Detection Of Parkinson's Disease

N.Bharath , Madanapalle institute of technology and science

***Abstract*--** **We aim of classifying the Parametric and Non Parametric models by using the collected dataset of Parkinson's disease. The Parkinson's data is tested with two respective models to determine which model provides the higher classification accuracy. In parametric modeling, Logistic Regression is used to classify the Parkinson's data. From Non parametric modeling, K-Nearest Neighbors and Random Forest Algorithm is used to classify the training and test data of Parkinson's disease. The classification is made using the parametric and non-parametric model with the collected Parkinson's data. With the classified value of data, classification accuracy on parametric and nonparametric model is resulted. Comparison of both Parametric and Non Parametric model is done evaluate the performance of the Parkinson's dataset.**

Keywords – Parametric Modeling, Non-Parametric Modeling, Logistic Regression, k-nearest neighbor, Random Forest.

I. INTRODUCTION

Parkinson's disease is a long term de-generative disorder of the central nervous system. The symptoms come generally over time, most obvious shaking, rigidity, slowness of movement and difficulty in walking. The cause of Parkinson's is generally unknown, but involved with both genetic and environmental factors. More than one million cases per year occur in India. This disease cannot be cured, but treatment may help. Meditation can help in control of symptoms of Parkinson's disease. Most people symptoms take years to develop, and they for years with disease. The famous personalities affected by Parkinson's disease are, [1] Boxer Muhammad Ali was

diagnosed with Parkinson's disease in 1984, as the result of sustaining several severe head injuries. [2] Pope John Paul II was diagnosed with Parkinson's disease in 2001. [3] Michael J. Fox was diagnosed with Parkinson's disease in 1991. [4] Adolf Hitler suffered from Parkinson's disease, the first symptoms were observed was 1937. [5] Dave Jennings an American Football Player died from complications with Parkinson's disease in 2013. In this paper, two techniques are used, [1] Data analytics technique which is used for analyzing the data and it refers to qualitative and quantitative techniques and process used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data, patterns and trends. The extensive use of mathematics and statistics are descriptive techniques and predictive models to gain knowledge from the data. Organizations may apply analytics to business data to

describe predict and improve the business performance; especially areas within analytics include predictive analytics, prescriptive analytics and retail analytics. Analytics is the discovery interpretation and communication of meaningful patterns in data. [2] Machine learning is the method of data analysis that automates analytical model building using algorithms that allows the computer to find hidden insights. Machine learning is Artificial Intelligence model in which known data is designed to predict the model. Instead of writing code, that solves a specific problem, machine learning developers create algorithms that are able to take in data and then built their own logic on their data. Analytics done by machine learning provides more accurate results. The machine learning is important because, It can analyze bigger, more complex data and deliver faster with more accurate results even on a larger scale and also because so many different industries are starting to relay on machine learning.

II. RELATED WORK

We present an assessment of the practical value of existing traditional and non-standard

measures for discriminating healthy people from people with Parkinson's disease (PD) by detecting dysphonia. We introduce a new measure of dysphonia, Pitch Period Entropy (PPE), which is robust to many uncontrollable confounding effects including noisy acoustic environments and normal, healthy variations in voice frequency. We collected sustained phonations from 31 people, 23 with PD. We then selected 10 highly uncorrelated measures, and an exhaustive search of all possible combinations of these measures finds four that in combination lead to overall correct classification performance of 91.4%, using a kernel support vector machine. In conclusion, we find that non-standard methods in combination with traditional harmonics-to-noise ratios are best able to separate healthy from PD subjects. The selected non-standard methods are robust to many uncontrollable variations in acoustic environment and individual subjects, and are thus well-suited to tele monitoring applications.[1]

Voice disorders affect patients profoundly, and acoustic tools can potentially measure voice function objectively. Disordered sustained vowels exhibit wideranging phenomena from nearly periodic to highly complex, a periodic vibrations, and increased "breathiness". Modeling and surrogate data studies have shown significant nonlinear and non-Gaussian random properties in these sounds. Nonetheless, existing tools are limited to analyzing voices displaying near periodicity, and do not account for this inherent biophysical nonlinearity and non-Gaussian randomness, often using linear signal processing methods insensitive to these properties. They do not directly measure the two main biophysical symptoms of disorder: complex nonlinear aperiodicity, and turbulent, aero acoustic, nonGaussian randomness. Often these tools cannot be applied to more severe disordered voices, limiting their clinical usefulness. This paper introduces two new tools to speech analysis: recurrence and fractal scaling, which overcome the range limitations of existing tools by addressing directly these two symptoms of disorder, together reproducing a "hoarseness" diagram. A simple bootstrapped

