Received December 26, 2020, accepted January 6, 2021, date of publication January 13, 2021, date of current version January 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051432

A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech

CHANGQIN QUAN, KANG REN, AND ZHIWEI LUO
Graduate School of System Informatics, Kobe University, Kobe 657-8501, Japan
Corresponding author: Changqin Quan (quanchqin@gold.kobe-u.ac.jp)

ABSTRACT Detection of voice changes in Parkinson's Disease (PD) patients would make it possible for early detection and intervention before the onset of disabling physical symptoms. This study explores static and dynamic speech features relating to PD detection. A comparative analysis of the articulation transition characteristics shows that the number of articulation transitions and the trend of the fundamental frequency curve are significantly different between HC speakers and PD patients. Motivated by this observation, we propose to apply Bidirectional long-short term memory (LSTM) model to capture time-series dynamic features of a speech signal for detecting PD. The dynamic speech features are measured based on computing the energy content in the transition from unvoiced to voiced segments (onset), and in the transition from voiced to unvoiced segments (offset). Under the two evaluation methods of 10-fold cross validation (CV) and splitting the dataset without samples overlap of one individual, the experimental results show that the proposed method remarkably improves the accuracy of PD detection over traditional machine learning models using static features.

INDEX TERMS Parkinson's disease, speech signal processing, deep learning, dynamic features, bidirectional long short term memory.

I. INTRODUCTION

Parkinson's disease (PD) is currently the second most frequent neurodegenerative disease, after Alzheimer disease [1]. Generally, there are two kinds of symptoms of PD, motor symptoms and non-motor symptoms. The main motor symptoms of PD are tremor, slowness of movement (bradykinesia), stiffness (rigidity), and poor balance (postural instability). Non-motor symptoms mainly include mood disorders, cognitive dysfunction, pain, sensory dysfunction, and dysautonomia [2]. Motor speech disorders are common among PD patients. Speech disturbances such as very quiet and hurried speech occur in more than half of the patients [3]. Analysis of speech signals is considered as an important non-invasive method for PD identification. Noninvasive identification and prediction technology of PD is attractive to clinicians and neuroscientist. In addition, detection of voice changes in PD patients would make it possible for early detection and intervention before the onset of disabling physical symptoms, giving a relevant effect on both the quality of life of patients and the healthcare system.

The development of modern speech processing technology mostly relies on interdisciplinary research in the areas of multimodal signal processing and artificial intelligence. A number of methods have been developed with the aim of solving various Human-computer interaction (HCI) problems [4]. Some results based on PD speech analysis have shown that people with Parkinson have shorter maximum phonation time, higher jitter and shimmer, decreased pitch range and increased phonation threshold pressure [5]. Based on these analyses, many machine learning (ML) models have been applied for PD detection [6]–[11].

ML based approaches cast the problem of PD detection from speech into a classification problem. The dominant ML methods generally fall under two broad categories: traditional ML based methods and Deep Learning (DL) based methods. Traditional ML based methods include Support Vector Machine (SVM) [12]–[14], K-Nearest Neighbor (KNN) [12], Naïve Bayes (NB) [12], [15], [16], Decision tree [12], Genetic Algorithm [14], and their combinations [17], [18]. Traditional ML based methods usually extract global static

The associate editor coordinating the review of this manuscript and approving it for publication was Valentina E. Balas.

features based on different measurements such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, jitter, shimmer values from voice signals. Then, feature selection techniques such as Least Absolute Shrinkage Selection Operator (LASSO), Minimum Redundancy Maximum Relevance, Relief and Local Learning-Base Feature Selection (LLBFS) are used to select the best features. After that, dimensionality reduction technology like Principal Component Analysis (PCA) is often applied to compress a dataset onto a lower-dimensional feature subspace for reducing the model complexity and avoiding overfitting.

Different from the feature selection techniques used in traditional ML based methods, one of the strong points of DL is precisely the hierarchical feature selection along the successive level of increasing abstraction in detecting patterns. Several studies have explored PD detection from speech based on DL, such as Convolutional Neural Network (CNN) [19]–[22].

Most of previous studies made their efforts on finding effective static features for PD speech classification; some studies used continuous speech features while ignoring the interdependencies in sequences of features.

Our contributions can be summarized as follows:

1) Explored static and dynamic speech features relating to PD detection. A comparative analysis of the articulation transition characteristics shows that the number of articulation transitions and the trend of the fundamental frequency curve are significantly different between HC speakers and PD patients. We applied a paired t-test to evaluate the difference on the number of articulation transitions between HC speaker and PD patient groups and got a p-value of 0.042 (<0.05), which indicates the difference did not occur by chance.

2) Proposed to apply Bidirectional long short term memory (LSTM) model to capture time-series dynamic features of speech signals for detecting PD. The dynamic speech features are measured based on computing the energy content in the transition from unvoiced to voiced segments (onset), and in the transition from voiced to unvoiced segments (offset). To the best of our knowledge, combining Bidirectional LSTM model and dynamic articulation transition features of speech has not yet been used for PD detection. Under the two evaluation methods of 10-fold cross validation (CV) and splitting the dataset without samples overlap of one individual, the experimental results showed that the proposed method remarkably improves the accuracy of PD detection over traditional ML models using static features.

The outline of the paper is as follows: Section 2 discusses related work. The features of speech for PD detection are introduced in Section 3. The framework of the model is shown in Section 4. The experimental study is introduced in Section 5. Section 6 is discussion and Section 7 is the conclusions.

II. RELATED WORK

Recent research is now increasingly focused on the use of DL architectures and algorithms to solve difficult speech signal processing (SSP) tasks. This section reviews significant speech features and DL related models and methods that have been employed for PD detection from speech.

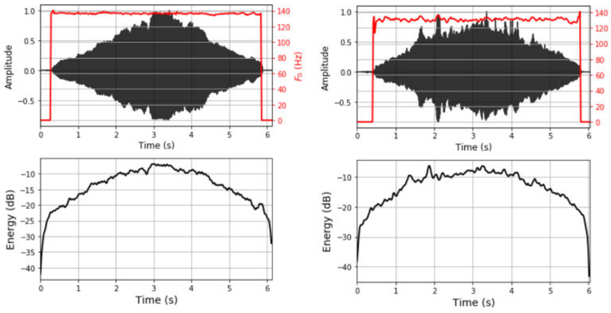
PD detection from speech can be regarded as a two-step task. The first step is to transfer the input speech signal to speech feature vectors or tensors that can be analyzed by DL models. Regarding the speech features of PD patients, several dimensions of speech are included, such as articulation, phonation, prosody, etc. [23]–[25]. Previous studies have explored articulation analysis with different acoustic measures including the triangular Vowel Space Area (tVSA), Vowel Articulation Index (VAI), Formant Centralization Ratio (FCR), etc. Skodda and Visser found that VAI is reduced in PD speakers compared with respect to the healthy controls (HC) group based on vowel articulation analyses [26]. After comparing sustained phonations of the Czech vowel /i/, repetition of short sentences, reading of a text with 80 words, and a monologue of approximately 90-second duration, Rusz *et al.* found that the monologue was the most suitable task to differentiate speech of early PD patients and HC speakers, giving classification accuracies of up to 80% [27]. Phonation is evaluated through a set of measures that include jitter, shimmer, the correlation dimension (D2), etc. In [28], phonation and articulation analyses are performed using recordings of sustained vowels; and an accuracy of 81% was reported. Arias-Vergara *et al.* found that the inclusion of features extracted from continuous speech, e.g., prosody, intelligibility, and articulation, could obtain satisfactory results to discriminate between PD patients and HC speakers [29]. Additionally, several features typically used in speech processing such as MFCC, energy content, pitch and others were also employed in PD detection.

