Novel Parkinson's Disease Classification using Voiced Speech with Advanced Spectrograms, Convolutional Autoencoders and Machine Learning

Sai Bharadwaj Appakaya
Department of Electrical Engineering
University of South Florida
Tampa, FL, USA
saibharadwaj@usf.edu

Ravi Sankar
Department of Electrical Engineering
University of South Florida
Tampa, FL, USA
sankar@usf.edu

In-Ho Ra
School of Computer, Information and Communication Engineering
Kunsan National University
Gunsan, South Korea
ihra@kunsan.ac.kr

ABSTRACT

Research using speech from people with Parkinson's disease (PD) has shown a significant progress in classification studies. Due to the ease of acquisition and availability of established research protocols from applications like speaker recognition, speech processing has generated substantial research interest. A plethora of research studies have focused on developing objective classification models using sustained phonations and features developed using known signal processing methods. In this study, we focused on using connected speech with pitch synchronous segmentation and convolutional autoencoders to develop a model that can automatically extract the features and provide reliable classification. This methodology also aims at bypassing data availability issues by making use of standardized TIMIT dataset for training autoencoders. With Logistic regression and Linear SVM, we achieved 85% classification accuracy using the features from autoencoders. A mean accuracy of 84% is obtained under leave one subject out (LOSO) classification indicating the performance reliability for completely new data.

KEYWORDS

Speech Processing, Pitch Synchronous Segmentation, Parkinson's Disease, Convolutional Neural Network (CNN), Support Vector Machine (SVM).

1 INTRODUCTION

Research using speech for Parkinson's disease (PD) detection is a field of growing popularity due to the ease in data acquisition and processing. PD is a neurodegenerative disease that affects the production of dopaminergic neurons [1]. This results in many motor and non-motor symptoms like tremors, freezing of gait (FoG), dysarthria, dysphonia, sleep disorders, dementia, and depression [2]. The subjective rating scales like Unified Parkinson's disease rating scale (UPDRS) and Hoehn & Yahr (HY) scale play a significant role in diagnosis of PD. The disease cause, diagnosis methods and disease progression are all under research. Except for postmortem autopsy, there are no definitive methods that can verify a diagnosis [3]. The symptomatic research for PD majorly depends on analysis of data from wearable sensors for motor symptoms and perceptual studies for non-motor symptoms [4].

Speech production is a result of the synchronized coordination between cognitive and motor systems. As PD affects both systems, it results in progressive impairment of speech [5]. Analysis of Parkinsonian speech towards a diagnostic aid has been the active research area. Research groups working in this field have focused more on sustained phonations and traditional signal processing based features [6]. While some of these studies have shown a lot of progress in classification, other studies have shown the necessity for using connected speech as sustained phonations cannot represent natural speech and hence cannot yield reliable results [7, 8]. Research based on connected speech predominantly rely on perceptual ratings from trained listeners [9, 10]. Even though these studies include multiple listeners and show high interrater reliability, they have been criticized for the subjectivity and potential lack of reproducibility.

Studies based on mathematical modeling make use of features extracted from segments of speech samples that are of fixed length, typically close to 25ms. These features are meant to numerically quantify various psychoacoustic parameters which are used to train different machine learning and deep learning algorithms [11-13]. Research following this methodology have experimented with different features, learning algorithms and hyper-parameters to identify the optimal solution. A subset of these studies has focused on using deep neural networks to improve PD classification. Studies have focused on utilizing advanced convolutional algorithms to train neural networks on auditory spectrograms [14, 15].

In this study, we proposed a protocol for creating an advanced version of spectrogram which can output images of fixed dimensions despite the variations in the utterance duration in time domain. We used these advanced spectrograms with convolutional autoencoders (CAEs) to extract low dimensional features to be used for PD classification. Studies using autoencoders have made use of regular spectrograms or targeted dimensionality reduction in feature space [13-16]. The efficacy of this architecture was tested using logistic regression and linear support vector machines (SVM).

Voiced portions of the connected speech bounded by silence or unvoiced speech were analyzed in this study as it contains the merits of both sustained phonations and connected speech. They are close to sustained phonations in terms of spectral structure and represent the natural speech as they are extracted from connected speech. The voiced portion is transformed into regular and advanced spectrograms then fed to CAE for extracting the features. Studies that make use of spectrograms for similar application have made use of image resizing or audio clipping to maintain uniform dimensions across the dataset. The spectrograms have been resized to fit into a fixed dimension in this study.

