2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)
4-7 March 2018
Las Vegas, Nevada, USA

DeepVoice: A Voiceprint-based Mobile Health Framework for Parkinson's Disease Identification

Hanbin Zhang[1], Aosen Wang[1], Dongmei Li[2] and Wenyao Xu[1]

***Abstract*— Parkinsons disease (PD) identification has attracted a lot of attention in recent years. However, there is still no standardized and convenient way to identify PD, because most researchers are only focusing on promoting identification accuracy. With the recent development of mobile health, a feasible mobile application to achieve PD identification is highly demanded with a small amount of information but can provide reliable results. To this end, we propose DeepVoice, a voiceprint-based PD identification application simultaneously integrating deep learning and mobile health. DeepVoice works by collecting a short period of monosyllabic voice through a mobile health App on a smartphone. Specifically, we propose the Joint Time-Frequency Analysis algorithm to enhance the voiceprint feature in spectrogram domain. We also develop a customized convolutional neural network (CNN) to complete the final identification. We evaluate our proposed DeepVoice on input data format, the length of the input voice and neural network architecture in a large PD dataset. Experimental results show DeepVoice could successfully achieve PD identification with an accuracy of 90.45±1.71% with only 10 seconds long audio segment. Our study also reveals that the smartphone-based mobile health application is feasible for PD identification.**

I. INTRODUCTION

Parkinsons disease (PD) has been traditionally characterized by the motor symptoms of the disease such as tremor, altered gait, and bradykinesia, with a well-established pathophysiology related to loss of dopaminergic neurons in substantia nigra [1]. Even though there is no cure for PD at present, there exists a variety of medications provide dramatic relief from the symptoms [2]. In fact, early detection and treatment play an important role in the process of treating PD. It not only helps patients lessen the pain but reduces complications to extend their lifespans. There are many cases misdiagnosed as spondylolysis and respiratory disease in early diagnosis. For example, Muhammad Ali, one of the most famous cases of PD, took four years to diagnose in his early life, therefore, delayed the best treatment period [3]. The reason behind is PD identification is complex and not widespread. People may not pursue a professional PD diagnose when they feel minor throat or muscle irritation. Consequently, a convenient and reliable PD identification application is urgently needed but still underexplored.

There are several related works applying artificial neural work into PD identification in recent two years. Pereira et al. [4] proposed to use the state-of-the-art CNN architecture after recording participants handwritten dynamics from the smartpen. Eskofier et al. [5] proposed to use the sensors to record the participants mobility pattern during a specific mobility task supervised by a movement disorder specialist. The common feature of the papers mentioned above is they consider accuracy as the first priority and ignore the accessibility of data acquisition process. To some extent, their data collection method is even as complex as some professional way: They either need special instruments [4] to support an experiment or an experiment needed to be supervised by a specialist [5]. Such complex data collection method also results in the small-scale dataset: the dataset in [4] and [5] contains hundreds of cases and [6] contains tens of cases. Therefore, their methods are not convenient and reliable in practical applications, even though their results seem being improved.

Fortunately, the widespread of the smartphones provide us this opportunity today. We propose a voiceprint-based PD identification application, named DeepVoice, which integrates deep learning technique and mobile health. Different from the methods in the literature, our DeepVoice adopts a smartphone system to collect and process data. Instead of using movement information or speech material, DeepVoice works by collecting a short segment of monosyllabic voice "Aaaaaah" as the input due to the two key reasons: **1).** Tremor is the most obvious symptom in PD but about 30% of patients whose symptom is not very obvious in early stage [7]. Such tremor caused by muscle contraction disorders may not be observed by human eyes but might be well captured by smartphone when a person is phonating using his throat muscles. **2).** PD patients always have a slow evolvement in language barrier in early stage but their volume and intonation will change first. Because there is no explicit model could tell how many features such kind of audio data will have, we design and implement a customized CNN in the cloud server to finish classification. Specifically, after audio data is collected and uploaded to the cloud server, a Joint Time-Frequency Analysis algorithm is applied to enhance the features. Those features, we also call voiceprint, will be sent to the input of the CNN in the form of the spectrogram. In the end, the cloud server will feedback the result, PD or not PD, back to the user.