With the extracted speech features, the second step of DL based methods is designing a classification framework based on the properties of the Neural Networks. In [30], the Multiple Artificial Neural Networks (ANNs) are used with 26 speech features for PD detection. Principal Component Analysis (PCA) and Self-Organizing Map (SOM) are applied for feature selection. The ANN architecture is configured with few neurons (5 and 10) and hidden layers (from 1 to 3). In [31], the deep neural networks (DNNs) is used for PD detection based on the Audio-Visual Emotion recognition Challenge (AVEC) feature set [32] and the Geneva Minimalistic Acoustic Parameter Set (GeMaps) [33]. Using 16 biomedical voice measures, DNN is also applied for PD severity prediction [34].

CNNs were originally developed for image recognition tasks [35] and they are been successfully applied to SSP domain. CNNs are essentially formed of multiple hidden layers where the convolution and pooling operations are performed on. In [19], two 9-layered CNNs on feature-level combination and on model-level combination are separately employed for PD detection, and feature correlations are computed for extracting the relationships between the features.

Recent studies have explored end-to-end DL approaches using CNN for audio classification. First audio signals are converted into time-frequency representations, and then recognized by a CNN model like the task of image recognition. In [20] it was introduced a method to model the transitions between voiced and unvoiced segments for PD detection in three different languages (Spanish, German, and Czech). The CNNs model is applied to extract speech features from two time–frequency representations: the short time Fourier transform and the continuous wavelet transform. In [21] the authors studied the spectrogram-based CNN model for PD detection in Lithuanian language.

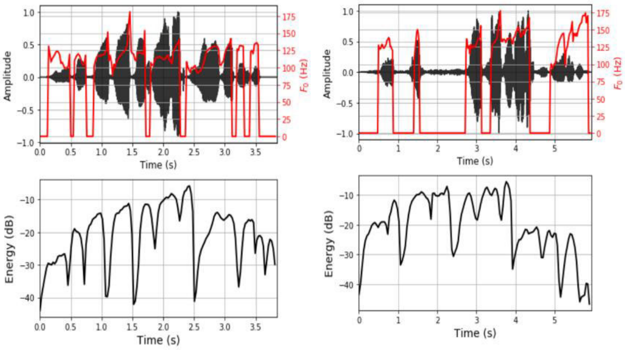
CNNs have exhibited some degree of invariance to small shifts of speech features along the frequency axis and the efficiency for Automatic Speech Recognition (ASR) [36]. However, a main issue with CNNs is the limitation for modeling long-distance contextual information even when using dilated convolutions [37]. Comparatively, RNNs (Recurrent Neural Networks) [38] are able to model long-distance contextual information by memorizing over previous computations and utilize this information in current processing. However, simple RNN based methods suffer from the vanishing gradient problem, which makes it hard to learn and tune the parameters of the earlier layers in the network. This limitation was overcome by various networks such as LSTM [39]. Particularly, Bidirectional LSTMs train two LSTMs on the input sequence. The first on the input sequence as-is and the second on a reversed copy of the input sequence, which provide additional context to the network and result in faster and even fuller learning on the problem. The effectiveness of Bidirectional LSTMs for PD detection from speech has not been referred in previous studies. In this study, we focus on exploring the Bidirectional LSTMs model to capture time-series dynamic features of speech for detecting PD.



(a) HC speaker

(b) PD patient

FIGURE 1. Speech signal, fundamental frequency, and energy of a sustained phonation of monophonic /a/ uttered by a HC speaker (a), and a PD patient (b).



(a) HC speaker

(b) PD patient

FIGURE 2. Speech signal, fundamental frequency, and energy of a sustained phonation of a short sentence uttered by a HC speaker (a) and a PD patient (b). Both subjects pronounced exactly the same short sentence.

III. SPEECH FEATURES FOR PD DETECTION

There are several dimensions of speech features relating to PD detection, including phonation, articulation, prosody, Intelligibility, etc. [40].

Phonation features are characterized by bowing and inadequate closure of vocal folds [23]. Phonation features are mainly related to perturbation measures such as jitter (temporal perturbation of the fundamental frequency), shimmer (temporal perturbation of the amplitude of the signal), Amplitude Perturbation Quotient (APQ), and Pitch Perturbation Quotient (PPQ) [41].

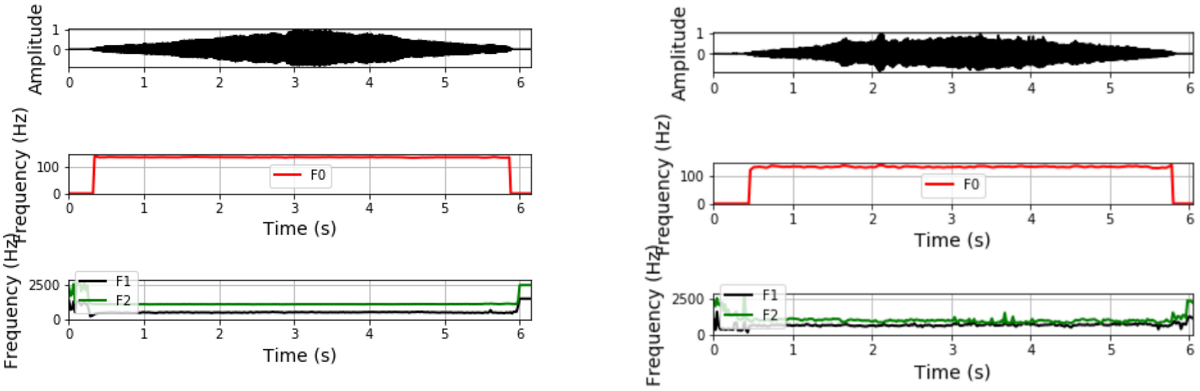
Fig. 1 and Fig. 2 show the contour of the fundamental frequency computed over monophonic /a/ and a short sentence (Both subjects pronounced exactly the same short sentence.) separately uttered by a HC speaker (a) and a PD patient (b). Both of them are mandarin native speakers. NeuroSpeech software [42] is utilized for phonation analysis.

From Fig. 1 and Fig. 2, phonation analysis shows that the contour of the HC sample is more stable than the contour obtained from the PD patient for both inputs of monophonic /a/ and a short sentence.

Articulation features are mainly related to reduced amplitude and velocity of lip, tongue, and jaw movements [43]. Articulation analysis can be performed with sustained vowels or with continuous speech signals [41]. With NeuroSpeech software [42], articulation feature is mainly based on the computation of the first two vocal formants F1 and F2, including the measures of the Vowel Space Area (VSA), Vocal Pentagon Area (VPA), and Formant Centralization Ratio (FCR). Fig. 3 shows the speech signals and fundamental frequency of a sustained articulation of monophonic /a/ uttered by a HC speaker (a) and a PD patient (b).

From Fig. 3, the sustained articulation analysis of monophonic /a/ shows that the contour of the HC speaker is more stable than the contour obtained from the PD patient.

