Parkinsonian Chinese Speech Analysis towards Automatic Classification of Parkinson's Disease

Hao Fang FANGH18@MAILS.TSINGHUA.EDU.CN
Chen Gong GONGC16@MAILS.TSINGHUA.EDU.CN
Chen Zhang ZHANGCHEN2020@TSINGHUA.EDU.CN
Yanan Sui YSUI@TSINGHUA.EDU.CN
Luming Li LILM@TSINGHUA.EDU.CN
National Engineering Laboratory for Neuromodulation, Tsinghua University

Editors: Emily Alsentzer[⊗], Matthew B. A. McDermott[⊗], Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy[‡], Stephanie L. Hyland[‡]

Abstract

Speech disorders often occur at the early stage of Parkinson's disease (PD). The speech impairments could be indicators of the disorder for early diagnosis, while motor symptoms are not obvious. In this study, we constructed a new speech corpus of Mandarin Chinese and addressed classification of patients with PD. We implemented classical machine learning methods with ranking algorithms for feature selection, convolutional and recurrent deep networks, and an end to end system. Our classification accuracy significantly surpassed state-of-the-art studies. The result suggests that free talk has stronger classification power than standard speech tasks, which could help the design of future speech tasks for efficient early diagnosis of the disease. Based on existing classification methods and our natural speech study, the automatic detection of PD from daily conversation could be accessible to the majority of the clinical population.

Keywords: Speech Disorder, Parkinson's Disease, Classification, Chinese

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disease in the world. The affected population keeps increasing as we expect an aging society. Dorsey et al. (2007) estimated that by 2030, over eight million people will suffer from PD, while about half of them speaking Chinese. Vaiciukynas et al. (2017) showed that intervention therapy in the early stage of PD could effectively alleviate the disease progression. Early diagnosis thus is crucial to lead to early medical treatments. However, it is still difficult to make early detection of patients with PD when their motor symptoms are not obvious. Speech disorders are common symptoms of PD. About 75%-95% of PD patients show speech impairments, such as mono-tone, mono-loudness, slurred speech, and loss of volume. These symptoms frequently occur at the onset of PD, long before the appearance of significant motor signs (Pawlukowska et al., 2018). Therefore, speech deficits could be treated as potential indicators for early diagnosis of the majority of the clinical population (Rusz et al., 2011).

In this study, We aimed to assess the speech disorder and separate the clinical group from the healthy deploying various

machine learning methods. We built a corpus of Chinese speech tasks, including one specific task of Chinese poem (structured sentences) reading. We implemented classical machine learning methods with ranking algorithms for feature selection, convolutional and recurrent deep networks, and an ene to end system. We investigated acoustic characteristics that could distinguish the speech disorders of patients with PD from healthy participants. The performances of our methods surpassed the state-of-the-art results. Our results suggested that both classical machine learning methods with feature selection and advanced deep learning tools could effectively capture Parkinsonian speech characteristics. This work presented the first study on classifying PD patients from healthy subjects using Chinese speech signals to the best of our knowledge.

We further explored possible ways to optimize speech tasks towards better classification/diagnosis. Our free talk based image description task yielded better classification accuracy comparing to standard tasks, which is consistent with the physician's experience and has been proved in Goberman et al. (2010). Our methods could be directly applied to clinical evaluations and potentially utilized for detecting patients with PD from daily speech. This assistive diagnostic system would be accessible to everyone when integrated into mobile applications.

2. Related Work

Previous studies showed the possibility of classification of patients with PD from their speech signals. The speech samples collected in these studies included sustained vowels, diadochokinetic (DDK), words, and sentences. Some studies achieved relatively good classification results from sustained vowels (Xu et al., 2018; Sakar et al., 2019; Gunduz, 2019), mainly the vowel /a/. DDK, words, and sentences contained more varieties in pitch and rhythm compared to sustained vowels, comprehensively presenting more information about the speech disorders. Classifications of the PD group were reported from words and/or sentence tasks in different languages, including Hebrew (Hauptman et al., 2019), French (Jeancolas et al., 2019), Spanish (López et al., 2019), German (Orozco-Arroyave et al., 2016a), Czech (Orozco-Arroyave et al., 2016a), and Lithuanian (Vaiciukynas et al., 2017).

