In the preprocessing step, the samples of gait data recognized as outliers are simply substituted with the median value of whole time series of gait data. In the final step, the gait data are classified using HMM classifier. The investigation of results shows that the proposed method is efficient for interpretation of PD. The proposed classifier based on the raw gait data can correctly identify more than 90 % of the 31 subjects. These results are comparable to the results corresponding to the nonlinear SVM classifier based on the two feature extracted from the same gait data [18]. The same gait dataset was also used in several studies. In [27] a linear model was proposed to investigate the stride interval time series in PD, but the results showed that such a linear model is only useful for interpretation of the gait signals during walking at a constant speed.

In this article because of lack of match between age and gender of the subjects in PD and healthy subjects, we could not investigate the effect of these two factors on the performance of proposed classifier. However, it was shown that the effect of gender on usual locomotion patterns is not considerable [28]. In another study, It was reported that the effect of age factor on walking is extremely complex [29]. However, studies show that the factor of neurological disease is more effective in changing the rhythm of gait than the factor of age [3]. In the current study, the size of gait dataset was low and so this limited us to test the generality of the proposed method on recognition of PD. In the future studies we hope to construct a larger gait dataset to evaluate the performance of current classifier on recognizing PD.

References

1. J. Sian, M. Gerlach, M. B. H. Youdim, and P. Riederer, "Parkinson's disease: A major hypokinetic basal ganglia disorder," *J. Neural Transmission*, vol. 106, no. 5–6, pp. 443–476, 1999.
2. J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol., Neurosurgery, Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

3. J.M. Hausdorff, S. L. Mitchell SL, R. Firtion, C. K. Peng, M. E. Cudkowicz, J. Y. Wei, A. L. Goldberger, "Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington's disease," *Journal of Applied Physiology*, vol. 82, no. 1, pp. 262–269, 1997.
4. J.M. Hausdorff, N. B. Alexander, "Gait disorders: evaluation and management," *New York, NY: Informa Healthcare*, 2005.
5. C. O. Sakar and O. Kursun, "Telediagnosis of Parkinson's disease using measurements of dysphonia," *Journal of Medical Systems*, vol. 34, no. 4, pp. 591–599, 2010.
6. A. Ozcift, "SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease," *Journal of Medical Systems*, vol. 36, no. 4, pp. 2141–2147, 2012.
7. H. Karimi Rouzbahani and M. R. Daliri, "Diagnosis of Parkinson's Disease in Human Using Voice Signals," *Basic and Clinical Neuroscience*, vol. 2, pp. 12–20, 2011.
8. J.M. Hausdorff, C. K. Peng, Z. Ladin, J. Y. Wei, A. L. Goldberger, "Is walking a random walk? Evidence for long-range correlations in stride interval of human gait," *Journal of Applied Physiology*, vol. 78, no. 1, pp. 349–358, 1995.
9. J.M. Hausdorff, P. L. Purdon, C. K. Peng, Z. Ladin, J. Y. Wei, A. L. Goldberger, "Fractal dynamics of human gait: stability of long-range correlations in stride interval fluctuations," *Journal of Applied Physiology*, vol. 80, no. 5, pp. 1448–1257, 1996.
10. J.M. Hausdorff, A. Lertratanakul, M. E. Cudkowicz, A. L. Peterson, D. Kaliton, A. L. Goldberger, "Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis," *Journal of Applied Physiology*, vol. 88, no. 6, pp. 2045–2053, 2000.
11. J.M. Hausdorff, "Gait dynamics, fractals and falls: finding meaning in the stride-to-stride fluctuations of human walking," *Human Movement Science*, vol. 26, no. 4, pp. 555–589, 2007.
12. F. Y. Liao, J. Wang, P. He, "Multi-resolution entropy analysis of gait symmetry in neurological degenerative diseases and amyotrophic lateral sclerosis," *Medical Engineering and Physics*, vol. 30, no. 3, pp. 299–310, 2008.
13. W. Aziz, M. Arif, "Complexity analysis of stride interval time series by threshold dependent symbolic entropy," *European Journal of Applied Physiology*, vol. 98, no. 1, pp. 30–40, 2006.
14. M. R. Daliri, "Chi-square distance kernel of the gaits for the diagnosis of Parkinson's disease," *Biomedical Signal Processing and Control*, vol. 8, pp. 66–70, 2013.
15. J. M. Hausdorff, M. E. Cudkowicz, R. Firtion, J. Y. Wei, and A. L. Goldberger, "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, 1998.
16. M. Sekine, M. Akay, T. Tamura, and Y. Higashi, "Fractal dynamics of body motion in patients with Parkinson's disease," *J. Neural Eng.*, vol. 1, no. 1, pp. 8–15, 2004.
17. M. Akay, M. Sekine, T. Tamura, Y. Higashi, T. Fujimoto, "Fractal dynamics of body motion in post-stroke hemiplegic patients during walking," *J Neural Eng.*, vol. 1, no. 2, pp. 111–116, 2004.
18. W. Yunfeng, S. Krishnan, "Statistical analysis of gait rhythm in patients with Parkinson's disease," *Neural Systems and Rehabilitation Engineering, IEEE Transactions*, vol. 18, no. 2, pp. 150–158, 2010.
19. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
20. W. Speier, C. Arnold, J. Lu, A. Deshpande, and N. Pouratian, "Integrating Language Information With a Hidden Markov Model to Improve Communication Rate in the P300 Speller" *Neural Systems and Rehabilitation Engineering, IEEE Transactions*, vol. 2, no. 3, pp. 678–684, 2014.
21. R. V. Andreão, B. Dorizzi, and J. Boudy, "ECG signal analysis through hidden Markov models" *Biomedical Engineering, IEEE Transactions on*, Vol. 53, no. 8, pp. 1541–1549, 2006.
22. O. Samanta, U. Bhattacharya, and S. Parui, "Smoothing of HMM Parameters for Efficient Recognition of Online Handwriting" *Pattern Recognition*, vol. 47, no. 11, pp. 3614–3629, 2014.
23. J. Chen and Y.-C. Jiang, "Development of hidden semi-Markov models for diagnosis of multiphase batch operation" *Chemical Engineering Science*, Vol. 66, no. 6, pp. 1087–1099, 2011.
24. G. B. Moody, R. G. Mark, and A. L. Goldberger, "PhysioNet: A webbased resource for the study of physiologic signals," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 70–75, 2001.
25. G. J. Hahn, S. S. Shapiro, *Statistical Models in Engineering*. New York: Wiley, 1994.
26. R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
27. T. Carletti, D. Fanelli, and A. Guarino, "A new route to non invasive diagnosis in neurodegenerative diseases," *Neurosci. Lett.*, vol. 394, no. 3, pp. 252–255, 2006.
28. A. Gabell and U. Nayak, "The effect of age on variability in gait" *Journal of Gerontology*, vol. 39, no. 6, pp. 662–666, 1984.
29. D. C. Kerrigan, M. K. Todd, C. U. Della Croce, L.A. Lipsitz, J. J. Collins, "Biomechanical gait alterations independent of speed in the healthy elderly: evidence for specific limiting impairments," *Archives of Physical Medicine and Rehabilitation* vol. 79, no. 3, pp. 317–322, 1998.