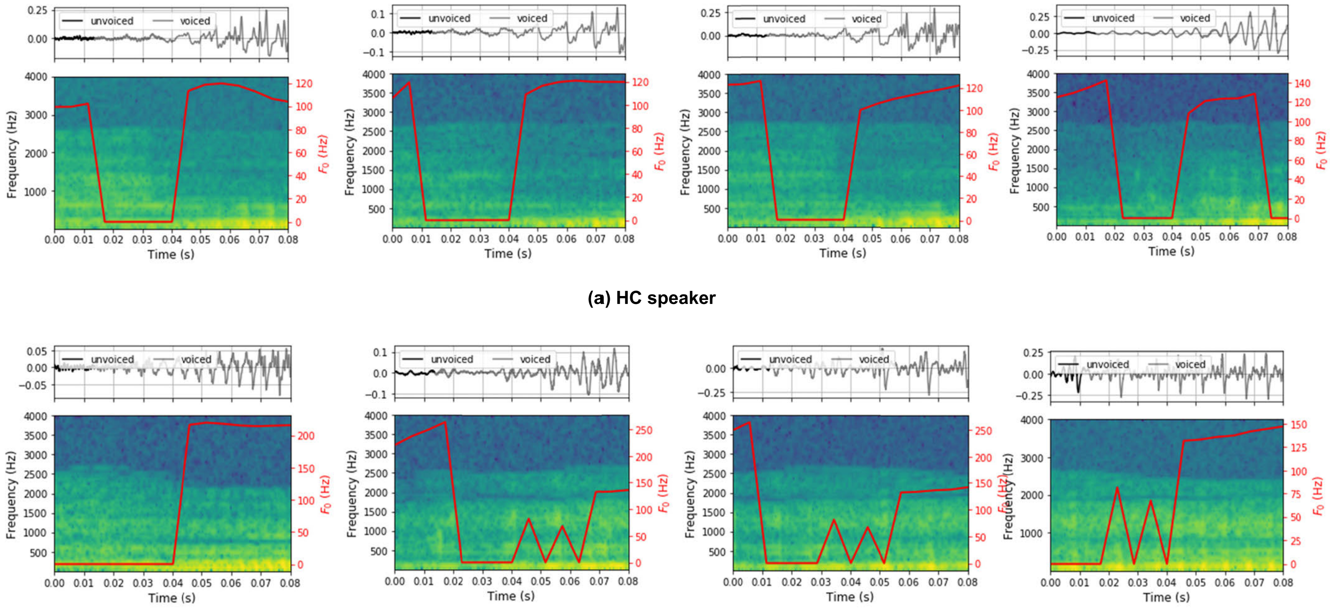
With regard to continuous speech signals, the articulation feature is measured based on computing the energy content in the transition from unvoiced to voiced segments (onset), and in the transition from voiced to unvoiced segments (offset) [44], [45]. The main hypothesis is: PD patients produce abnormal unvoiced sounds and have difficulty to begin and/or to stop the vocal fold vibration [46]. It can be observed on speech signals by modeling the frequency content of the unvoiced frames and the transitions between voiced and unvoiced sounds.



(a) HC speaker

(b) PD patient

FIGURE 3. Speech signal, fundamental frequency, and energy of a sustained articulation of monophonic /a/ uttered by a HC speaker (a) and a PD patient (b).



(a) HC speaker

(b) PD patient

FIGURE 4. Comparison of articulation transitions on the speech signal and fundamental frequency of a short sentence uttered by a HC speaker (a) and a PD patient (b). Both subjects pronounced exactly the same short sentence.

Fig. 4 shows the comparison of articulation transitions on the continuous speech signal and fundamental frequency of a short sentence uttered by a HC speaker (a) and a PD patient (b). Both subjects pronounced exactly the same short sentence. The duration of the continuous speech is about 3.9s. There are 17 articulation transitions for the HC speaker, and 39 articulation transitions for the PD patient.

A comparative analysis of the articulation transition characteristics shows that the number of articulation transitions and the trend of the fundamental frequency curve are significantly different between HC speakers and PD patients. We applied a paired t-test to evaluate the difference on the number of articulation transitions between HC speaker and PD patient groups (44 speech samples for each group, all subjects pronounced exactly the same short sentence). We got a p-value of 0.042 (<0.05), which indicates the difference did not occur by chance. This observation motivates us to model the articulation transition dynamic features for PD detection.

IV. THE MODEL DESCRIPTION

The architecture of the proposed Bidirectional LSTMs model using dynamic speech features for PD detection is shown in Fig. 5.

We applied Bidirectional LSTMs model to capture time-series characteristics of speech signals for detecting PD. The Bidirectional LSTMs model takes the dynamic time-series Articulation Features (AFs) of the speech signal as inputs. The AFs for each articulation transition contain 58 measures,

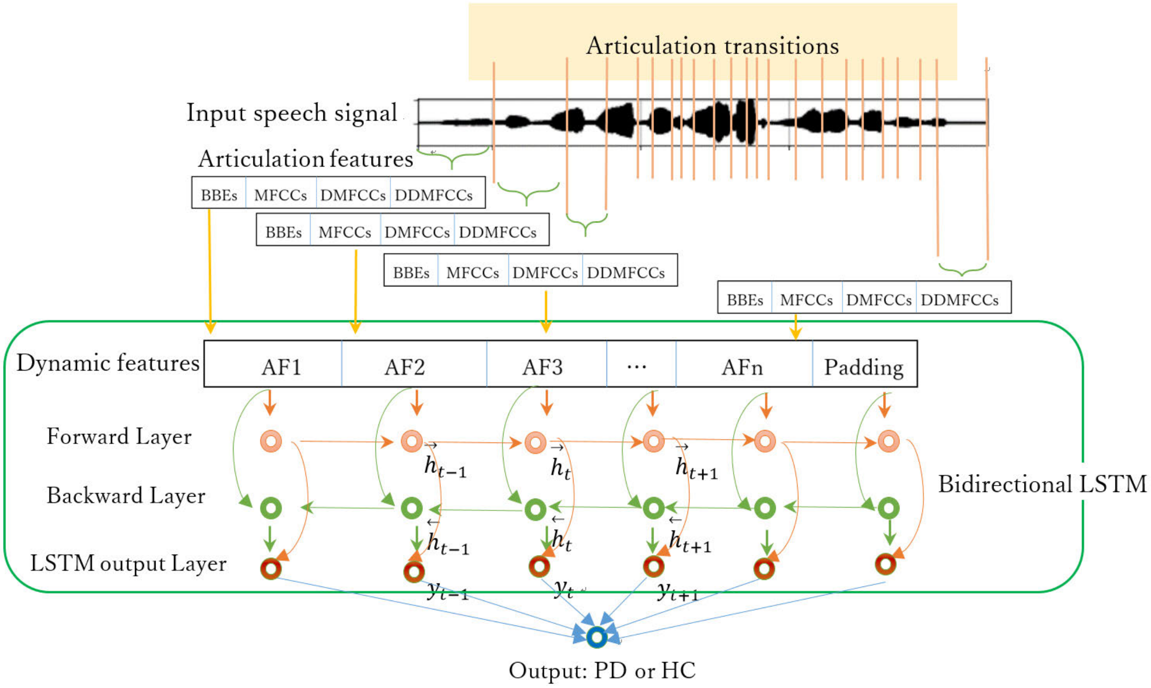


FIGURE 5. Architecture of the Bidirectional LSTMs model using dynamic speech features for PD detection. The input of the model is the dynamic features of a speech signal. The output of the model is the category of PD or HC for the input speech signal.

including 22 BBEs (Bark band energies), 12 MFCCs (Mel-Frequency Cepstral Coefficients), 12 DMFCCs (the first derivative of the MFCCs), and 12 DDMFCCs (the second derivative of the MFCCs) [42]. All sequences of the dynamic features will be zero-padded to the same length before they are fed into the Bidirectional LSTMs model.