classifier then uses these two features to distinguish normal from disordered voices. On a large database of subjects with a wide variety of voice disorders, these new techniques can distinguish normal from disordered cases, using quadratic discriminate analysis, to overall correct classification performance of 91.8 ± 2.0%. The true positive classification performance is 95.4 ± 3.2%, and the true negative performance is 91.5 ± 2.3% (95% confidence).This is shown to outperform all combinations of the most popular classical tools. Given the very large number of arbitrary parameters and computational complexity of existing techniques, these new techniques are far simpler and yet achieve clinically useful classification performance using only a basic classification technique. They do so by exploiting the inherent nonlinearity and turbulent randomness in disordered voice signals. They are widely applicable to the whole range of disordered voice phenomena by design. These new measures could therefore be used for a variety of practical clinical purposes.[2]

Ambulatory monitoring of motor symptoms in Parkinson's disease (PD) can improve our therapeutic strategies, especially in patients with motor fluctuations. Previously published monitors usually assess only one or a few basic aspects of the cardinal motor symptoms in a laboratory setting. We developed a novel ambulatory monitoring system that provides a complete motor assessment by simultaneously analyzing current motor activity of the patient (e.g., sitting, walking, etc.) and the severity of many aspects related to tremor, bradykinesia, and hyperkinesias. The monitor consists of a set of four inertial sensors. Validity of our monitor was established in seven healthy controls and six PD patients treated with deep brain stimulation (DBS) of the sub thalamic nucleus. The patients were tested at three different levels of DBS treatment. Subjects were monitored while performing different tasks, including motor tests of the Unified PD Rating Scale (UPDRS). Output of the monitor was compared to simultaneously recorded videos. The monitor proved very accurate in discriminating between several motor activities. Monitor output correlated well with blinded UPDRS ratings during different

DBS levels. The combined analysis of motor activity and symptom severity by our PD monitor brings true ambulatory monitoring of a wide variety of motor symptoms one step closer.[3]

III. ARCHITECTURE DIAGRAM

In this paper, the architecture diagram is defined with the flow of process which is used to refine the raw data and used for predicting the Parkinson's data. The first step is, preprocessing the collected raw data into understandable format. Standardization of data is used to standardize. Standardization is the technique which is used to transform the various format of data into the common format and by this the attributes of data are being choose for standardization. Normalization of data is made here by, min-max technique where the data is normalized using min and maximum of data values. The Parkinson's data is evaluated with the Parametric model using the Logistic Regression algorithm and the classification accuracy of this model is found. The Non-Parametric model using the KNearest Neighbor and Random Forest algorithm on Parkinson's data to predict the data and classification accuracy of this model is found. Finally, the result of both Parametric and Non-Parametric model model is compared on the basis of classification accuracy

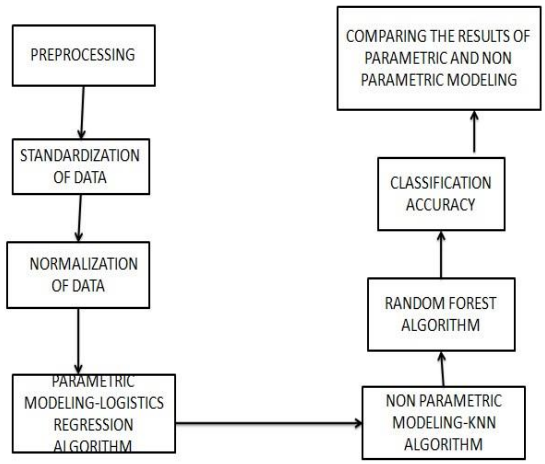


Figure 1 : Architecture diagram

IV. IMPLEMENTATION

The Parkinson's dataset is collected and analyzed to determine which model suitable for classification. Standardization and Normalization techniques are used to arrange the data in an orderly manner. According to the result obtained from the

standardization and normalization the analysis is made and the data is treated with Parametric and Non Parametric model which are used for classification. In parametric modeling, the Logistic Regression is used to analyze the classification accuracy of the data and in Non parametric modeling, K Nearest Neighbor algorithm and Random Forest algorithm has been used for classification accuracy. Comparisons of both the parametric and non parametric models are evaluated.