As features, time-frequency representations of acoustic signals are frequently used in machine learning studies for human speech. With the advantage of approximating the human auditory system's response (Logan et al., 2000), Mel-Frequency Cepstral Coefficients (MFCCs) are one of the most commonly used features in these studies (Benba et al., 2015; Xu et al., 2018; Jeancolas et al., 2019; Gaballah et al., 2019).

Both classical machine learning and deep learning methods had been introduced in the classification of PD patients' speech. The Support Vector Machine (SVM) was widely selected and frequently performed well (Garcia et al., 2018; Arias-Vergara et al., 2018). Some studies also chose k nearest neighbor (KNN) and random forest (RF) to classify PD speech (Sakar et al., 2013; Zhang, 2017; Polat, 2019). In recent years, many studies chose deep learning methods, especially Convolutional Neural Network (CNN), as classifier (Vásquez-Correa et al., 2017; Correa et al., 2018; Gunduz, 2019).

However, publicly available speech datasets and classification studies of patients with PD in Chinese are not found. Some people discussed the intonation contrast of PD patients speaking Mandarin Chinese (Liu et al., 2019) and Cantonese (Ma et al., 2010), without further attempts on classification.

3. Data Collection

We collected a new Chinese speech corpus containing speech samples recorded from 34 patients with PD and 34 healthy controls (HCs) via different phones and DVs. This uncontrolled device and environment recording condition rather than well-controlled experimental settings was more applicable in daily life, making it accessible to the majority of the disorder group. It also helped to train classifiers which would be easier to generalize and more robust in real application. The set of PD patients included 20 males aged 45 to 73 years (mean 56.70 ± 8.35 years) and 14 females aged 41 to 68 years (mean 58.29 ± 6.94 years). All PD patients were clinically diagnosed by experienced neurologists. The HC group included 16 males aged 20-55 years (mean 41.22 ± 14.89 years) and 18 females aged 21-74 years (mean 48.55 ± 11.65 years). We noticed that our participant groups were not precisely matched in age and gender. However, Sapir et al. (1999) proved that age and gender were not relevant to speech abnormalities. All subjects signed an informed consent form before their participation.

Speech signals of each subject were sampled at 48kHz for three tasks:

1) Image description (describe the image content after watching it for 30sec);

2) DDK test (quickly repeat /lalala-tatata-dadada/ for three times);

3) Text reading (read two ancient Chinese poems, composed of eight seven-word sentences).

4. Methods

4.1. Preprocessing

We first removed non-speech episodes and non-subject speech episodes from all recordings. Then, speech segments were extracted for each task respectively, according to the following criteria:

1) For image description, segments were extracted between speech pauses;

2) For DDK test, each segment included one /lalala-tatata-dadada/ sample;

3) For text reading, each segment included one seven-word sentence.

We excluded poor quality segments (e.g., containing significant noise, deviating from task requirements, etc.), and acquired 4820 speech segments in total for all 68 subjects (task 1: 2098, task 2: 773, task 3: 1949). Each segment was given a label indicating whether it belonged to PD patients or healthy subjects.

We calculated 128 MFCCs within 2000Hz frequency for each segment, using a sliding window of 2048 points (about 42.67msec) with an overlap of 75%. Among all 128 MFCCs, the 5^{th} to 44^{th} MFCCs were chosen, which covered the major frequency range of human speech. Thus, each segment was represented by a $40 \times n$ matrix, where 40 indicated the number of selected MFCCs (the 5^{th} to 44^{th} of 128 MFCCs), and n denoted the number of time bins sized around 10.67msec. The calculation of MFCCs was implemented with LibROSA library (McFee et al., 2015) in Python.

4.2. Classical Machine Learning Methods

Classical machine learning methods classify each segment via its features. The feature extraction and selection are essential, determining the performance of the classifier. We examined several commonly used classical machine learning methods in this study, including kNN, RF, and SVM with three different kernels: polynomial kernel, linear kernel, and Gaussian radial basis function (RBF) kernel. These methods were im-

plemented with the Scikit-learn library (Pedregosa et al., 2011) in Python.