The Bidirectional LSTMs network is a combination of Bidirectional recurrent neural networks (RNNs) with LSTM cells. As illustrated in Fig. 5, Bidirectional LSTMs compute the forward hidden sequence $\vec{h}$, the backward hidden sequence $\overleftarrow{h}$, and the output sequence y by iterating the forward layer from $t = (1, \ldots, N)$, and the backward layer from $t = (N, \ldots, 1)$ (N denotes the max length of input sequences) and then updating the output layer as follows:

$$\vec{h}_t = S(W_{AF\vec{h}}AF_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1}AF_t + b_{\vec{h}}) \tag{1}$$

$$\overleftarrow{h}_t = S(W_{AF\overleftarrow{h}}AF_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1}AF_t + b_{\overleftarrow{h}}) \tag{2}$$

$$y_t = W_{\vec{h}y}\vec{h}t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \tag{3}$$

where W denotes weight matrices, b denotes bias vectors, and S is the hidden layer function on each element of a vector.

In the Bidirectional LSTM network, each neural network unit is an LSTM cell (Fig. 6):

$$f_t = \sigma(W_{AFf}AF_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4}$$

$$i_t = \sigma(W_{AFi}AF_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{5}$$

$$o_t = \sigma(W_{AFo}AF_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{6}$$

$$c_t = f_tc_{t-1} + i_t\tanh(W_{AFc}AF_t + W_{hc}h_{t-1} + b_c) \tag{7}$$

$$h_t = o_t\tanh(c_t) \tag{8}$$

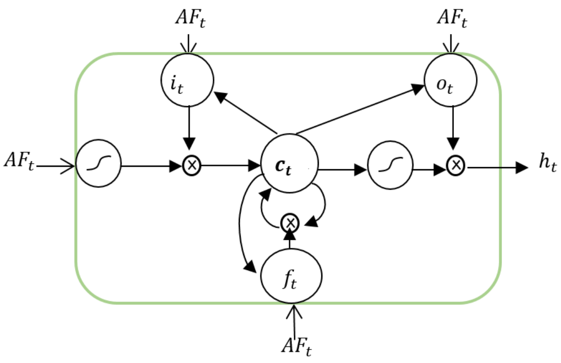


FIGURE 6. Illustration of the long short-term memory (LSTM) cell.

where σ is the logistic sigmoid function, and f_t, i_t, o_t, c_t are the forget gate, input gate, output gate, and cell state, respectively at the time step t.

The Bidirectional LSTM network outputs are fed to a fully connected layer to get the category output of PD or HC.

V. DATASETS AND EXPERIMENTAL SETUP

A. DATASETS AND PREPROCESSING

A mixed gender (25 female, 20 male) database collected contains 45 subjects (15 HC and 30 PD cases) who are hired as volunteers at the GYENNO SCIENCE Parkinson

Disease Research Center.[1] PD cases consist of patients who are suffering from PD with HY (Hoehn and Yahr) stage 1-5. Individual ages vary between 37 and 75. For all subjects, 5-6 voice samples including sustained monophonic /a/ of approximately 5-second duration and a short sentence of approximately 5-second duration are recorded, including 268 samples totally.

The voice signals are acquired by a smartphone and saved as ".wav" files with a frequency 96kHZ. NeuroSpeech software [42] is utilized to extract the speech features.

Several traditional ML and DL models are performed using different speech features and their combinations for comparison. For traditional ML models, Principal Component Analysis (PCA) is applied to compress the dataset onto a lower-dimensional feature subspace.

The experiments are employed under two evaluation methods: 1) 10-fold cross validation (CV); 2) Splitting the dataset into training and testing sets without samples overlap of one individual to ensure unbiased results.

Scikit-learn 0.22.1 is applied to implement the traditional ML models, and Keras 2.2.4 (https://keras.io/) is applied to implement the DL models. The configurations of machine is as follows. GPU: Quadro M1200/PCIe/SSE2; CPU: Intel® Core^{TM} i7-7820HQ CPU @ 2.90GHz × 8; System: Ubuntu 18.04.2 LTS 64-bit Memory, 16 GiB.

B. THE EXPERIMENTS

1) TRADITIONAL ML MODELS FOR PD DETECTION

The parameter settings for traditional ML models are summarized in Table 1. Scikit-Learn [47]'s default settings are applied for the parameters that are not listed in Table 1.

TABLE 1. Parameter settings for ML algorithms.

ML models	Parameters
Decision Tree (DT)	Scikit-learn [47] default settings
Multilayer perceptron (MLP)	Hidden layer sizes =100, 200, 500
K-NearestNeighbor (KNN)	Number of neighbors = 3, 5, 10
Gaussian Naïve Bayes (GNB)	Scikit-learn [47] default settings
Support Vector Machine (SVM)	Kernel = linear, Polynomial, Radial Basis Function (RBF), Sigmoid

Using different static speech features, we compare several traditional ML models. Table 2 lists the dimensions of the speech features and the related component dimensions after PCA.

The evaluation metrics include Accuracy, F-score, Specifity, Sensitivity, Matthews Correlation Coefficient (MCC), Fit_time, and Score_time. The formulations of these metrics are given as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

TABLE 2. The dimensions of speech features and the related component dimensions after PCA.

	Phonation	Articulation	Phonation + Articulation
Dimension of features	29	488	517
Dimension of components after PCA	15	30	30

$$F-score = \frac{2 \times Specifity \times Sensitivity}{Specifity + Sensitivity} \tag{10}$$

$$Specifity = \frac{TN}{TN + FP} \tag{11}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{13}$$

where *TP*, *TN*, *FP*, *FN* are the numbers of true positives, true negatives, false positives, and false negatives. Sensitivity and specificity are statistical measures of correctly classified positive and negative instances. *F-score* is the harmonic mean of precision and recall. *MCC* is a metric used for quantifying the quality of binary classifications with a value between −1 and +1. While a value of +1 indicates a perfect prediction, −1 when there is the disagreement between prediction and actual labels, and 0 when the classification is no better than a random prediction.

Fit_time is the time for fitting the estimator on the training set for each CV split. *Score_time* is the time for scoring the estimator on the testing set for each CV split.

With the input speech signals of monophonic /a/ and a short sentence, Table 3 and Table 4 respectively show the best results achieved by the traditional ML models using different parameter settings (Table 1) and speech features (Articulation, Phonation, and Articulation + Phonation). The results are evaluated by 10-fold CV. The best results are mainly based on the results of *Accuracy*, *F-score*, and *MCC*.

With the input speech signal of monophonic /a/, as shown in Table 3, the best classification *Accuracy* (73.35%), *F-score* (79.67%), and *MCC* (0.3773) are obtained by SVM with linear kernel and using the static articulation features. But the *Fit_time* and *Score_time* of SVM are more than DT and GNB.

Using the static articulation features, Table 3 also shows that MLP, KNN, and SVM obtained better results than using the static phonation features, while DT and GNB obtained better results using the static phonation features.

With the input speech signal of a short sentence, Table 4 shows that the best classification *Accuracy* (73.46%) and *MCC* (0.3909) are achieved by DT using the static phonation features. The best *F-score* (80.96%) is obtained by SVM with RBF kernel and using the static phonation features. But the *Fit_time* of SVM is more than DT.

Additionally, with the input either the speech signal of monophonic /a/ or a short sentence, the combination of the

[1] Ethical Approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and the "Law of the People's Republic of China on Medical Practitioners" (1998) declaration and its later amendments or comparable ethical standards.

TABLE 3. Experimental results on traditional ML models using static features on the input speech signal of monophonic /a/ (10-fold CV).

ML model (feature, parameter setting)	Acc. (%)	F-score (%)	Spe. (%)	Sen. (%)	MCC	Fit_time (s)	Score_time (s)
DT (Phonation)	61.37	69.70	71.30	70.00	0.1304	0.0014	0.0014
MLP (Articulation, hidden layer sizes =500)	68.19	75.59	77.84	75.56	0.2767	0.1668	0.0017
KNN(Articulation, Number of neighbors = 10)	66.70	74.96	75.08	76.67	0.2254	0.0012	0.0030
GNB (Phonation)	61.26	71.60	69.25	75.56	0.0781	0.0003	0.0014
SVM(Articulation, Kernel = Linear)	**73.35**	**79.67**	79.36	82.22	**0.3773**	0.0285	0.0019

TABLE 4. Experimental results of traditional ML models using static features on the input speech signal of a short sentence (10-fold CV).

