Stacked auto-encoder based Time-frequency features of Speech signal for Parkinson disease prediction

Biswajit Karan, Sitanshu Sekhar Sahu, and Kartik Mahto
Department of Electronics and Communication Engineering, Birla Institute of technology, Mesra, Ranchi

Abstract: **Proper classification between normal and Parkinson affected people is an important topic in recent years. From the last two decades, the number of methods has been proposed for the classification of Parkinson's affected and healthy people. Most of them based on a shallow structured network classifier. In this study stacked auto-encoder deep neural network framework is introduced to classify Parkinson affected and healthy people voice signals. The present study uses a spectrogram and scalogram of speech signals as input to the stacked autoencoder deep network. The extracted features are tested with a support vector classifier (SVM) and a Softmax classifier. Highest classification accuracy of up to 87 % with a spectrogram and 83 % with scalogram are obtained using Softmax classifier. The softmax classifier performed better than SVM. The proposed deep neural network may be a new window for further research.**

Index term: ***Parkinson disease, stacked auto encoder, Time-frequency image, STFT, CWT, deep learning.***

1. INTRODUCTION

PD is a neurodegenerative disorder privileged in old age. Millions of people worldwide affected by this disease [1]. The people with PD have less control to utter their voice. In several aspects, the voice signal is affected. Their voice becomes more slurred, breathy, hoarse, or softer. In the medical field, these speech changes are defined in terms of dysarthria (motor speech), hyphophonia (weak voice) and tachyphenia(fast-talking). The PD affected voice signal lost their irregularity. This effect can be detected using TF features. Several studies have been conducted for the assessment of PD people. In this study, various speech signal processing algorithm has been utilized to extract useful features for PD assessment. These features and various machine learning algorithms directly affect the accuracy and reliability of the PD diagnosis system. Most of the studies are based on the start of art acoustic features [2, 3, 4], spectral and cepstral features[5,6], Empirical mode decomposition based features, various and wavelet features [7,8]. From all previous studies, it has been found the most of the study based on the time domain, frequency domain, or spectral and cepstral domain. A very view study has been conducted based on TF frequency representation. T.Villa-Canas et al [9] used TF based technique based on Wigner Ville distribution and reported classification accuracy 72 % between Parkinson and healthy speakers. Convolution neural network approach is presented by J. C. Vasquez-Correa et al.[10] in three different languages. Autoencoder based TF features concept for PD detection is introduced the first time in this study. We proposed TF autoencoder based features which effectively captures both time and frequency domain aspect of the speech signal. The proposed autoencoder based TF features are extracted using STFT and CWT based TF representation (TFR). The extracted features from TFR representation are fed into the Softmax classifier. We also compare the extracted features using, SVM. According to the results, the proposed approach provides the accuracies up to 87 % as compared to the SVM classifier. The study is steps towards the robust classification of speech impairment.

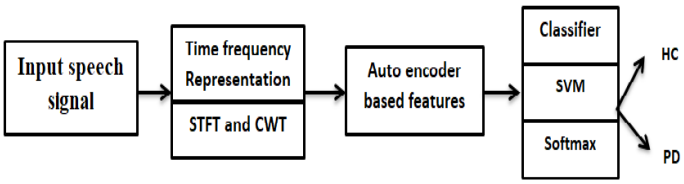

Figure1: Proposed methodology

2. MATERIALS AND METHODS

Data Sources: In this work PC-GITA database [9] is used which consists of a voice sample of 50 PD and 50 normal people. All speakers are Colombian native speakers. The database is well balanced in terms of age and gender. The speakers performed different speech tasks, sustained phonation, word, and monologue, read a text, sentence and

Biswajit Karan is with the Birla Institute of Technology, Mesra, Ranchi India (phone: 09470549164; e-mail: bkaranetc@gmail.com).

Sitanshu Sekhar Sahu is with the Birla Institute of Technology, Mesra, Ranchi, India (e-mail sitanshusekhar@gmail.com).

Kartik Mahto is with the Birla Institute of Technology, Mesra, Ranchi India (e-mail: kartik_mahto@rediffmail.com).