4.2.1. Feature Extraction

The size of MFCCs matrices varied due to the differences in duration of speech segments. To build fixed-length feature vectors for classical machine learning methods, we compressed the $40 \times n$ matrices along the second dimension (temporal dimension). We calculated four time-domain statistics (average value, standard deviation, skewness, and kurtosis) for each MFCC to encompass the segment-level information, as what Orozco-Arroyave et al. (2016b) had done. To reduce the loss of time-variant information in matrix compression, we also added the first and the second derivatives of MFCCs (MFCCs$^{(1)}$ and MFCCs$^{(2)}$) and the same time-domain statistics of them. Concatenating all the MFCCs and their derivatives and statistics, we finally aquired a 1-D feature vector of length 480 ($40\ MFCCs \times 3\ derivatives \times 4\ statistics$) for each segment.

4.2.2. Feature Selection and Classification

We performed feature selection in the entire feature space to remove irrelevant and redundant features. Firstly, we applied nine filtering methods according to Li et al. (2017) to all features. Among those filtering methods, fisher score, reliefF, and trace ratio were based on similarity. These methods assessed features' importance by their ability to preserve data similarity, especially referring to the data manifold structure encoded by an affinity matrix. RFS, ls_l21 (Liu et al., 2012), and ll_l21 were based on sparse learning. These methods considered minimizing both biases of fitting models and sparse regularization terms. Gini index, f-score, and t-score were based on statistics, assessing feature importance with statistical measures. Each method provided a feature-rank, which described the importance of features in classifying patients with PD from HCs. In this sense, features with higher ranks played more significant roles in classification.

Secondly, we performed feature selection and PD classification with RBF-kernel based SVM classifiers simultaneously. Each time we selected the top m ($m = 1, 2, ..., 480$) features according to their ranking orders generated from one of the nine filtering methods as inputs for the classifier. The leave-one-subject-out (LOSO) strategy was deployed during classification: segments from one subject were excluded as test samples, and the rest were used for training. This strategy guaranteed that segments from the same subject only appeared either in the training or test set, eliminating the risk of identity confounding (Neto et al., 2019). As a result, LOSO provided a more objective and fair evaluation of classifiers.

We traversed the nine feature-ranks and all possible values of m. The classification performance for each of these combinations guided the selection of the best feature subset. We also examined the performance of KNN, RF, and linear- and polynomial-kernel based SVM classifiers, using grid search strategy for parameter tuning.

4.3. Deep Learning Methods

Deep learning methods allowed us to use the original MFCCs matrices (see Section 4.1) as inputs to differentiate PD patients without manually extracting and selecting features, thus avoiding the loss of time-variant information when calculating the four statistics in previous classical machine learning methods. We extracted samples from a sliding window of 40 time bins (which corresponded to the speech signal of about 426.7msec) with an overlap of 75% for each of the original $40 \times n$ MFCCs matrices. In total, 66438 samples

sized 40×40 were obtained as the inputs for the two types of deep neural networks we designed to perform the classification.

4.3.1. 6-layer Convolutional Neural Network

CNN has been widely applied in tasks regarding images in various domains. CNN introduces convolutional and pooling layers as hidden layers. With its shared-weights architecture, CNN has a great response to the sliding and deforming of images. Accordingly, CNN is suitable for processing the time-frequency representation of speech signals. The CNN architecture proposed here included six layers, described in Table 1.

Table 1: Description of the CNN structure

Layer	Kernel Size	Output Size
Input	-	$40 \times 40 \times 1$
Conv	$3 \times 3 \times 16$	$40 \times 40 \times 16$
Conv	$3 \times 3 \times 16$	$40 \times 40 \times 16$
MaxPool	2×2	$20 \times 20 \times 16$
Conv	$3 \times 3 \times 32$	$20 \times 20 \times 32$
MaxPool	2×2	$10 \times 10 \times 32$
Conv	$3 \times 3 \times 64$	$10 \times 10 \times 64$
MaxPool	2×2	$5 \times 5 \times 64$
MaxPool	5×5	$1 \times 1 \times 64$
Flatten	-	64
FC	64×8	8
Dropout	$p = 0.5$	-
FC	8×2	2

4.3.2. Self-attention based Long Short-Term Memory Network

Long Short-Term Memory (LSTM) networks are a kind of Recurrent Neural Networks (RNNs) dealing with long-term sequences (Hochreiter and Schmidhuber, 1997). Inspired by the physiological process of human decision-making after listening to a speech segment where attention is involved, we constructed a self-attention based LSTM architecture. The MFCCs matrix was fed into the LSTM layer chronologically frame by frame, producing a 2-D dimensional output matrix V of size $40 \times Hidden\_size$. We designed the attention layer according to the structure of the Transformer proposed by Google (Vaswani et al., 2017). We chose the last output vector as the feature vector extracted by the encoder, fed it into the classifier to make a decision. The LSTM structure is described in Table 2.