ML model (feature, parameter setting)	Acc. (%)	F-score (%)	Spe. (%)	Sen. (%)	MCC	Fit_time (s)	Score_time (s)
DT (Phonation)	**73.46**	79.92	80.41	80.00	**0.3909**	0.0010	0.0016
MLP (Phonation, hidden layer sizes =500)	70.44	76.27	80.55	73.33	0.3563	0.2494	0.0017
KNN(Phonation + Articulation, Number of neighbors =3)	69.51	78.86	72.87	87.78	0.2362	0.0013	0.0032
GNB (Phonation)	66.54	70.01	82.25	63.33	0.3388	0.0003	0.0015
SVM(Phonation, Kernel = RBF)	71.10	**80.96**	73.18	91.11	0.2668	0.0032	0.0015

static phonation and articulation features does not contribute much to PD detection.

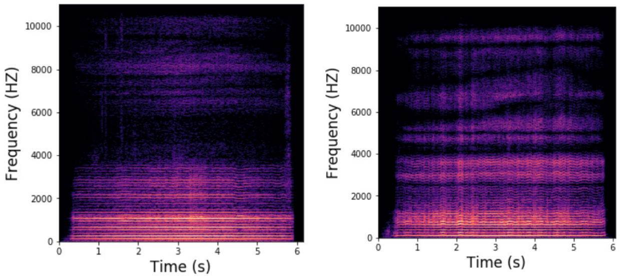
2) DL MODELS FOR PD DETECTION

Previous studies have explored CNN model based methods using static speech features [19], [22] and end-to-end DL architecture [20], [21] for the task of PD detection. In this study, using the input speech signals of monophonic /a/ and a short sentence, we investigate the performance of PD detection by using dynamic speech features under two basic DL models (CNNs and RNNs).

For comparison purposes, the performance of an end-to-end DL architecture using CNN model is evaluated. The audio pre-processing is carried out using librosa [48]. Three main time-frequency representations are extracted: a) linear-scaled short-time Fourier transform (STFT) spectrogram b) Mel-scaled STFT spectrogram, and c) Constant-Q transform (CQT) spectrogram. Fig. 7 shows the STFT spectrograms of the speech signal of monophonic /a/ uttered by a HC speaker (a) and a PD patient (b).

Unlike [20] and [21], onset and offset transitions detection, speech signal rolling and filtering are not included in the pre-processing of the raw data. The input speech signals are transformed to time-frequency representation matrices to feed to a 3-layer CNN model.

Table 5 summarizes the parameter settings of the CNN model and the RNN model. As the main objective of this work is to investigate the role of using dynamic speech features for PD detection, the network structure and the parameters are pre-defined. For the CNN model, three activation functions (Relu, Tanh, and Sigmoid) in convolution layer are tested. The convolutions are performed only in the temporal axis. For the RNN model, three network structures (LSTM, Bidirectional LSTM, and Bidirectional GRU) are tested.



(a) HC speaker

(b) PD patient

FIGURE 7. STFT spectrograms of the speech signal of monophonic /a/ uttered by a HC speaker (a) and a PD patient (b).

Under different parameter settings (Table 5), Table 6 shows the best results of DL models (CNNs and RNNs) using dynamic speech features and end-to-end DL using CNN model. The results are evaluated on 10-fold cross validation (CV). The best results are mainly based on the results of *Accuracy*, *F-score*, and *MCC*.

It is can be found from Table 6, the best results of *Accuracy*, *F-score*, and *MCC* are achieved by the Bidirectional LSTM model using dynamic articulation features on a short sentence. But it took more computation time for a single epoch than CNNs. In comparison with traditional ML models using static features (Table 3 and Table 4), the basic DL

TABLE 5. Parameter settings of DL models.

	Parameters		
CNN model	Num. of filters	Kernel size	Pooling size
Convolution layer 1	64	5	3
Convolution layer 2	128	3	3
Convolution layer 3	256	3	3
Dropout rate after each max pooling layer		0.2	
Activation function in convolution layer		Relu / Tanh / Sigmoid	
Input shape of dynamic features			
	Phonation	Articulation	
Input: monophonic /a/	(569 , 7)	(53 , 58)	
Input: a short sentence	(200 , 7)	(38 , 58)	
RNN model	LSTM / Bidirectional LSTM / Bidirectional GRU		
Number of units		100	
Dropout rate		0.2	
Recurrent dropout rate		0.2	
Num. of timesteps / Input_dim of dynamic features			
		Phonation	Articulation
Input: monophonic /a/		569 / 7	53 / 58
Input: a short sentence		200 / 7	38 / 58
End-to-end DL using CNN			
Time-frequency representation	a) linear-scaled short-time Fourier transform (STFT) spectrogram b) Mel-scaled STFT spectrogram c) constant-Q transform (CQT) spectrogram.		
Common			
Num. of epochs		100	
Training batch size		32	
Optimization		Adam	
Learning rate		1e-4	

models using dynamic features significantly improve the performance.

In Table 6, the results of the DL models using dynamic speech features are higher on a short sentence than on Monophonic /a/, but different results are obtained by the end-to-end DL using CNN model, i.e. the results on Monophonic /a/ are much higher than the results on a short sentence.

It is also observed that with the input of sustained monophonic /a/, the end-to-end DL using CNN model obtained slightly better results than the DL models using dynamic speech features. However, the results of the end-to-end DL using CNN model obtained on a short sentence are much lower than the DL models using dynamic speech features.

As leave-one-out cross validation may result in biased results in performance evaluation in case of having multiple recordings per individual [49]. In the following experiments, we split the dataset into two parts: training set (89 samples) and testing set (45 samples) with ratios 6:4. And all the voice samples of one individual is used only for training or only for testing without overlap.

Based on the results of Table 6, we further explore the performance of DL models by hyperparameter tuning. As a reference, Table 7 shows the experimental results of traditional ML models using static articulation features.

With the same architecture of the networks (Table 5) of the CNNs and Bidirectional LSTM, Talos [50] is utilized to perform hyperparameter tuning for DL models on the training set. Table 8 lists the hyperparameter search space for DL models.

Table 9 presents the best experimental results based on hyperparameter tuning for DL models. In comparison with traditional ML models using static articulation features (Table 7), the basic DL models using dynamic speech features showed significant improvement, especially the Bidirectional LSTM model using dynamic speech features, giving an accuracy of 75.56%. The results also showed that the end-to-end DL using CNN model obtained an accuracy of 71.11% with the input of monophonic /a/, giving a much better result than the CNN model using dynamic speech features. This indicates that time-frequency representations are useful as learning features for PD detection. Concerning the time complexity, more fitting time is required for Bidirectional LSTM model than the CNNs model.

Under the two evaluation methods of 10-fold CV (Table 6) and splitting the dataset without samples overlap of one individual (Table 9), the experimental results showed that the proposed method improves the accuracy of PD detection over traditional ML models using static features and the end-to-end DL using CNN model.

Fig. 8 illustrates the accuracy increasing trends of the Bidirectional LSTM model as the number of epochs increase for training and validation.