Our contributions are summarized as three-fold:

- We develop DeepVoice, a voiceprint-based PD identification application combining both Deep Learning and mobile health.
- We evaluate our proposed DeepVoice in a large PD dataset.
- We perform extensive experiments to validate our Deep-

[1]H. Zhang, A. Wang and W. Xu are with the Department of Computer Science and Engineering, University at Buffalo (SUNY), Buffalo, NY 14260, USA.

[2]D. Li is with the School of Medicine and Dentistry, University of Rochester Medical Center, Rochester, NY 14627, USA

978-1-5386-2405-0/18/$31.00 ©2018 IEEE

Voice design. The relationship between identification accuracy and length of input audio can even provide us the insight to further improve the user experience.

The rest of the paper is organized as follows. In section II, we introduce our data collection method, data preprocessing algorithm, the proposed CNN architecture and present the workflow of DeepVoice. In section III, we evaluate our DeepVoice and analyze the results. Last, we conclude our work and highlight future directions in Section IV.

II. METHOLOGY

A. Data Preprocessing

We apply a mobile health application [8] on Apple APP Store to record the voice data. In particular, each participant is asked to phonate "Aaaaah" for 10 seconds and each audio sample is recorded by smartphone using a sampling rate at 44.1Kbps. Generally, this sampling rate is too large to become an input for CNN, because a ten seconds long audio will even contain about 441000 samples in total.

One basic alternative is to directly perform downsampling and reshape the data into an appropriate size as input. The motivation behind is the frequency band of the voice signal and the tremor are generally 300∼3400Hz and 5∼7Hz respectively. Therefore, the sample rate 44.1Kbps is much higher than what we need. However, the disadvantage of this method is that we will lose some unknown but significant knowledge in high frequency that plays an important role in PD identification.

Joint Time-Frequency Analysis is another promising method. It can output the spectrogram visualization as the input of CNN. Concretely, Discrete-time STFT (Short-time Fourier transform) [9] can be applied to derive the time-frequency graph. We divide audio data into 20 chunks and perform Fourier transform to each one. The complex result after Fourier transform is added into a matrix, which records magnitude and phase for each point in time and frequency, which can be expressed as:

$$STFT\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}. \quad (1)$$

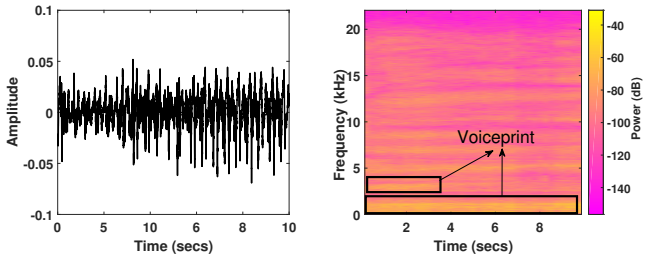
The magnitude squared of the STFT yields the spectrogram as:

$$\text{spectrogram}\{x(t)\}(m,\omega) \equiv |X(m,\omega)|^2. \quad (2)$$

This method can preserve more high frequency information and present participant's voiceprint better, as shown in Fig. 1. We will evaluate two architecture in section III.

B. Customized CNN Architecture

The network architecture of DeepVoice refers to the design from AlexNet [10]. For our first trial, we apply AlexNet as a whole into our application but the result seems not good. Specifically, the accuracy in the training set is much higher than the one in the testing set, which is called overfitting. The prominent reason is the number of features in our spectrogram is limited compared to those pictures used in Large Scale Visual Recognition Challenge (ILSVRC) [11].



(a) The Time Domain image for 10 seconds long audio data. (b) The Spectrogram for 10 seconds long audio data.

Fig. 1. A comparison between different input data.

To this end, we propose to customize the network architecture by scaling down the number of convolutional layers and the filters in each convolutional layer to suppress the overfitting issue. We also remove the first two fully connected layers which bring severe overfitting risk. In addition, even we do not find the pooling layer make a big influence on our result (no more than 3%), a configuration 2×2 with stride 2 gives us the best result.

We implement and verify our customized CNN using the Neural Network Toolbox from Matlab. The proposed architecture is shown in Fig. 2, which consists of 2 convolutional layers, 2 ReLU layers, 1 maxPooling layer, 1 fully connection layer and 1 softmax output layer.