In our previous studies, we showed that features extracted from pitch synchronous (PS) segments of the voiced portions in connected speech have higher probability in classifying PD [17, 18]. In this study, performance between PS segmentation and block processing has been compared. Due to the reduced availability of data, the CAE has been trained on voiced portions extracted from TIMIT database. The data used for PD analysis contains recordings of two Italian passages read by both people with PD and healthy control (HC). Next section contains further information regarding the datasets. Details regarding the methodology including pre-processing are discussed in Section III followed by results and discussion in Section IV and conclusion in Section V.

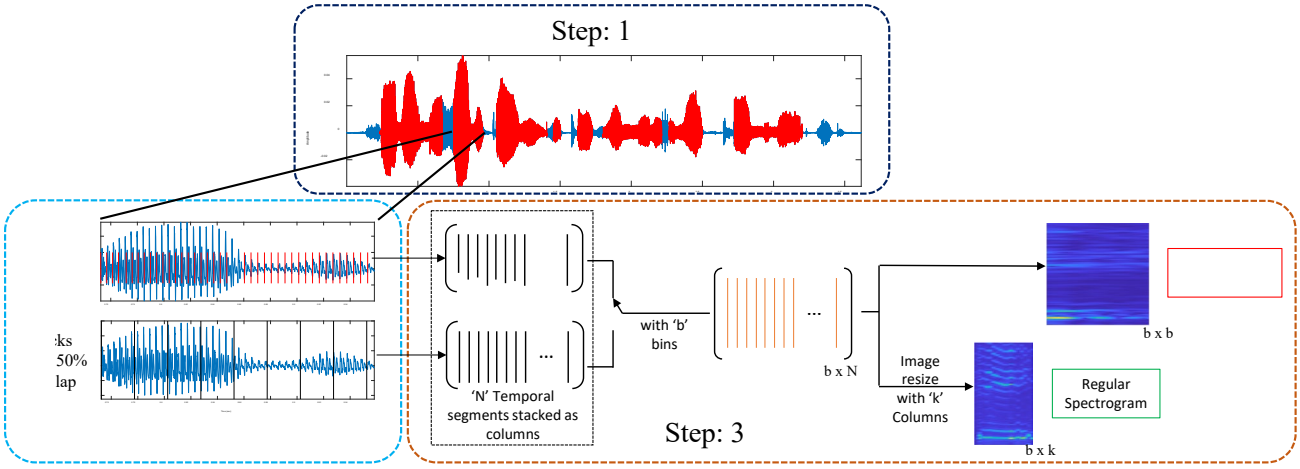

Figure 1: Pre-processing steps

2 DATA

Two datasets have been used in this study. The TIMIT Dataset has been used for training the CAE. This dataset contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States[19]. All the recordings have been made at 16 kHz in noise free conditions.

The PD dataset was collected by Giovanni *et al.* and analyzed using automatic Speech-to-Text system for PD classification. This dataset was made available on IEEE DataPort for research purposes [20]. It contains various speech tasks including reading of two phonetically balanced Italian passages by 50 subjects with 28 PD (19 male and 9 female) and 22 HC (10 male and 12 female). All the recordings were made at 16 kHz under noise-free conditions. None of the patients reported speech or language disorders unrelated to their PD symptoms prior to their study and they were receiving antiparkinsonian treatment [21]. The HY scale ratings for all the patients were < 4 except for two patients with stage 4 and one patient with stage 5. The duration of each passage recording varied between 1 to 4 minutes with a mean of 1.3 minutes.

3 METHODOLOGY

The methodology adopted in this study can be divided into four parts: Pre-processing, autoencoder training, feature extraction and classification.

3.1 Pre-processing

The pre-processing had been carried out in three steps for both datasets as shown in Fig. 1. In the first step, the voiced portions of passage recordings were extracted. These voiced portions were transformed into images in the second and third steps. In second step, each voiced portion was segmented using block processing or PS segmentation. In third step, magnitude spectrum of each block or PS segment is extracted. For regular spectrogram, the magnitude spectrum is resized to have 'k' columns. For, advanced spectrogram a second stage Fourier transform (FT) for each bin across the 'N' segments was calculated as shown in Fig. 2. In the Fig. 2, the matrix on left is magnitude spectra of time domain segments. The array of elements in each row represent the energy in the respective frequency bin for the time segments. Each row in the matrix on right represent the magnitude spectrum of the corresponding row in the left matrix. As the frequency resolution (number of bins) was kept the same, 'b', the right matrix becomes a square matrix with same number of rows and columns ('b x b'). The frequency for all the magnitude spectra was between 0 and 8 kHz (Nyquist frequency).