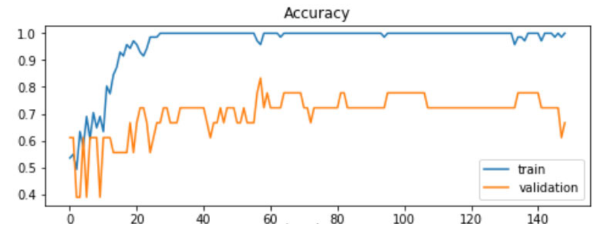


FIGURE 8. Illustration of the accuracy increasing trends of Bidirectional LSTM model as the number of epochs increase for training and validation.

VI. DISCUSSION

Detection of voice changes in PD patients through ML methods has been shown to be a promising way for Parkinson's early detection. In the task of PD detection from speech, the performance of a ML based method is mainly affected by the speech features and the architecture of ML models. In this study, we explored static and dynamic speech features relating to PD detection. A comparative analysis of the articulation transition characteristics shows that the number

TABLE 6. Experimental results of DL models using dynamic speech features and end-to-end DL using CNN model (10-fold CV).

Input speech	Input signals	Models (features/learning method, activation function, model)		Acc. (%)	F-score (%)	Spe. (%)	Sen. (%)	Mcc	Epoch_time (s)
Monophonic /a/	Speech (.wav)	CNNs	Articulation, Activation: relu	74.29	80.75	83.14	80.85	0.4326	0.0398
		RNNs	Articulation, Bidirectional LSTM	77.36	83.85	88.75	80.36	0.4751	0.1959
	Spectrogram (.npy)	End-to-end DL using CNN	Mel-scaled spectrogram, Activation: relu	80.99	87.00	92.22	83.15	0.5319	0.1019
A short sentence	Speech(.wav)	CNNs	Articulation, Activation: relu	83.52	87.48	87.78	88.53	0.6478	0.0511
		RNNs	Articulation, Bidirectional LSTM	**84.29**	**88.52**	91.11	87.34	**0.6603**	0.1572
	Spectrogram (.npy)	End-to-end DL using CNN	linear-scaled STFT spectrogram, Activation: relu	65.24	74.95	80.56	71.20	0.1765	0.1853

TABLE 7. Experimental results of traditional ML models using static articulation features with the input speech signal of a short sentence (training: testing = 6:4, without overlap of the voice samples of one individual).

ML model (parameters)	Acc. (%)	F-score (%)	Spe. (%)	Sen. (%)	MCC
DT	48.89	62.30	63.33	61.29	-0.1697
MLP (hidden layer sizes=200)	55.56	76.27	80.55	73.33	-0.0347
KNN (Number of neighbors=5)	60.00	67.74	70.00	65.63	-0.0434
GNB	44.44	59.02	60.00	58.06	-0.2715
SVM(Kernel=poly)	**64.44**	78.38	97.67	65.91	-0.1066

TABLE 8. Hyperparameter search space for DL models.

DL models Hyperparameter	Hyperparameter search space
CNNs	
Num. of filters	32, 64, 128, 200
Dropout rate after each max pooling layer	0, 0.1, 0.2, 0.5, 0.5
Activation function in convolution layer	Relu, sigmoid, tanh, elu
Bidirectional LSTM	
Num. of units	20, 50, 100, 200
Dropout rate	0, 0.1, 0.2, 0.5
Recurrent dropout rate	0, 0.1, 0.2, 0.5
Common	
Optimizer	Adam, Nadam, Adagrad, SGD
Learning rate	1, 0.1, 1e-2, 1e-3, 1e-4
Num. of epochs	10, 30, 50, 100, 150
Training batch size	10, 30, 50, 100, 150, 200

of articulation transitions and the trend of the fundamental frequency curve between HC speakers and PD patients are significantly different. We applied a paired t-test to evaluate the difference on the number of articulation transitions between HC speaker and PD patient groups and got a p-value of 0.042 (<0.05), which indicates the difference did not occur by chance. This observation motivated us to model the dynamic speech features for PD detection.

In the experiments of using static speech features, we compared traditional ML models including DT, MLP, KNN, GNB and SVM for PD detection. The experimental results based on 10-fold CV showed that, with the input speech signal of monophonic /a/, the best classification accuracy (73.35%) is obtained by SVM with linear kernel and using static articulation features. With the input speech signal of a short sentence, DT obtained the highest classification accuracy (73.46%) by using static phonation features.

In the experiments of using dynamic speech features, we experiment the basic DL models including CNNs and Bidirectional LSTM. With the input speech signal of a short sentence, both CNNs and Bidirectional LSTM achieved significant improvement in classification accuracy (83.52% for CNNs and 84.29% for Bidirectional LSTM).

In addition, we implemented the performance comparison between DL models (CNNs and Bidirectional LSTM) using dynamic speech features and the end-to-end DL using CNN model. With the input of sustained monophonic /a/, the end-to-end DL obtained better results than the DL models using dynamic speech features, but the *Accuracy* showed a decrease when using the input of a short sentence. A Multi-scale CNN [51] that can catch the time-frequency feature

TABLE 9. Experimental results of DL models using dynamic speech features and end-to-end DL using CNN model (training: testing = 6:4, without overlap of the voice samples of one individual).

DL models (Input contents, learning method)	Hyperparameter setting	Acc. (%)	F-score (%)	Spe. (%)	Sen. (%)	MCC	Epoch_time (s)
CNNs (a short sentence, dynamic feature extraction)	(Num. of filters, Kernel size, Pooling size, Activation function, Dropout rate)	66.67	80.00	100	66.67	0.0	0.0131
The 1st CNN layer	(200, 5, 3, sigmoid, 0.2)						
The 2nd CNN layer	(64, 3, 3, sigmoid, 0.0)						
The 3rd CNN layer	(128, 3, 3, elu, 0.2)						
Optimizer	Aadam						
Learning rate	0.1						
Num. of epochs	50						
Training batch size	150						
end-to-end DL using CNN (Monophonic /a/, Mel-scaled spectrogram)	(Num. of filters, Kernel size, Pooling size, Activation function, Dropout rate)	71.11	78.69	80.00	77.42	0.3394	0.0708
The 1st CNN layer	(64, 5, 3, sigmoid, 0.2)						
The 2nd CNN layer	(128, 3, 3, relu, 0.2)						
The 3rd CNN layer	(256, 3, 3, relu, 0.2)						
Optimizer	Aadam						
Learning rate	1e-4						
Num. of epochs	150						
Training batch size	30						
Bidirectional LSTM (a short sentence, dynamic feature extraction)	(Num. of units, Dropout rate, Recurrent dropout rate)	**75.56**	**80.70**	76.67	85.19	**0.4811**	0.1097
The 1st Bidirectional LSTM layer	(20, 0.1, 0.2)						
The 2nd Bidirectional LSTM layer:	(200, 0.1, 0.1)						
Optimizer	Adagrad						
Learning rate	0.1						
Num. of epochs	150						
Training batch size	20						
Num. of timesteps	38						
Input_dim	58						

representations at different time scales, or adding the process of onset and offset transitions detection [20], or adding speech signal rolling and filtering [21] may improve the performance. However, a heavy audio pre-processing stage is required before the DL model training.

As leave-one-out cross validation may result in biased results in performance evaluation. By splitting the dataset into training and testing sets without samples overlap of one individual, we further explore the performance of DL models by hyperparameter tuning. In comparison with traditional ML models using static articulation features, the basic DL models showed great improvement using dynamic speech features, especially the Bidirectional LSTM model. In addition, with the input of monophonic /a/, the end-to-end DL using CNN model showed a much better result than the CNN model using dynamic speech features. The results suggest that a PD detection system would benefit from combining the two methods (Bidirectional LSTM model using dynamic speech features and the end-to-end DL using CNN model) to make robust predictors and improve the system flexibility on the input content.

As more complex network architectures (e.g. a deep hybrid model with more layers, or a deep reinforcement learning model) have not been experiment in this study, it is can see still space for further improvement of the DL model architecture.