A. PRE-PROCESSING

Data preprocessing is the data mining technique that involves transforming raw data into understandable format. In Figure 2, The Raw data is highly susceptible to noise, missing values, and inconsistency. Real world data is often incomplete, inconsistent, and is likely to contain errors. Data preprocessing is the proven method for resolving such issues. In order to improve the quality of the data consequently, the mining results of raw data is preprocessed so the efficiency process improved. Data preprocessing is one of the most critical steps in a data mining process which deals with preparation and transformation of initial dataset.

Data preprocessing includes data cleaning, data transformation, data reduction, discretization and gathering the concepts of hierarchies.

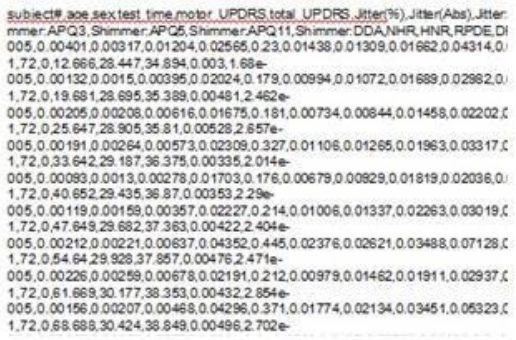


Figure 2 : Data Pre-Processing

B. STANDARDIZATION

Data standardization is the process by which similar data is collected in various formats is transformed to a common format that enhances the comparison process, allows for collaborative research and large scale analytics.

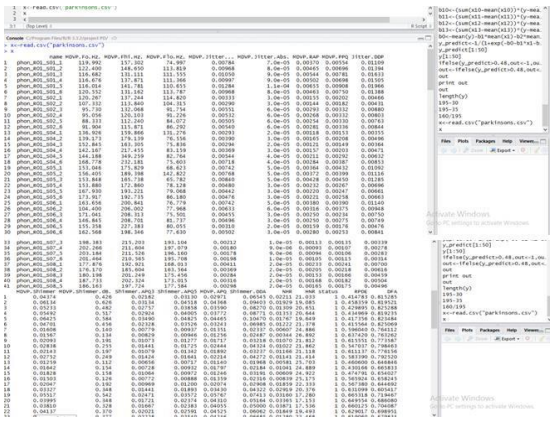


Figure 3 : Standardization

In Figure 3, the standardization is done using the formula which contains the value of mean and standard deviation. By standardizing, the data is predicted and analyzed for the next step of modeling. In this paper, out of 24 attributes, we choose 13 attributes and standardize the data to a common format.

Formula used for standardization:

X-mean(X)/ standard deviation(X)

C. NORMALIZATION

Database normalization or simply normalization is the process of organizing columns (attributes) and tables (relations) of a relational database. It reduces the data redundancy and improves data integrity. In Figure 4, the purpose of normalization is to prove more consistent and operational relation. Min-Max normalization is often known as feature scaling where the values of a numeric range of a feature of data, i.e. a property, are reduced to scale between 0 and 1. Therefore, in order to calculate z, i.e. the normalized value of a member of the set of observed values of x, the following formula is used,

Formula used for normalization

Z = x-min(x) / max(x) – min(x)

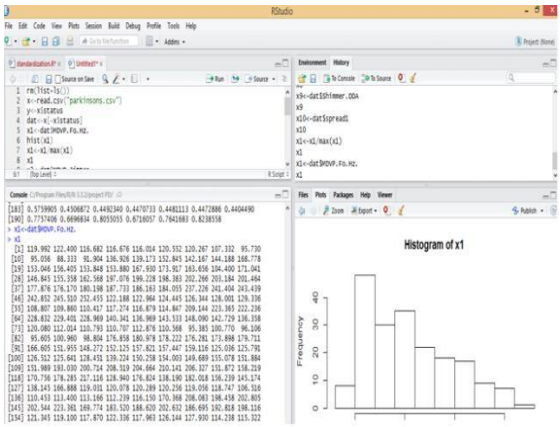
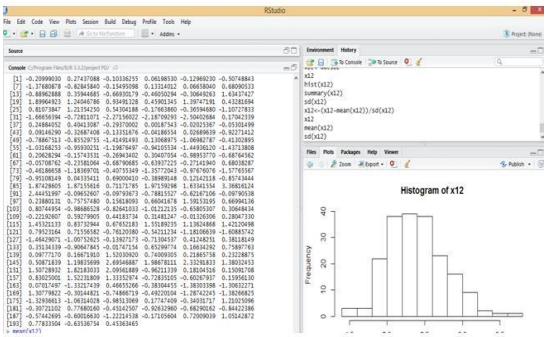