Table 2: Description of the LSTM structure

Layer	Kernel Size	Output Size
Input	-	40×40
LSTM	Hidden Size = 512, Layer Number = 3	40×512
Attention	Self-Attention	40×512
Last output	-	512
MaxPool	2	256
FC	256×64	64
FC	64×8	8
Dropout	$p = 0.4$	-
FC	8×2	2

4.3.3. End-to-End System

The end-to-end (E2E) system usually means the network that can utilize the original signal without additional processing. In this work, we also constructed a deep neural network that directly adopted the original waveform signal as the input. We hoped this design could capture preciser time information than using MFCCs as input. We used a time-convolutional layer structure as the front-end network instead of the calculation of MFCCs in Section 4.1, inspired by the CLDNN structure proposed by Sainath et al. (2015). We used the LSTM network as the back-end network to make a final decision. The E2E architecture proposed here is described in Table 3.

4.3.4. Training

We utilized the same LOSO strategy to divide training and test sets, and further sepa-

Table 3: Description of the E2E structure

Layer	Kernel Size	Output Size
Input	-	1×21500
Unfold	Window = 21500, Stride = 500	40×2000
1d-Conv	200×40, Stride = 25	$40 \times 73 \times 40$
MaxPool	73×1	$40 \times 1 \times 40$
Flatten	-	40×40
Log-ReLU	-	40×40
LSTM	Hidden Size = 512, Layer Number = 3	40×512
Last output	-	512
MaxPool	2	256
FC	256×64	64
Dropout	p = 0.4	-
FC	64×8	8
FC	8×2	2

rated 20% samples from the training set for validation. We chose BCE loss as the loss function, Adam as the optimizer, and set the learning rate to 0.001. The training process was stopped when accuracy on the validation set reached 0.99 or the maximum number of epochs reached (in this case, we chose the number of epochs when the highest validation accuracy was achieved).

4.4. Evaluation

We evaluated the performances of classifiers at segment-level. The classical machine learning methods directly provided a segment-level classification. While in deep learning methods, we fed the 40×40 MFCCs matrix (named 'sample' of a speech 'segment') into the classifier, thus made a sample-level classification. We then averaged the prediction probabilities of samples belonged to the same segment. The segment's label was assigned according to the averaged probability. To improve robustness and generalizability, we excluded samples with the top and bottom 30% probability values from averaging. This strategy generated a segment-level classification result, making it available to compare the performance of deep learning and classical machine learning methods.

The classifiers were evaluated through the prediction accuracy (ACC) by calculating the ratio of correctly classified segments to all segments. Confusion matrix and Area Under the receiver operating characteristics Curve (AUC) were also used as the extension evaluation criterion.

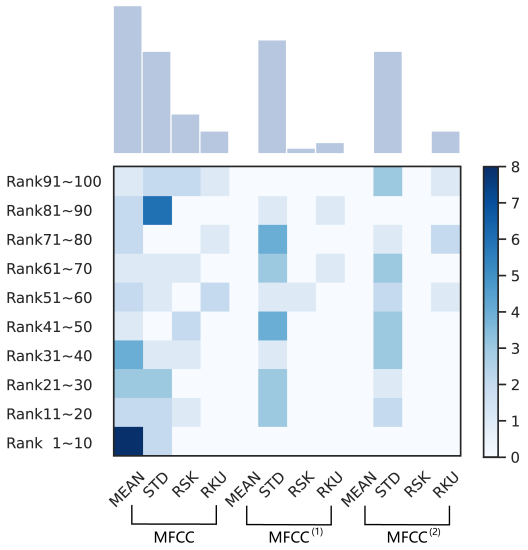

Figure 1: Distribution of selected features

5. Results and Discussion

For classical machine learning methods, using ls_121, the sparse learning-based feature selection method led to the achievement of the best performance comparing to others. We selected the top 100 features with ls_l21 as the input feature subset for classification. This group of features reached execellent performance (exceeded 99% of maximum AUC) with a small number of features. Figure 1 presents the distribution of this feature subset. MFCCs accounted for a large proportion