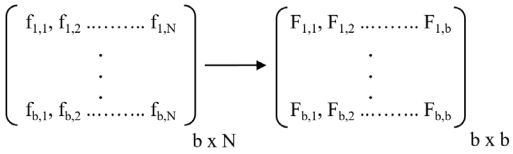

Figure 2: Advanced spectrogram creation from magnitude spectrum

In this way, both types of spectrograms are computed and saved as images for training autoencoders for each voiced portion.

3.2 Autoencoders

Autoencoders are very powerful tools that have been used for applications like dimensionality reduction and anomaly detection. They map a high dimensional input (x) to a low dimensional latent space (z) in bottle neck. This z is then used to reconstruct the original input. The first half when x is mapped to z is called the encoder ($g : x \rightarrow z$) and the second part where image is reconstructed is decoder ($f : z \rightarrow x$). The training is done based on a loss computed from the difference between the input image and its reconstructed counterpart. The z is used as a set of features representing each input for PD classification.

The autoencoders trained for this study contained 2 hidden layers in encoder and decoder. These layers contained 32 and 64 convolutional neurons, respectively. Experiments with a greater number of hidden layers and more neurons in each layer increased complexity but did not reduce loss. Leaky ReLU with $\alpha = 0.2$ was used as activation for all layers. autoencoders were used for dimensionality reduction to have the images represented by a set of 16 numbers as shown in Fig. 3. Fig. 4 and Fig. 5 show a sample voiced portion represented as advanced and regular spectrograms with both segmentation types. These figures also show the original inputs created along with their reconstructions.

3.3 Classification

After pre-processing, approximately 51000 images were created from TIMIT dataset. With two types of segmentations and two types of input images (spectrograms), a CAE was trained four different times. After CAE training, the images from PD dataset were processed and their respective 16-dimensional vectors were extracted. These vectors were used to train the two classifiers: Logistic regression and Linear SVM.

In logistic regression, the probability of an input belonging to a class is given by the equations:

$$h_\theta(z) = (\theta_0 + \theta_1 z_1 + \cdots + \theta_{16} z_{16}) \quad (1)$$

$$\Pr[1|\, z] = \frac{1}{1 + e^{-h_\theta(z)}} \quad (2)$$

The loss is calculated using

$$J(\theta) = \frac{1}{m}\left[\sum -y\, log\big(h_\theta(z)\big) + (1-y)\log\big(1 - h_\theta(z)\big)\right] \quad (3)$$

Where, m is the number of samples and 'y' is the correct labels.

In Linear SVM, a hyperplane is constructed in the feature space which has the largest distance from the data point that is closest in each class. The convergence condition is for minimization of the cost function given below:

$$\min_{\theta,\xi} \frac{1}{2}\theta^T\theta + C\sum \xi \quad (4)$$

Where, ξ is the slack variables which penalizes data points which violate the margin requirements and θ is the weight vector

The PD dataset yielded approximately 12000 datapoints/images. These images were split into training and testing groups with 80% and 20% of the data, respectively. To avoid conclusions drawn over single anomalies, the classification step was repeated 10 time and the mean values of the metrics are used for discussion.

4 RESULTS AND DISCUSSION

In all the four combinations, the autoencoders converged after around 40 iterations/epochs while trained upon TIMIT dataset. These encoder models delivered good reconstruction when images from PD dataset were used. Fig. 4 shows a sample image from PD dataset reconstructed using their corresponding autoencoders. For the regular spectrograms, x-axis is the frequency axis.

The overall results from this study are given in Table 1. This includes the mean of the test accuracies obtained over 10 different iterations. Both classifiers are trained using data from male and female groups individually and then combined. Across the different epochs, the variance in the results was negligible with a standard deviation ranging between ±2%. Between logistic regression and linear SVM, performance variance is minimal ranging between ±1.5%.

Table 1: Mean Accuracies

		Logistic Regression			Linear SVM		
		M	F	Both	M	F	Both
Regular	Block	81.87	71.56	77.52	82.28	73.15	78.3
	PS	78.68	71.6	74.08	78.26	73.32	74.5
Advanced	Block	89.27	79.39	80.77	89.04	79.18	80.53
	PS	**89.36**	**85.67**	**85.07**	**89.12**	**86.13**	**85.06**