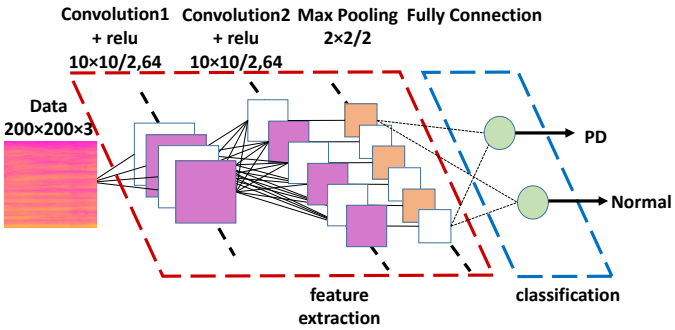


Fig. 2. The Architecture of CNN in DeepVoice: It includes 2 convolutional layers, 2 ReLu layers, 1 maxpooling layer, 1 fully connection layer and 1 softmax output layer.

C. DeepVoice Framework

Fig. 3 shows the entire framework of DeepVoice. It is consist of two parts: Data Collection module at the user-end and Classification module at the serve-end. Data collection module comprises a smartphone and a mobile health application to collect users' data. The Classification module is a server on the cloud to finish the data preprocessing and classification.

Above all, DeepVoice works as follows: To begin with, a mobile health application is running on the smartphone to record the "Aaaaah" voice phonated by the participant; A Joint Time-Frequency Analysis is applied after the data is sent to the server; In the end, a pre-trained customized CNN

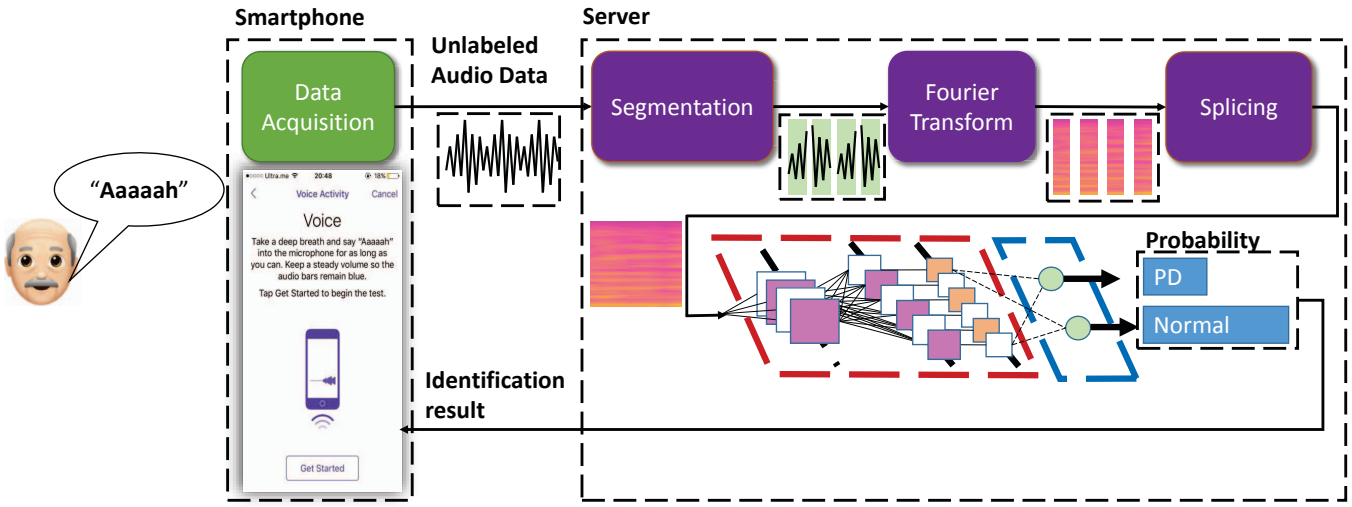


Fig. 3. The DeepVoice framework includes a smartphone (user end) and a remote server (service end). The smartphone finishes the data collection and the data preprocessing. The server finishes the classification.

will complete the classification and feedback the final result to the participant.

III. Experiments and Evaluation

A. Experimental Setup

We choose a dataset from mPower [12], which is a mobile application-based study piloting new approaches to monitoring key indicators of Parkinson Disease progression that has already surveyed more than thousands of people since 2015. We download 500 PD and 500 control cases respectively, and design three groups of comparison experiments to evaluate the performance of DeepVoice. Concretely, we evaluate the influence of data size of each sample, input data architecture and neural network architecture. In order to provide a reliable analysis, we adopt hold-out method that conducts each experiment by 10 runnings with random initialization and with a ratio of four to one for the size of the training set and testing set.