978-1-7281-4458-0/20/$31.00 ©2020 IEEE

rapid repetition of /pa/, /ta/, /ka/. In this study sustained phonation of vowels and words is considered. All total 1500 utterances (750 PD, 750 healthy) are considered for vowels (each speaker uttered five vowels three times) and 2500 utterances (1250 PD, 1250 healthy) are considered for words (25 words each speaker).

Time-frequency representation:

Short time Fourier transformation (STFT):

It is a time-frequency analysis method based on Fourier transformation. Here the signal is segmented by multiplying signal with a translated version of the window function. Then Fourier analysis is performed on each segment. The short-time Fourier transformation of signal p(t) is given by

$$f(\tau,\omega) = \int_{-\infty}^{+\infty} \mathrm{p}(\tau).\omega(\tau - \mathrm{t})\, \mathrm{e}^{-j\omega\tau}\, \mathrm{d}\,\tau \qquad (1)$$

Where $\omega(\tau - \mathrm{t})$ -window function

p(t)- speech signal

From Figure-2 it is observed that, in Parkinson's speech spectrogram, the periodicity is destroyed disordered voice. It is noticed in the spectrogram that dominant frequency components are shifted across the high-frequency region due to reduced quality in speech signal.

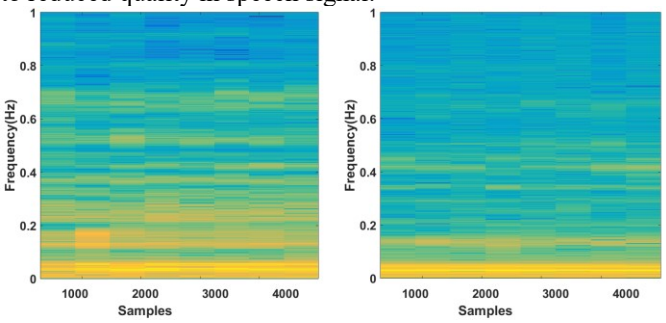

Figure 2: Spectrogram of Parkinson's affected and healthy speech.

The visible difference can be noticed in the spectrogram of Parkinson and healthy speech.

Continuous wavelet transformation (CWT)

To achieve good time and frequency resolution it might be better to use different types of tile at low and high frequency. Such representation can be achieved using wavelet transformation. It is similar to STFT, but instead of using a fixed window, the variable window size is used. The speech signal is decomposed into 256 scales of step size. Morlet wavelet is used to plot the scalogram because it is closed to human hearing perception [11]. Scalogram has the ability to represent distinguish characteristics of normal and disordered voice into a more precise TF plane.

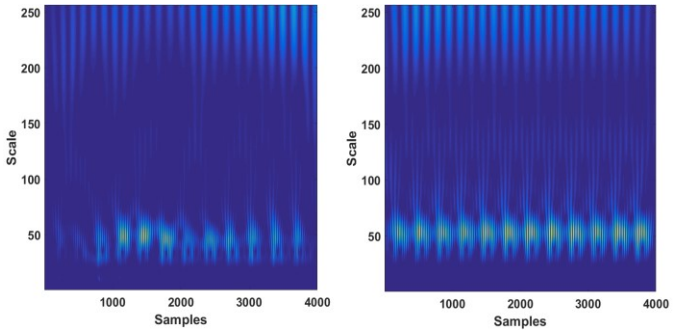

Figure 3: Scalogram of Parkinson affected and healthy speech

Since the Parkinson affected speech signal is seriously disturbed because of the presence of amplitude and fundamental frequency perturbation. In figure -3, the scalogram of Parkinson's affected speech shows more randomness in the energy distribution of spectral sub-band than healthy speech.

Stacked auto-encoder (SAE)

SAE is an unsupervised feature learning neural network with three layers, namely, the input layer that represents inputs, the hidden layer that represents learned features, and the output layer with the same dimension of the input layer that represents a reconstruction of same inputs [12].