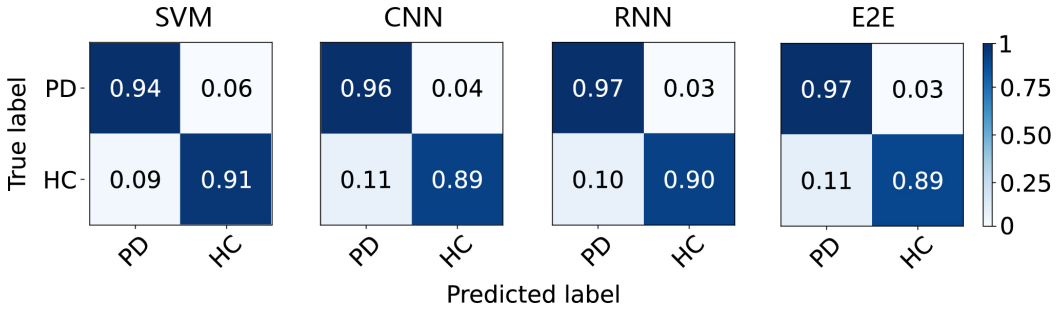

Figure 2: Confusion matrices of four methods

of selected features at higher ranks, whereas the proportion of MFCCs$^{(1)}$ and MFCCs$^{(2)}$ raised in lower rank intervals. The total number of each statistic selected for classification is shown in the upper part of Figure 1. The average value of MFCCs played a major role in classification, indicating the lower speech volume of PD patients. The widely distributed standard deviation of MFCCs and its derivatives indicated the importance of time-variant information in speech signals. This result revealed the contribution of each feature in differentiating the PD patients.

Among the three classical machine learning classifiers mentioned in Section 4.2.2, RBF-kernel based SVM with the selected feature subset achieved the highest AUC of 0.978 and ACC of 0.930, which was significantly better than those acquired with all 480 features (AUC = 0.966 and ACC = 0.906), showing the importance of feature selection. Figure 2 (a) shows the confusion matrix for RBF-kernel based SVM, with both the sensitivity and specificity higher than 0.89. After grid search, the best parameters we found in RBF-kernel were C = 4 and gamma = $1/(\#features \times variance\ of\ training\ samples)$.

Other classifiers with the top 100 feature subset provided by ls_l21 also performed well, though worse than RBF-kernel SVM: AUC = 0.955 and ACC = 0.880 for RF, AUC = 0.923 and ACC = 0.868 for kNN, AUC = 0.931 and ACC = 0.854 for linear kernel based SVM, and AUC = 0.954 and ACC = 0.875 for polynomial kernel based SVM.

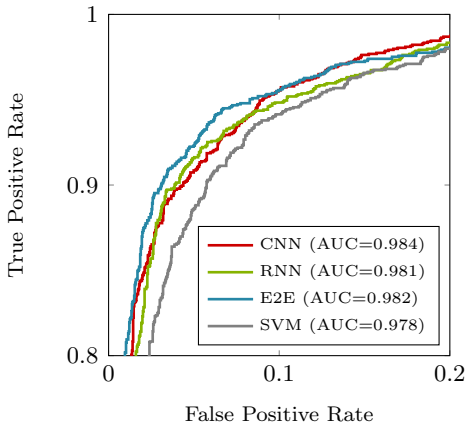

Figure 3: ROC curve of four methods

For deep learning methods, the 6-layer CNN, self-attention based LSTM and E2E system all exhibited better performance. CNN performed the best, with an AUC of 0.984 and ACC of 0.938. Self-attention based LSTM reached an AUC of 0.981 and ACC of 0.942, while E2E system reached an AUC of 0.982 and ACC of 0.945. Figure 2 (b), (c) and (d) presents the confusion matrices for these three methods respectively. The better result may due to the detailed usage of signal information. In classical machine learning methods, we roughly represented the signal by the statistics of MFCCs, losing some time-variant informa-

tion of the speech signal. Orozco-Arroyave et al. (2015) successfully classify PD speech from voiced/unvoiced transitions, showing that parkinsonian speech disorder can be detected from the rapid changes of speech. The original MFCCs matrix can provide a higher time-domain resolution than its statistics, which may benefit the classification. The local view of each method's ROC curve is presented in Figure 3, providing a visual comparison of these four methods.