B. Comparison Among Different Data Sizes

It is necessary for a mobile health application to balance the accuracy and the convenience. Thus, we would like to explore how many seconds data can guarantee DeepVoice at a high accuracy. particularly, we collect 10 seconds audio data from each case in total and divide the audio data in a size range from 1 to 10 seconds long to compare the performance with the spectrogram as input.

The results for different data sizes are shown in Fig. 4. Results of tests are layered over a 1.96 SEM (95% confidence interval) in the black patch and a 1 SD (standard deviation) in the colorful patch. From Fig. 4, we can observe that as the data size increases, the identification accuracy is also promoted rapidly when the data size is less than 5 seconds long. If we continue to increase the length of the data size, the identification accuracy becomes saturated.

Furthermore, this experiment also indicates that 5 seconds long data contains enough information, which will be the best choice when we deploy our model in the mobile health applications. As for the strong sensing ability of our customized CNN network, we could achieve an identification accuracy of 82.95±1.92% even with 1 second long data.

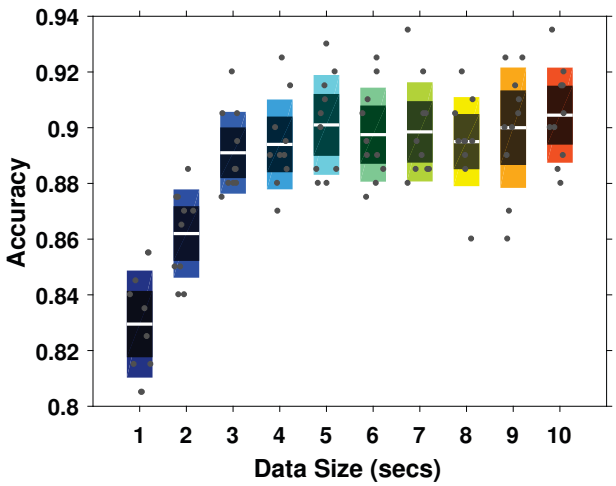
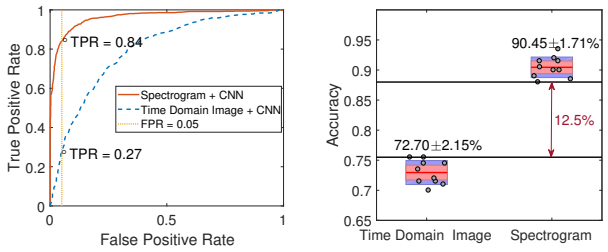


Fig. 4. The accuracy comparison for different data size with CNN. Points are layered over a 1.96 SEM in black patch and a 1 SD in colorful patch.

C. Time Domain Image V.S. Spectrogram

Data preprocessing is another significant procedure influencing the identification accuracy. In this experiment, we compare the performance between spectrogram and Time Domain image (Fig. 1). Spectrogram has been considered as the state-of-the-art input architecture in speech recognition.

The results are shown in Fig. 5. In Fig. 5(a), We can find that the curve of Spectrogram can entirely contain the one of Time Domain image, which could assert that the performance of Spectrogram is better. In Fig. 5(b), the spectrogram case demonstrates huge superiority with an accuracy of 90.45±1.71% over the Time Domain image with



(a) ROC curve for different input architecture.

(b) The accuracy comparison for different input architecture.

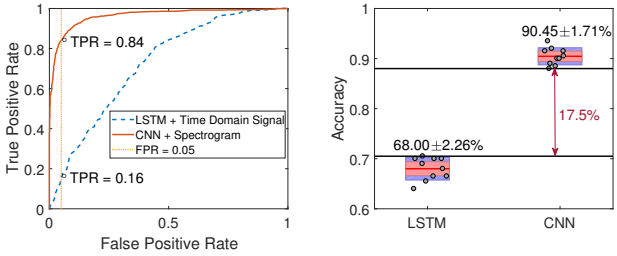
Fig. 5. The performance comparison for different input architecture with CNN. Figure (a) plots the average ROC curve for the input as Spectrogram and Time Domain image separately. Figure (b) presents the accuracy of each test. Points are layered over a 1.96 SEM in the pink patch and a 1 SD in the blue patch.

an accuracy of 72.70±2.15% in identification. Moreover, even the worst case in Spectrogram shows a 12.5% lead in accuracy over the best case in Time Domain image.