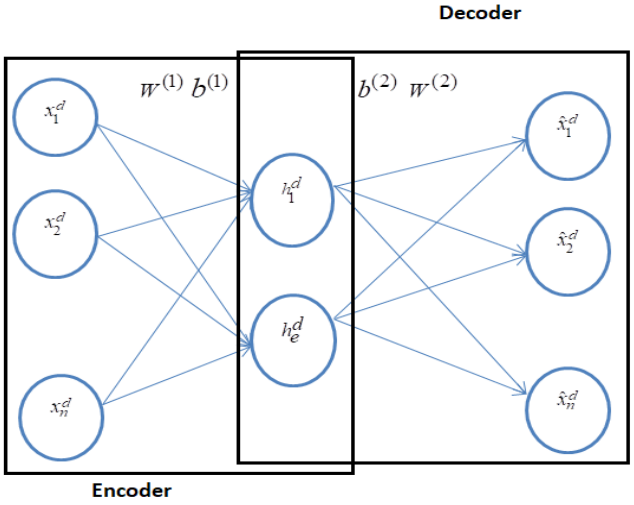

Figure 4: Basic encoder and decoder

The input and hidden layers form the encoder network responsible for transforming the original inputs into hidden representation codes, whereas the hidden and output layers form the decoder network responsible for reconstructing the original inputs from the learned hidden representation shown in figure-4.

3. Proposed Experimental framework

Proposed Set up is carried out using two-component the auto-encoder and Softmax classifier as shown in figure 4. The original speech signal is converted into a time-frequency

image using STFT and CWT. Before giving output the TF image to the encoder, images are resized to 28 x 28 pixels. Then TF images are fed through the auto-encoder network to extract high-level features. As shown in figure-4, the Softmax classifier is placed after the second auto-encoder. The stacked auto-encoder in combination with the Softmax classifier provides better performance. The SAE approach disadvantageous in terms of time consumption but shows high performance in the testing phase because of high-level features provided by the last encoder.

Classification of the speech signal with the Softmax classifier and with two encoders is implemented with the following steps.

Step 1: Divide the dataset into 80 % training and 20 % testing. During splitting no speaker data in training set is included in testing and vice versa.

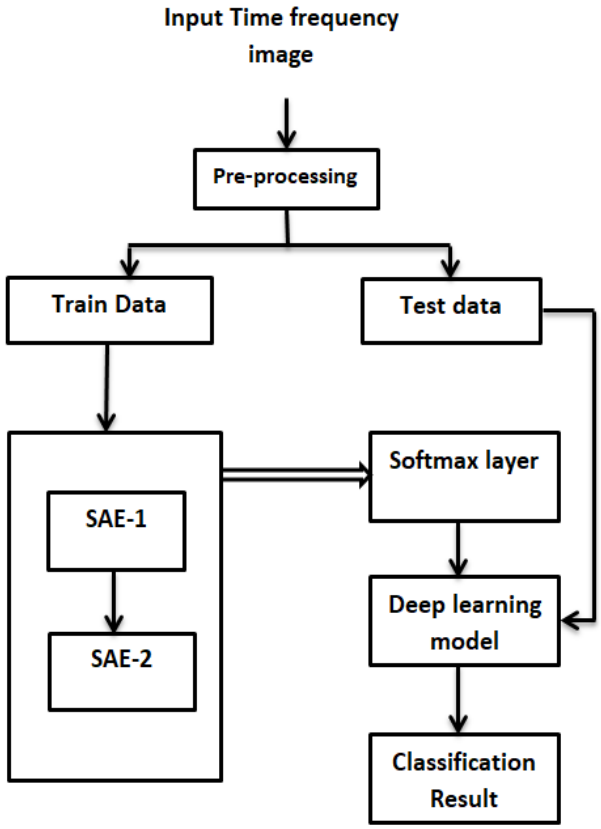

Figure 4: Flow chart of the proposed TF SAE network

Step 2: The training dataset is given to the first encoder and trained.

Step 3: The output of the hidden layer of the first encoder is fed to the second auto-encoder and trained as step 2.

Step 4: The output values of the second SAE are given to the Softmax classifier.

Table1: Parameter used to train the stacked auto-encoder

Meta-parameter	Value
Autoencoder-1 hidden size	**500**
Autoencoder-2 hidden size	**250**
Epoch	**200**
Sparse regularization	**4**
L2 Regularization	**0.004**

4. RESULTS AND DISCUSSION

Experiments are performed as shown in figure 4 using the TF image of the speech signal. 80 % data is used for training the model and the rest 20 % has used the data for testing the model. The proposed SAE deep learning framework performed well because of interlayer communication. Second, it is more effective in the presence of noise. The study is done in MATLAB 2017b environment. The parameters used for SAE are mentioned in table 1. The dataset used in this study was divided into training and testing data. Hold out validation is used to make the standard evaluation. Each encoder in the model was trained using sparse regularization parameters for optimization, and L2 regularizations parameter to preventing the over fitting.