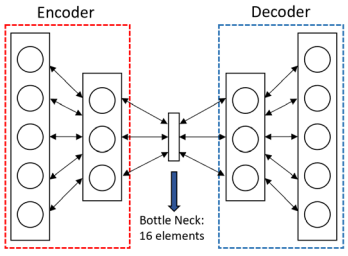

Figure 3: Autoencoder architecture

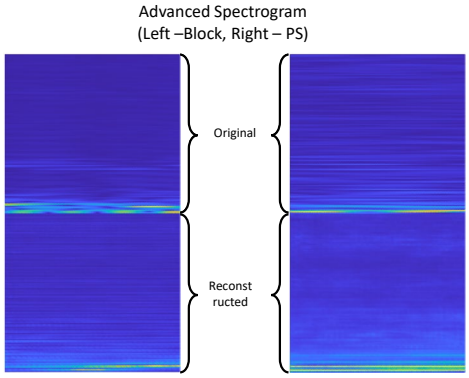

Figure 4: Advanced spectrogram example

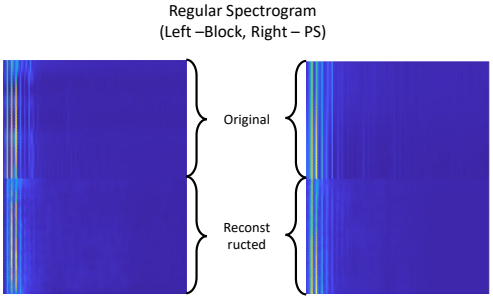

Figure 5: Regular spectrogram example

Table 2: Performance Comparison

	Logistic Regression			Linear SVM		
	M	F	Both	M	F	Both
	PS - Block					
Regular	-3.192	0.124	-3.442	-4.018	0.166	-3.798
Advanced	0.092	**6.278**	**4.308**	0.084	**6.948**	**4.532**
	Advanced-Regular					
Block	7.394	7.828	3.248	6.758	6.026	2.232
PS	**10.678**	**13.982**	**10.998**	**10.86**	**12.808**	**10.562**

From the results, advanced spectrograms with PS segmentation has better result performance accuracy. Performance comparison between PS and block processing is done in Table 2. Difference between accuracy from PS segmentations and block processing for regular and advanced spectrograms shows that PS segmentation has better performance with positive values. Comparison between regular and advanced spectrograms is shown in last two rows of Table 2. With block processing or PS segmentation, advanced spectrograms have better performance which is denoted by the positive values. Overall, PS segmentation with advanced spectrograms yielded the best performance. The ROC curves for logistic regression and linear SVM are shown in Fig. 6.

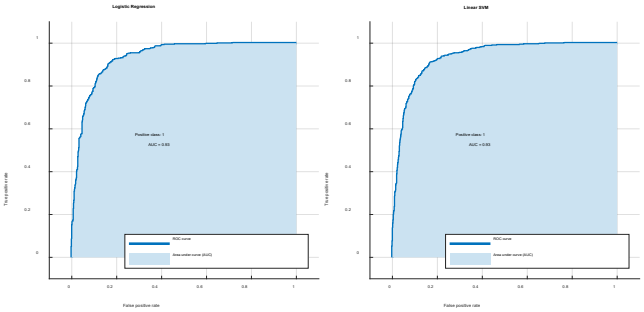

Figure 5: ROC curves for both classifiers

To further evaluate the model performance with PS segmentation and conditions, leave one subject out (LOSO) analysis was evaluated. LOSO analysis is conducted for every PD subject individually. For each one of the 28 PD subjects, a random HC subject was chosen and set aside as test data. For training, a balanced dataset containing 21 randomly chosen PD out of the remaining 27 and remaining 21 HC were used. The accuracies identified for each one of the 28 subjects is shown in Fig. 7. Classification accuracy was identified to be more than 75% for 26 subjects out of 28.

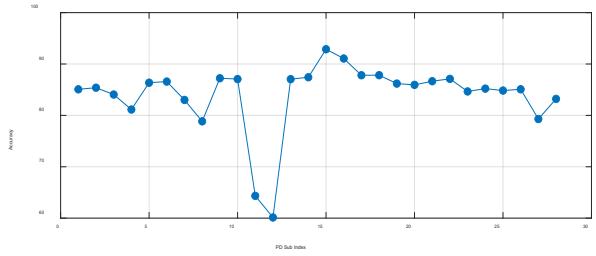

Figure 6: Subject-wise classification accuracies from LOSO analysis

5 CONCLUSIONS

In this study, an advanced spectrogram has been proposed and tested using connected speech samples collected from 50 subjects. Similar images extracted from a standardized speech database have been used to train autoencoders that can map the image to 16 dimensional vectors. These images from PD dataset have been mapped to 16 dimensional vectors, which were used as inputs to train logistic regression and linear SVM classifiers to classify between PD and HC classes. The performance has been compared between regular and advanced spectrograms with pitch synchronous and block processing. The results showed that advanced spectrogram created using PS segmentation had better classification performance than advanced spectrogram created using block processing and regular or advanced spectrograms created using block processing. For both classifiers, classification accuracy of 85% is observed while the area under ROC was 0.93. The LOSO analysis conducted to test the efficacy with a subject's data completely left out of training resulted in more than 75% accuracy for 26 subjects and more than 80% accuracy for 24 subjects out of 28 subjects. Overall, this analysis showed that the autoencoder based features extracted from advanced spectrogram has better chance of classifying between PD and HC. Future studies have been planned to include more data and test the classification performance using data collected by different research groups. Similar analysis using sustained phonations was also planned for further investigation.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2014333).