This study included three speech tasks, as described in Section 3. Accordingly, we tested the classification performance on each task individually. The same feature selection procedure in Section 4.2.2 was applied. Figure 4 shows the AUC calculated for classification performance with different numbers of selected features using RBF-kernel based SVM.

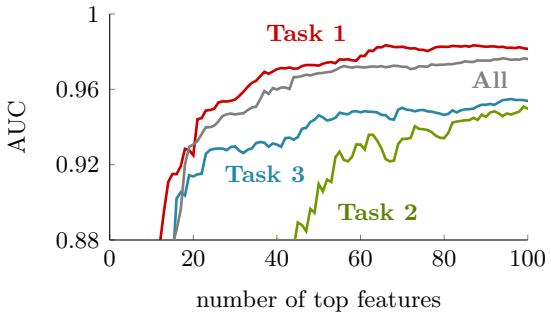

Figure 4: AUC of each tasks

The AUC of task 2 (DDK) is lower than task 1 (image description) and task 3 (text reading), which is consistent with the result shown in Orozco-Arroyave et al. (2016a). Sentences involved different words, pauses, and rhythms, while DDK in this study mainly contained the single vowel /a/. Thus DDK covered less variation in voice and only reflected a narrower range of vocal space. This performance difference might support the assumption that MFCCs are more suitable for complicated tasks or speech, as discussed in Moro-Velazquez et al. (2020).

Besides, task 1 achieved a significantly better classification result compared to task 3. The reason might be that text reading only involved fixed contents, while image description allowed more freedom with fewer limitations. Another possible explanation of this difference could be that subjects were more focused on recalling the image's content when they participated in task 1, thus relaxed the control of some muscles and exhibited more severe speech disorder. As a result, the difference between PD patients and healthy subjects' speech was shown more obviously. This trick of attracting patients' attention is also used in other tests to illuminate their movement disorder. PD patients could be classified from natural speech without specifically designed tasks, indicating the potential to detect PD patients from speech obtained in daily life and assist early clinical diagnosis of PD.

Though worse than task 1, task 3 also achieved a strong classifying ability and was better than task 2. The reading texts we designed in task 3 were two ancient Chinese poems, which are very familiar to Chinese people. Subjects could read the poems in their most comfortable way as a conditional reflex, significantly reduced the time cost. Besides, it is uncertain whether image description may be influenced by cognitive function. Reading tasks were more focused on speech function and can provide an equal number of training data, which is suitable for machine learning.

We visualized the selected features using t-distributed stochastic neighbor embedding (t-SNE). In detail, we performed principal component analysis (PCA) to reduce the features from 480-d to 50-d and retain more than 90% of the effective information. Then we performed t-SNE on PCA result. Figure 5

showed the separability of features from the two groups.

We further compared our RBF kernel-based SVM results in single speech task with previous studies, as shown in Table 4. Our method (in bold) surpassed previous state-of-the-art results.

6. Conclusions and Future Work

In this study, we built a speech dataset with different tasks in Mandarin Chinese; investigated features for speech disorders related to Parkinson's disease; developed high-accuracy methods for the classification of patients with PD and healthy subjects.

Both classical machine learning methods with feature selection and deep learning methods we developed have state-of-the-art performance. In addition, we tested the classification performance for different speech tasks. The result suggests that free talk has stronger classification power than standard tasks, which could aid the design of future speech tests for efficient early diagnosis of the disease.

For future work, more specifically designed neural networks may further improve accuracy and generalizability. The pipeline of feature selection and the neural network structure proposed in this study could also be applied to speech tasks in other languages. High quality and large volume data on speech tasks from patients and healthy subjects are crucial for practical applications. A better understanding of the speech disorders' mechanism would also help us in developing more effective diagnostic tools.

Acknowledgments

This work is sponsored by The National Key Research and Development Program of China (2016YFC0105502), NSFC (81527901), and Shenzhen International Cooperative Research Project (GJHZ20180930110402104). All the authors are affiliated with the National Engineering Laboratory for Neuromodulation, School of Aerospace Engineering, Tsinghua University. Luming Li is also affiliated with: Precision Medicine Healthcare Research Center, Tsinghua-Berkeley Shenzhen Institute, Tsinghua University; IDG / McGovern Institute for Brain Research, Tsinghua University; Institute of Epilepsy, Beijing Institute for Brain Disorders. Correspondence to Yanan Sui and Luming Li.