Table 2: Experimental results for TF image classification in terms of Accuracy (%) using SVM and Softmax classifier.

TFR	SVM classifier	Softmax classifier
STFT	83 %	87 %
CWT	78.93 %	82 %

Table 2, shows that TF based stacked autoencoder framework results using STFT and CWT image. First, the SAE based TF features are tested with a Support vector machine classifier. 83 % and 78.93 % accuracy is obtained using STFT and CWT respectively. Further, the accuracy is improved by using the Softmax classifier. STFT based TF representation gives better accuracy as compare to CWT.

Figure 5:Histogram of classification accuracy comparison.

The effectiveness of the TF feature with the softmax classifier is shown in figure 5. From the result, it is inferred that TF

features obtained from STFT representation are superior to CWT. Softmax classifier possesses good stability in both types of TFR.

5. CONCLUSION

A robust approach for PD prediction is proposed in this study. In order to classify speech signals of PD and healthy people, a stacked auto-encoder based approach is implemented using the TF image. The highest accuracy 87 % is obtained using SAE based TF features and softmax classifier. Among two TFR, the spectrogram performed better than the scalogram. From the results, it is inferred that softmax classifiers better distinguish the speech sample of Parkinson's affected and healthy people. The proposed approach may be a good alternative for the prediction of Parkinson's disease.

6. Acknowledgment

This work has been carried out under the sponsorship of DST-SERB under grant no- ECR/2017/000345.

7. References

1. Roberts, Angela, and Danielle Post. "Information content and efficiency in the spoken discourse of individuals with Parkinson's disease." Journal of Speech, Language, and Hearing Research 61.9 (2018): 2259-2274.

2. Little, MA, McSharry PE, Hunter EJ, Spielman J, Ramig, LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng. 2009; 56(4):015-1022.

3. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel Speech Signal Processing Algorithms for High Accuracy Classification of Parkinsons Disease. IEEE Trans Biomed Eng 2012; 59:1264–71.

4. Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, et al. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE J Biomed Heal Informatics 2013; 17:828–34.

5. Orozco-Arroyave JR, Hönig F, Arias-Londoño JD, Vargas-Bonilla JF, Nöth E. Spectral and cepstral analyses for Parkinson's disease detection in Spanish vowels and words. Expert Syst 2015;32:688–97.

6. Orozco-Arroyave JR, Hönig F, Arias-Londoño JD, Vargas-Bonilla JF, Daqrouq K, Skodda S, Rusz J, Nöth E. Automatic detection of Parkinson's disease in running speech spoken in three different languages. The Journal of the Acoustical Society of America. 2016 Jan;139(1):481-500.

7. Karan, Biswajit, Sitanshu Sekhar Sahu, and Kartik Mahto. "Parkinson disease prediction using intrinsic mode function based features from speech signal." Biocybernetics and Biomedical Engineering (2019).

8. Tsanas, Athanasios, et al. "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity." Age (years) 64.8.1 (2010): 63-6.

9. Orozco-Arroyave JR, Arias-Londoño JD, Vargas-Bonilla JF, González-Rátiva MC, Nöth E. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. Lr 2014 Proc Ninth Int Conf Lang Resour Eval 2014:342–7.

10. Villa-Cañas, T., et al. "Time-frequency approach in continuous speech for detection of Parkinson's disease." 2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA). IEEE, 2015.

11. Vásquez-Correa, Juan Camilo, Juan Rafael Orozco- continuous speech for detection of Parkinson's disease." 2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA). IEEE, 2015.

12. Vásquez-Correa, Juan Camilo, Juan Rafael Orozco-Arroyave, and Elmar Nöth. "Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease." INTERSPEECH. 2017.

13. Adem, Kemal, Serhat Kiliçarslan, and Onur Cömert. "Classification and diagnosis of cervical cancer with softmax classification with stacked autoencoder." Expert Systems with Applications 115 (2019): 557-564.