This is because PD will bring direct damage to the voice system including stiffness, quaver, and obscure sound. Instead of using the traditional Time Domain sequence, spectrogram features, we also call them as "voiceprint", could be better presented in spectrogram domain.

D. LSTM v.s. CNN

In the experiment, we expect to explore the impact of neural network structures in DeepVoice. In particular, we compare the performance between long short-term memory (LSTM) and CNN, as LSTM [13] has already been considered as the state-of-the-art neural network in speech recognition field.



(a) ROC curve for different neural network architecture.

(b) The accuracy comparison for different neural network architecture.

Fig. 6. The performance comparison for different neural network architecture with CNN. Figure (a) plots the average ROC curve for the architecture as CNN and LSTM separately. Figure (b) presents the accuracy of each test. Points are layered over a 1.96 SEM in the pink patch and a 1 SD in the blue patch.

We can observe in the Fig. 6(a) that the CNN case demonstrates huge superiority over the LSTM, as the ROC curve of the CNN entirely contains the one of the LSTM. Furthermore, even the worst case of CNN shows a 17.5% lead in accuracy over the best case in LSTM, as shown in Fig. 6(b).

This is because LSTM is difficult to model the semantic-level relationship here. One of the appeals of LSTM is its ability to connect previous information with the present, such like using the previous word to predict the next in speech recognition. However, the kind of input in DeepVoice is nothing but continually saying "Aaaaah" for ten seconds thus it glosses over the advantage of LSTM.

IV. CONCLUSIONS AND FUTURE WORK

A convenient and reliable PD identification application is still underexplored. In this paper, we presented DeepVoice, which was a voiceprint-based PD identification application using a smartphone. We implemented Deepvoice including a Joint Time-Frequency Analysis algorithm and a customized CNN architecture and evaluated DeepVoice using the dataset from mPower with the best accuracy of 90.45±1.71%.

Currently, our work focuses on studying voice information from the participants. In the future, we plan to study other types of data such as gesture and motion to make a joint prediction. In addition, the influence of demographic information including age, gender and the treatment of the Parkinson's disease on identification accuracy will be further explored.

REFERENCES

[1] A. D. Trister, E. R. Dorsey, and S. H. Friend, "Smartphones as new tools in the management and understanding of parkinson's disease," *npj Parkinson's Disease*, vol. 2, p. 16006, 2016.

[2] J. Jankovic, "Parkinsons disease: clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[3] R. L. Brey, "Muhammad ali's message: Keep moving forward," *Neurology Now*, vol. 2, no. 2, p. 8, 2006.

[4] C. R. Pereira, S. A. Weber, C. Hook, G. H. Rosa, and J. P. Papa, "Deep learning-aided parkinson's disease diagnosis from handwritten dynamics," in *Graphics, Patterns and Images (SIBGRAPI), 2016 29th SIBGRAPI Conference on*. IEEE, 2016, pp. 340–346.

[5] B. M. Eskofier, S. I. Lee, J.-F. Daneault, F. N. Golabchi, G. Ferreira-Carvalho, G. Vergara-Diaz, S. Sapienza, G. Costante, J. Klucken, T. Kautz, *et al.*, "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for parkinson's disease assessment," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 655–658.

[6] W.-Y. Cheng, A. Scotland, F. Lipsmeier, T. Kilchenmann, L. Jin, J. Schjodt-Eriksen, D. Wolf, Y.-P. Zhang-Schaerer, I. F. Garcia, J. Siebourg-Polster, *et al.*, "Human activity recognition from sensor-based large-scale continuous monitoring of parkinsons disease patients," in *Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017 IEEE/ACM International Conference on*. IEEE, 2017, pp. 249–250.

[7] N. Quinn, "Parkinsonism–recognition and differential diagnosis." *BMJ: British Medical Journal*, vol. 310, no. 6977, p. 447, 1995.

[8] ResearchKit, 2017, http://researchkit.org/.

[9] M. Garrido, "The feedforward short-time fourier transform," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 9, pp. 868–872, 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[12] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, *et al.*, "The mpower study, parkinson disease mobile data collected using researchkit," *Scientific data*, vol. 3, p. 160011, 2016.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.