References

T. Arias-Vergara, J. C. Vasquez-Correa, J. R. Orozco-Arroyave, P. Klumpp, and E. Noth. Unobtrusive monitoring of speech impairments of parkinson's disease patients through mobile devices. In *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Table 4: ACC comparison with previous studies

Article	Language	Vocal task		
		Talking[1]	DDK	Reading
Hauptman et al. (2019)	Hebrew	0.741	0.741	0.721
Jeancolas et al. (2019)	French	0.74	0.78	-
López et al. (2019)	Spanish	0.84	0.78	0.80
Vaiciukynas et al. (2017)	Lithuanian	-	-	0.859
Ours	Chinese	**0.940**	**0.835**	**0.911**

[1] Talking task includes free talk, monologue, and image description

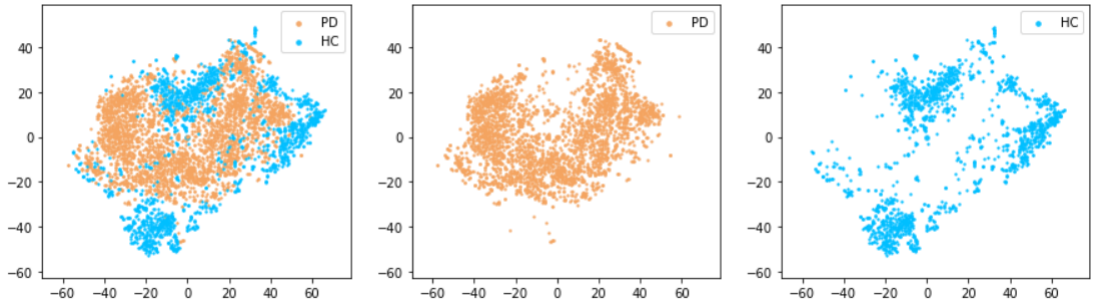

Figure 5: t-SNE result of selected features

Achraf Benba, Abdelilah Jilbab, Ahmed Hammouch, and Sara Sandabad. Voiceprints analysis using mfcc and svm for detecting patients with parkinson's disease. In *2015 International conference on electrical and information technologies (ICEIT)*, pages 300–304. IEEE, 2015.

Juan Camilo Vásquez Correa, Tomas Arias, Juan Rafael Orozco-Arroyave, and Elmar Nth. A multitask learning approach to assess the dysarthria severity in patients with parkinson's disease. In *Interspeech 2018*, 2018.

ERl Dorsey, R Constantinescu, JP Thompson, KM Biglan, RG Holloway, K Kieburtz, FJ Marshall, BM Ravina, G Schifitto, A Siderowf, et al. Projected number of people with parkinson disease in the most populous nations, 2005 through 2030. *Neurology*, 68(5): 384–386, 2007.

Amr Gaballah, Vijay Parsa, Monika Andreetta, and Scott Adams. Objective and subjective speech quality assessment of amplification devices for patients with parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(99):1226–1235, 2019.

Nicanor Garcia, Juan Camilo Vásquez Correa, Juan Rafael Orozco-Arroyave, and Elmar Nth. Multimodal i-vectors to detect and evaluate parkinson's disease. In *Interspeech 2018*, 2018.

Alexander M Goberman, Michael Blomgren, and Erika Metzger. Characteristics of speech disfluency in parkinson disease. *Journal of Neurolinguistics*, 23(5): 470–478, 2010.

Hakan Gunduz. Deep learning-based parkinson's disease classification using vocal feature sets. *IEEE Access*, 7:115540–115551, 2019.

Yermiyahu Hauptman, Ruth Aloni-Lavi, Itshak Lapidot, Tanya Gurevich, Yael Manor, Stav Naor, Noa Diamant, and Irit Opher. Identifying distinctive acoustic and spectral features in parkinson's disease. *Proc. Interspeech 2019*, pages 2498–2502, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Laetitia Jeancolas, Graziella Mangone, Jean-Christophe Corvol, Marie Vidailhet, Stéphane Lehericy, Badr-Eddine Benkelfat, Habib Benali, and Dijana Petrovska-Delacretaz. Comparison of telephone recordings and professional microphone recordings for early detection of parkinson's disease, using mel-frequency

cepstral coefficients with gaussian mixture models. *Proc. Interspeech 2019*, pages 3033–3037, 2019.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):1–45, 2017.

Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l2, 1-norm minimization. *arXiv preprint arXiv:1205.2631*, 2012.

Lei Liu, Meng Jian, and Wentao Gu. Prosodic characteristics of mandarin declarative and interrogative utterances in parkinson's disease. *Proc. Interspeech 2019*, pages 3870–3874, 2019.

Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *Ismir*, volume 270, pages 1–11, 2000.

José Vicente Egas López, Juan Rafael Orozco-Arroyave, and Gábor Gosztolya. Assessing parkinson's disease from speech using fisher vectors. *Proc. Interspeech 2019*, pages 3063–3067, 2019.

Joan K-Y Ma, Tara L Whitehill, and Susanne Y-S So. Intonation contrast in cantonese speakers with hypokinetic dysarthria associated with parkinson's disease. *Journal of Speech, Language, and Hearing Research*, 53:836–849, 2010.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

Laureano Moro-Velazquez, Jesus Villalba, and Najim Dehak. Using x-vectors to automatically detect parkinson's disease from speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1155–1159. IEEE, 2020.

Elias Chaibub Neto, Abhishek Pratap, Thanneer M Perumal, Meghasyam Tummalacherla, Phil Snyder, Brian M Bot, Andrew D Trister, Stephen H Friend, Lara Mangravite, and Larsson Omberg. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ digital medicine*, 2(1):1–6, 2019.

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

Juan Rafael Orozco-Arroyave, Florian Hönig, Julián D Arias-Londoño, Jesús Francisco Vargas-Bonilla, Sabine Skodda, Jan Rusz, and Elmar Nöth. Voiced/unvoiced transitions in speech as a potential bio-marker to detect parkinson's disease. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Juan Rafael Orozco-Arroyave, JC Vdsquez-Correa, Florian Hönig, Julián D Arias-Londono, Jesús Francisco Vargas-Bonilla, Sabine Skodda, Jan Rusz, and E Noth. Towards an automatic monitoring of the neurological state of parkinson's patients from speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6490–6494. IEEE, 2016b.

Wioletta Pawlukowska, Aleksandra Szylińska, Dariusz Kotlega, Iwona Rotter,

and Przemysław Nowacki. Differences between subjective and objective assessment of speech deficiency in parkinson disease. *Journal of Voice*, 32(6):715–722, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Kemal Polat. A hybrid approach to parkinson disease classification using speech signal: the combination of smote and random forests. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–3. IEEE, 2019.

Jan Rusz, Roman Cmejla, Hana Ruzickova, and Evzen Ruzicka. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease. *The journal of the Acoustical Society of America*, 129 (1):350–367, 2011.

Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals. Learning the speech front-end with raw waveform cldnns. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

B. E Sakar, M. M Isenkul, C. O Sakar, and A Sertbas. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical Health Informatics*, 17(4): 828–834, 2013.

C Okan Sakar, Gorkem Serbes, Aysegul Gunduz, Hunkar C Tunc, Hatice Nizam, Betul Erdogdu Sakar, Melih Tutuncu, Tarkan Aydin, M Erdem Isenkul, and Hulya Apaydin. A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform. *Applied Soft Computing*, 74:255–263, 2019.

Shimon Sapir, A Pawlas, L Ramig, S Countryman, C O'BRIEN, M Hoehn, and LA Thompson. Speech and voice abnormalities in parkinson disease: relation to severity of motor impairment, duration of disease, medication, depression, gender and age. *NCVS Status and Progress Report*, 14:149–161, 1999.

Evaldas Vaiciukynas, Adas Gelzinis, Antanas Verikas, and Marija Bacauskiene. Parkinson's disease detection from speech using convolutional neural networks. In *International Conference on Smart Objects and Technologies for Social Good*, pages 206–215. Springer, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

J.C. Vásquez-Correa, Juan Rafael Orozco-Arroyave, and Elmar Nth. Convolutional neural network to model articulation impairments in patients with parkinson's disease. In *Interspeech 2017*, 2017.

Zhijing Xu, Juan Wang, Ying Zhang, and Xiangjian He. Voiceprint recognition of parkinson patients based on deep learning. *arXiv preprint arXiv:1812.06613*, 2018.

Y. N. Zhang. Can a smartphone diagnose parkinson disease? a deep neural network method and telediagnosis system implementation. *Parkinsons Disease*, 2017:1–11, 2017.