In addition to the DL model architecture, the speech features also have an effect on the classification performance. Previous studies have experiment several different static speech features (such as MFCCs, energy content, pitch).

TABLE 10. Summary of the recent studies of PD detection using ML models.

Ref	Datasets	Features / Time–frequency representation	ML models for Comparison	The best model Acc. (%)
Lahmiri et al. (2017) [12]	147 PD and 48 HC, 195 samples	22 static features (dysphonia measurements)	Linear discriminant analysis (LDA), KNN, Naı¨ve Bayes (NB), Regression trees (RT), Radial basis function neural networks (RBFNN), SVM, and Mahalanobis distance	SVM 92±2
Saloni et al. (2015) [13]	23 PD and 8 HC (36 second sustained vowel 'a'), 195 samples	15 static features	-	SVM 100
Shahbakhi et al. (2014) [14]	23 PD and 8 HC, 195 samples	22 static features	-	Genetic Algorithm and SVM 94.50 ± 3.54
Li et al. (2017) [16]	20 PD and 20 HC, 1040 samples	26 static features	Single classifier with Classification and Regression Tree (CART), Ensemble learning algorithm with CART, SVM, KNN, NB	Ensemble learning algorithm with CART 90.00
Gunduz et al. (2019) [19]	188 PD and 64 HC	752 static features	CNN, SVM	A 9-layered CNN 86.9
Vásquez-Correa et al. (2017) [20]	1) 50 PD and 50 HC (Spanish) 2) 88 PD and 88 HC (German) 3) 20 PD and 15 HC (Czech)	the short time Fourier transform (STFT), the continuous wavelet transform (CWT)	-	CNN with CWT on Czech 89.4
Wodzinski et al. (2019) [21]	50 PD and 50 HC	Spectrogram of the voice signal	-	ResNet 91.7
Vaiciukynas et al. (2018) [22]	194 PD and 74 HC, sentence segment, transitions between words when split on syllable, pairs of words, and full sentence	9 static features	-	A 4-layered CNN 85.9
Nilashi et al. (2016) [54]	1)42 subjects, 5875 samples 2)23 PD and 8 HC, 195 samples	16 static features	Principal Component Analysis (PCA)-Neural Networks (NN), PCA-Adaptive Neuro-Fuzzy Inference System (ANFIS), Expectation Maximization (EM)-PCA-ANFIS	PCA-ANFIS 99.72
Grover et al. (2018) [34]	42 subjects, 5875 samples	16 static features	-	Deep Neural Network (DNN) 81.67 (Severity prediction)

In this study, the articulation transition dynamic features are extracted such that the DL models can capture time-series characteristics of a continuous speech signal. With more speech features and their combinations, the performance improvement can be expected.

The difference on the experiment subjects, the languages, the content of the input speech, and the preprocessing strategies make it difficult to compare the performance directly. This study did not compare the results with other related studies directly. Instead, under the same experimental environment, our experiments covered most ML models that have been applied in the previous studies. It would be more objective to compare the performance of different ML models.

Table 10 summarizes the recent studies of PD detection using ML models for referring to. Several studies [12]–[14], [19] used the same dataset [52] to conduct comparative performance analysis of different ML models and feature selection approaches in distinguishing between HC and PD patients. Based on dysphonia symptoms, Lahmiri *et al.* [12]'s study showed that the SVM classifier achieved higher average performance than other traditional ML models by means of 10-fold CV evaluation. This result is in consistent with our results even though the raw speech signals are from two different languages. DL based methods are not discussed in this study.

Shahbakhi *et al.* [14] proposed a feature selection method using genetic algorithm (GA) for PD detection. Similarly, Saloni and Gupta [13] proposed a feature combination method for finding superior feature subsets. Both studies reported high classification accuracies based on the evaluation of splitting the dataset into two parts (training: testing = 3:1). However, the overlap of the voice samples of one individual between the training and testing subsets may result in biased results in performance evaluation.

Using Sakar *et al.*'s dataset [53], in Li and Wang [16], a single classifier with Classification and Regression Tree (CART) based sample selection algorithm and an ensemble learning algorithm were proposed for PD classification. Under the evaluation of leave-one-subject-out (LOSO), the best accuracy reported in this study was 90%, but when using Leave-one-out evaluation method (LOO), the average accuracy dropped to 74.50%.

Gunduz [19] proposed two frameworks (feature-level combination and model-level combination) based on deep CNNs (9-layered) to classify PD using sets of speech features. The best accuracy (86.9%) was reported by using the model-level combination method. But there are only 1.2% improvement in terms of accuracy when comparing with the accuracy (85.7%) obtained by the baseline SVM model. By contrast, our method outperformed the SVM model more than 11% in terms of accuracy.

Vásquez-Correa *et al.* [20] and Wodzinski *et al.* (2019) [21] explored end-to-end deep learning approaches using CNN for audio classification. Traditional feature engineering is not required in the end-to-end DL framework; instead, an audio signal pre-processing (the detection of the onset and offset transitions [20], signal rolling and filtering [21]) stage is required before the DL model training in order to achieve high performance.

Vaiciukynas *et al.* [22] proposed to use spectrograms and short-term features for PD detection in Lithuanian based on a 4-layered CNN model. This study focused on the performance comparison of various sentence segments in PD detection.

Nilashi *et al.* [54] and Grover *et al.* [34] respectively focused on predicting Parkinson's disease progression and severity of PD using DL methods. The multi-label PD classification will be included in our future work.

VII. CONCLUSION

In this paper, using speech signals, a deep learning based method is proposed for PD detection. The proposed method innovatively combines the dynamic articulation transition features with Bidirectional LSTM model to capture time-series characteristics of continuous speech signals.

Under the two evaluation methods of 10-fold cross validation (CV) and splitting the dataset without samples overlap of one individual, the experimental results showed that the proposed method remarkably improves the accuracy of PD detection over traditional machine learning models using static features.

For future work, we will apply the proposed method for stage classification of PD to explore its applicability in the multi-label classification task, and consider a more complex DL network architectures (e.g. a deep hybrid model with more layers, or a deep reinforcement learning model) to improve the performance.

REFERENCES

[1] K. Wirdefeldt, H. Adami, P. Cole, D. Trichopoulos, and J. Mandel, "Epidemiology and etiology of Parkinson's disease: A review of the evidence," *Eur. J. Epidemiol.*, vol. 26, no. S1, pp. S1–S58, 2011.

[2] A. Q. Rana, U. S. Ahmed, Z. M. Chaudry, and S. Vasan, "Parkinson's disease: A review of non-motor symptoms," *J. Expert Rev. Neurotherapeutics*, vol. 15, no. 5, pp. 549–562, 2015.

[3] S. Perez-Lloret, L. Nègre-Pagès, A. Ojero-Senard, P. Damier, A. Destée, F. Tison, M. Merello, and O. Rascol, "Oro-buccal symptoms (dysphagia, dysarthria, and sialorrhea) in patients with Parkinson's disease: Preliminary analysis from the French COPARK cohort," *Eur. J. Neurol.*, vol. 19, no. 1, pp. 28–37, 2012.

[4] V. Delić, Z. Peric, M. Secujski, N. Jakovljevic, J. Nikolic, D. Miškovic, N. Simic, S. Suzic, and T. Delic, "Speech technology progress based on new machine learning paradigm," *Comput. Intell. Neurosci.*, vol. 219, pp. 1–19, Jun. 2019, doi: 10.1155/2019/4368036.

[5] K. Chenausky, J. MacAuslan, and R. Goldhor, "Acoustic analysis of PD speech," *Parkinsons Disease*, vol. 2011, pp. 1–13, Oct. 2011, doi: 10.4061/2011/435232.