Figure 4 : Normalization

Figure 4.1 : Output of Normalization

The Output of Normalization Process is given in figure 4.1



D. PARAMETRIC MODELING

Parametric modeling is also known as White-Box model. A learning model that summarizes the data with the set of parameters of a fixed size is called as the parametric model.Parametric modeling is creating a model from some known facts about any population, these facts are called parameters. The parameters are used to find the mean and standard deviation. The standard normal distribution has the mean of zero (0) and standard deviation of one (1). A Parametric model captures all its information about the data within its parameters. For predicting a future data value from the current state of the model is done using its parameters. In the parametric model, all the parameters are in the finite dimensional parameter spaces. Many algorithms can be used in parametric modeling, in this paper the logistic regression algorithm has been used to analyze the classification accuracy of data. The main advantage of the parametric model is easier to understand and interpret the results and they do not require as much training data and can work well even if the fit to the data is not perfect.

1) . LOGISITIC REGRESSION

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating the probabilities. Logistic regression is a type of predictive model that can be used when the target variable is a categorical variable. In Figure 5, the coefficient of logistic regression algorithm must be estimated from the training data. This is done using the maximum-likelihood estimation. The maximumlikelihood estimation is the common learning algorithm used by variety of machine learning algorithms, although it does makes assumption about the distribution of the data. The best coefficient would result in the model that would predict the value which is very close to 1 for the default class and the value very close to 0 for the other class. The intuition for maximum-likelihood for logistic regression is that a search procedure seeks value for the coefficients that

minimizes the error in probability predicted by the model to those in the data.

Formula used to calculate the logistic regression:

Y = 1 / [1 + e^-(beta0 + beta1x)]

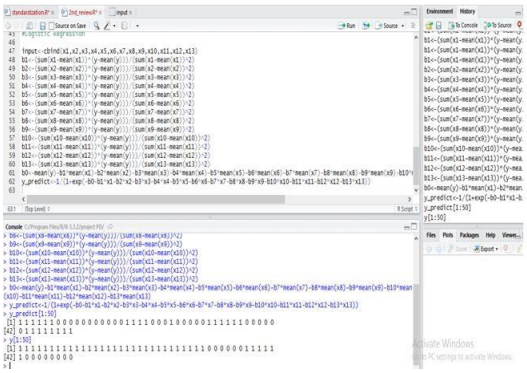


Figure 5 : Logistic regression

2). CLASSIFICATION ACCURACY FOR PARAMETRIC MODELING

Classification accuracy is the starting point. It is the number of correct predictions made divided by the total number of predictions made, multiplied by 100 to turn into a percentage. 80-85% classification is practically feasible as per reported in papers. In Figure 6, the classification accuracy for parametric model, has been predicted the Parkinson's data with the percentage of 82. As per survey on the logistic regression, the prediction on data proving the result the above 80% is considered as the efficient.

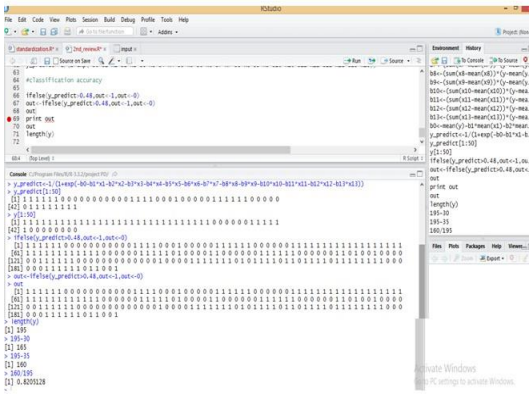


Figure 6 : Classification accuracy for Parametric model

E. NON PARAMETRIC MODELING

Non Parametric modeling is also known as Black-Box model. Algorithms that do not take strong assumption about the form of mapping function are called as the Non parametric model. This model does not rely on data belonging to any particular distribution. Non Parametric model is an error-minimization technique. These techniques do not assume that the structure of model is fixed. Non parametric methods seek to best fit the training data in constructing the mapping function, maintaining some

ability to generalize unseen data. The main advantage of Non parametric model is flexibility where it s capable of fitting a large number of functional norms and can result in high performance models for prediction.