REFERENCES

[1] H. Tohgi, T. Abe, and S. Takahashi, "Parkinson's disease: diagnosis, treatment and prognosis," *Nihon Ronen Igakkai zasshi. Japanese journal of geriatrics,* vol. 33, no. 12, pp. 911-915, 1996.

[2] M. MS Lima *et al.*, "Motor and non-motor features of Parkinson's disease–a review of clinical and experimental studies," *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders),* vol. 11, no. 4, pp. 439-449, 2012.

[3] T. G. Beach and C. H. Adler, "Importance of low diagnostic accuracy for early Parkinson's disease," *Movement Disorders,* vol. 33, no. 10, pp. 1551-1554, 2018.

[4] M. Bachlin *et al.*, "Wearable assistant for Parkinson's disease patients with the freezing of gait symptom," *IEEE Transactions on Information Technology in Biomedicine,* vol. 14, no. 2, pp. 436-446, 2009.

[5] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behavioural neurology,* vol. 11, no. 3, pp. 131-137, 1998.

[6] J. S. Almeida *et al.*, "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognition Letters,* vol. 125, pp. 55-62, 2019.

[7] R. I. Zraick, T. M. Dennie, S. D. Tabbal, T. J. Hutton, G. M. Hicks, and P. S. O'Sullivan, "Reliability of speech intelligibility ratings using the Unified Parkinson Disease Rating Scale," *Journal of Medical Speech-Language Pathology,* vol. 11, no. 4, pp. 227-241, 2003.

[8] F. Klingholtz, "Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels," *The Journal of the Acoustical Society of America,* vol. 87, no. 5, pp. 2218-2224, 1990.
[9] S. Anand and C. E. Stepp, "Listener perception of monopitch, naturalness, and intelligibility for speakers with Parkinson's disease," *Journal of Speech, Language, and Hearing Research,* vol. 58, no. 4, pp. 1134-1144, 2015.
[10] C. Kuo and K. Tjaden, "Acoustic variation during passage reading for speakers with dysarthria and healthy controls," *Journal of communication disorders,* vol. 62, pp. 30-44, 2016.
[11] C. Che, C. Xiao, J. Liang, B. Jin, J. Zho, and F. Wang, "An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease," in *Proceedings of the 2017 SIAM International Conference on Data Mining*, 2017, pp. 198-206: SIAM.
[12] V. J. Kadam and S. M. Jadhav, "Feature ensemble learning based on sparse autoencoders for diagnosis of Parkinson's disease," in *Computing, Communication and Signal Processing*: Springer, 2019, pp. 567-581.
[13] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and E. Nöth, "Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 717-720: IEEE.
[14] A. Lauraitis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Detection of Speech Impairments Using Cepstrum, Auditory Spectrogram and Wavelet Time Scattering Domain Features," *IEEE Access,* 2020.
[15] L. Zahid *et al.*, "A Spectrogram-Based Deep Feature Assisted Computer-Aided Diagnostic System for Parkinson's Disease," *IEEE Access,* vol. 8, pp. 35482-35495, 2020.
[16] B. Karan, S. S. Sahu, and K. Mahto, "Stacked auto-encoder based Time-frequency features of Speech signal for Parkinson disease prediction," in *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, 2020, pp. 1-4: IEEE.
[17] S. B. Appakaya and R. Sankar, "Classification of Parkinson's disease Using Pitch Synchronous Speech Analysis," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 1420-1423: IEEE.
[18] S. B. Appakaya and R. Sankar, "Parkinson's Disease Classification using Pitch Synchronous Speech Segments and Fine Gaussian Kernels based SVM," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 236-239: IEEE.
[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *STIN,* vol. 93, p. 27403, 1993.
[20] D. Giovanni and G. Francesco. "Italian Parkinson's Voice and Speech" [Online]. Available: http://dx.doi.org/10.21227/aw6b-tg17
[21] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system," *IEEE Access,* vol. 5, pp. 22199-22208, 2017.