1) *K-NEAREST NEIGHBORS ALGORITHM*

K-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. In Figure 7, the k-Nearest Neighbors algorithm that makes the prediction on the k most similar training patterns for the new data instance. The method does not assume anything about the form of the mapping function other than the patterns that are close likely have a similar output variable. The input consists of the K closest training examples in the feature space .K- Nearest Neighbors uses the voting system for the accurate result. The waited voting system should not go with the built-in function.

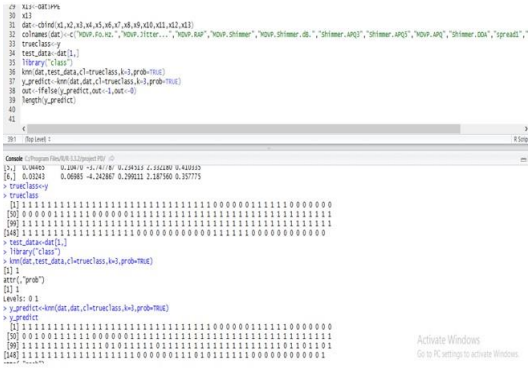


Figure 7, K-Neraest Neighbors Algorithm

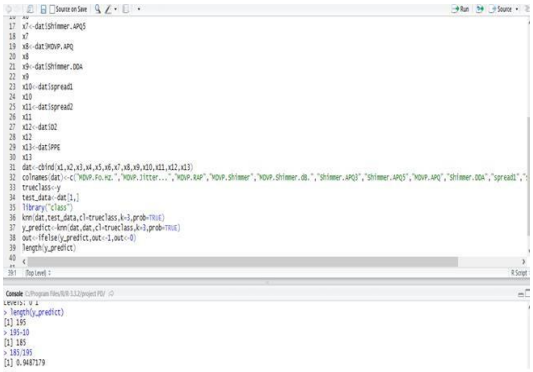


Figure 7.1 : Output for K – Nearest Neighbor Algorithm

2) *RANDOM FOREST ALGORITHM:*

Random Forest or Random Decision Forest is an ensemble learning method for classification and regression. In Figure 8, At each node: M variables are selected at random out of all M possible variables (independently for each node).The best splits on the selected m variable are found. Decision tree are individual learners that are combined. They are one of the most popular learning methods commonly used

for data exploration. One type of decision tree CART is used. CART is top-down binary, recursive partitioning, that divides the space into sets of disjoint rectangular regions. The region should be pure wrt response variable. Simple model is fit in each region where the majority votes for classification, constant value for regression. The major advantage of Random Forest algorithm, for many dataset it provides a highly accurate classifier and it runs efficiently on the large datasets. Provides an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

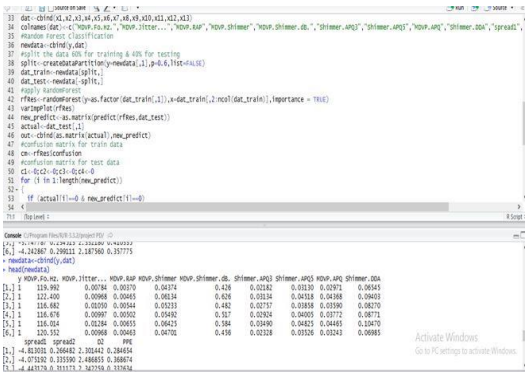


Figure 8 : Random Forest Algorithm

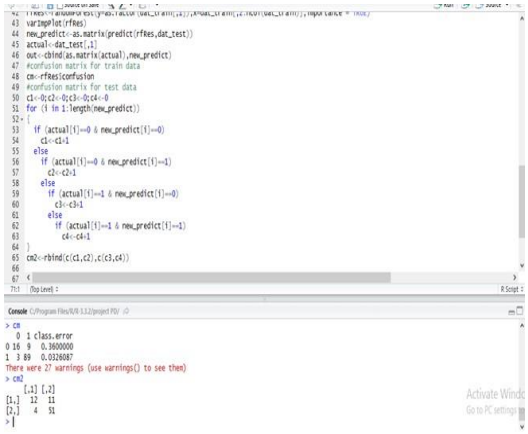


Figure 8.1 : Output for Random Forest Algorithm

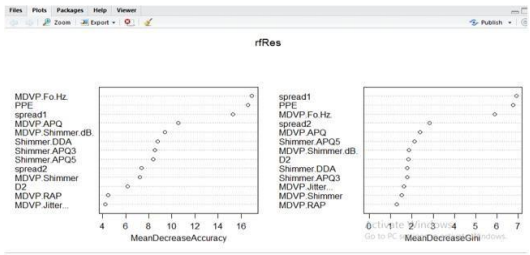


Figure 8.2 : Graphical Representation of Random Forest Algorithm

3) CLASSIFICATION ACCURACY FOR NON PARAMETRIC MODELING

The classification accuracy for Non parametric model for both the K-Nearest Neighbor algorithm and Random Forest algorithm is classified. For, K-Nearest Neighbor algorithm the classification accuracy is predicted with the percentage of 94. For, Random Forest algorithm the classification

accuracy is predicted with the percentage of 71. When comparing to both the algorithms of Non parametric model, the K-Nearest Neighbor algorithm provides the better classified output than the Random Forest algorithm.

F. COMPARING THE RESULTS OF PARAMETRIC AND NON PARAMETRIC MODELING

The comparison of result is made on both the Parametric and Non Parametric model with the basis of classification accuracy. Parametric model provided the classification accuracy of 82 percent, which is the efficient classified, resulted for the Parkinson's data. In the Non parametric model the classification accuracy is calculated and predicted with the two algorithms, K-Nearest Neighbor algorithm provided the classification of 94 percent and Random Forest algorithm provided with 71 percent. Thus the Parkinson's data works efficient on parametric modeling than comparing to the Non-parametric modeling.

V. CONCLUSION

From the results produced we conclude that parametric and non parametric model can be applied for any datasets. When compared with non parametric model frail variations are analyzed in parametric model according to our classification. In this paper, conclusion is made that efficiency of both parametric and non parametric models depends on the size of data and the number of attributes used for classification. Any dataset can be applied with any model for predicting the data, but only the certain model provide the efficient classified accuracy.

VI. FUTURE WORK

The further symptoms to predict Parkinson's disease can be analyzed using the data sets obtained. Provides better data for research people and doctor's who work with Parkinson's disease. Using the predicted classification accuracy of the Parkinson's data, the doctors and researchers would be benefited in diagnosing the disease and for the people around the world would gain the knowledge on the Parkinson's disease.

VII. REFERENCES

[1]Suitability of dysphonia measurements for tele monitoring of Parkinson's disease by Max A. Little, Member IEEE,[1] Patrick E. McSharry, SeniorMemberIEEE,[1] EricJ.Hunter,[2] JenniferSpielman,[2]

Lorraine[2,3] IEEE Eng. 2009 Apr; 56(4):
1015.
10.1109/TBME.2008.2005954

[2] Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection Max A Little*1,2,3,Patrick EMcSharry1,Stephen J Roberts2,Declan Moroz3 26 June 2007, **6**:23 doi: 10.1186/1475-925X-6-23

[3] Ambulatory Monitoring of Activities and Motor Symptoms in Parkinson's Disease by Daphne G.M.Zwartjes*, Tjitske Heida*, Member, IEEE*, Jeroen Pva G. Geelen,and Peter H. Veltink*, Senior Member, IEEE,* VOL. 57, NO. 11, NOVEMBER 2010.

[4] Rahn DA, Chou M, Jiang JJ, Zhang Y. Phonatory impairment in Parkinson'disease: Evidence from nonlinear dynamic analysis and perturbation analysis. J Voice. 2007;21:64–71. [PubMed]

[5] Cnockaert L, Schoentgen J, Auzou P, Ozsancak C, Defebvre L, Grenez F. Lowfrequency vocal modulations in vowels produced by Parkinsonian subjects. Speech Comm. 2008;50:288– 300.

[6] Zhang J, Small M. Complex network from pseudoperiodic time series: Topology versus dynamics. Phys Rev Lett. 2006;96 [PubMed]

[7] Zwirner P, Murry T, Woodson GE. Phonatory Function of Neurologically Impaired Patients. J Comm Disord. 1991;24:287–300. [PubMed]