[6] C. O. Sakar and O. Kursun, "Telediagnosis of Parkinson's disease using measurements of dysphonia," *J. Med. Syst.*, vol. 34, no. 4, pp. 591–599, Aug. 2010, doi: 10.1007/s10916-009-9272-y.

[7] M. Can, "Neural networks to diagnose the Parkinson's disease," *Southeast Eur. J. Soft Comput.*, vol. 2, no. 1, pp. 68–75, 2013.

[8] J. S. Almeida, P. P. R. Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, and V. H. C. de Albuquerque, "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognit. Lett.*, vol. 125, pp. 55–62, Jul. 2019.

[9] F. Åström and R. Koker, "A parallel neural network approach to prediction of Parkinson's disease," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 12470–12474, Sep. 2011, doi: 10.1016/j.eswa.2011.04.028.

[10] C. Ma, J. Ouyang, H. Chen, and X. Zhao, "An efficient diagnosis system for Parkinson's disease using kernel-based extreme learning machine with subtractive clustering features weighting approach," *Comput. Math. Methods Med.*, vol. 2014, Nov. 2014, Art. no. 985789, doi: 10.1155/2014/985789.

[11] S. Lahmiri, "Parkinson's disease detection based on dysphonia measurements," *Phys. A, Stat. Mech. Appl.*, vol. 471, pp. 98–105, Apr. 2016, doi: 10.1016/j.physa.2016.12.009

[12] S. Lahmiri, D. A. Dawson, and A. Shmuel, "Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures," *Biomed. Eng. Lett.*, vol. 8, no. 1, pp. 29–39, Feb. 2018, doi: 10.1007/s13534-017-0051-2.

[13] R. K. S. Saloni and A. K. Gupta, "Detection of Parkinson disease using clinical voice data mining," *Int. J. Circuits, Syst. Signal Process.*, vol. 9, pp. 320–326, Jan. 2015.

[14] M. Shahbakhi, D. T. Far, and E. Tahami, "Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine," *J. Biomed. Sci. Eng.*, vol. 7, no. 4, pp. 147–156, 2014, doi: 10.4236/jbise.2014.74019.

[15] D. Meghraoui, B. Boudraa, T. Merazi-Meksen, and M. Boudraa, "Parkinson's disease recognition by speech acoustic parameters classification," in *Modelling and Implementation of Complex Systems*. Cham, Switzerland: Springer, 2016, pp. 165–173.

[16] Y. Li, L. Yang, P. Wang, C. Zhang, J. Xiao, Y. Zhang, and M. Qiu, "Classification of Parkinson's disease by decision tree based instance selection and ensemble learning algorithms," *J. Med. Imag. Health Informat.*, vol. 7, no. 2, pp. 444–452, Apr. 2017.

[17] E. Vaiciukynas, A. Verikas, A. Gelzinis, and M. Bacauskiene, "Detecting Parkinson's disease from sustained phonation and speech signals," *PLoS ONE*, vol. 12, no. 10, Oct. 2017, Art. no. e0185613, doi: 10.1371/journal.pone.0185613.

[18] H. K. Rouzbahani and M. R. Daliri, "Diagnosis of Parkinson's disease in human using voice signals," *Basic Clin. Neurosci.*, vol. 2, no. 3, pp. 12–20, 2011.

[19] H. Gunduz, "Deep learning-based Parkinson's disease classification using vocal feature sets," *IEEE Access*, vol. 7, pp. 115540–115551, 2019, doi: 10.1109/ACCESS.2019.2936564.

[20] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with Parkinson's disease," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 314–318.

[21] 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 *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Berlin, Germany, Jul. 2019 pp. 717-720.

[22] E. Vaiciukynas, A. Gelzinis, A. Verikas, and M. Bacauskiene, "Parkinson's disease detection from speech using convolutional neural networks," in *Smart Objects and Technologies for Social Good* (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), vol. 233, B. Guidi, L. Ricci, C. Calafate, O. Gaggi, and J. Marquez-Barja, Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-319-76111-4_21.

[23] A. Pompili, A. Abad, P. Romano, and I. P. Martins, "Automatic detection of Parkinson's disease: An experimental analysis of common speech production tasks used for diagnosis," in *Proc. 20th Int. Conf. Text, Speech, Dialogue (TSD)*, Prague, Czech Republic, Aug. 2017, pp. 411–419.

[24] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *J. Speech Lang. Hearing Res.*, vol. 12, no. 2, pp. 246–269, 1969.

[25] D. G. Hanson, B. R. Gerratt, and P. H. Ward, "Cinegraphic observations of laryngeal function in Parkinson's disease," *Laryngoscope*, vol. 94, no. 3, pp. 53–348, 1984.

[26] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in Parkinson's disease," *J. Voice*, vol. 25, no. 4, pp. 72–467, 2011.

[27] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, Klempir, V. Majerova, P J. icmausova, J. Roth, E. Ruzicka, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *J. Acoust. Soc. Amer.*, vol. 134, no. 3, pp. 81–2171, 2013.

[28] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolanos, J. D. Arias-Londono, J. F. Vargas-Bonilla, T. Haderlein, and E. Nöth, "Phonation and articulation analysis of Spanish vowels for automatic detection of Parkinson's disease," in *Text, Speech and Dialogue*. Cham, Switzerland: Springer, 2014, pp. 374–381.

[29] T. Arias-Vergara, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Parkinson's disease and aging: Analysis of their effect in phonation and articulation of speech," *Cognit. Comput.*, vol. 9, no. 6, pp. 731–748, Dec. 2017, doi: 10.1007/s12559-017-9497-x.

[30] L. Berus, S. Klancnik, M. Brezocnik, and M. Ficko, "Classifying Parkinson's disease based on acoustic measures using artificial neural networks," *Sensors*, vol. 19, no. 1, pp. 1–15, 2019, doi: 10.3390/s19010016.

[31] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H. Ghomi, "Parkinson's disease diagnosis using machine learning and voice," in *Proc. IEEE Signal Process. Med. Biol. Symp. (SPMB)*, Philadelphia, PA, USA, Dec. 2018, pp. 1–7, doi: 10.1109/SPMB.2018.8615607.

[32] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Barcelona, Spain, Oct. 2013, pp. 3–10.

[33] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Barcelona, Spain, Oct. 2013, pp. 835–838.

[34] S. Grover, S. Bhartia, A. Yadav, and K. R. Seeja, "Predicting severity of Parkinson's disease using deep learning," *Procedia Comput. Sci.*, vol. 132, pp. 1788–1794, May 2018.

[35] A. Krizhevsky, S. Ilya, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[36] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2015.

[37] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, May 2019.

[38] K.-I. Funahashi and Y. Nakamura, "Approximation of dynamical systems by continuous time recurrent neural networks," *Neural Netw.*, vol. 6, no. 6, pp. 801–806, Jan. 1993.

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

[40] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease," *J. Commun. Disorders*, vol. 76, pp. 21–36, Nov. 2018.

[41] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, A. Vann, N. Vogler, T. Bocklet, M. Cernak, J. Hannink, and E. Nöth, "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digit. Signal Process.*, vol. 77, pp. 207–221, Jun. 2018.

[42] *NeuroSpeech*. Accessed: May 7, 2020. [Online]. Available: https://github.com/jcvasquezc/NeuroSpeech

[43] H. Ackermann and W. Ziegler, "Articulatory deficits in parkinsonian dysarthria: An acoustic analysis," *J. Neurol., Neurosurgery Psychiatry*, vol. 54, no. 12, pp. 1093–1098, Dec. 1991.

[44] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolanos, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Honig, and E. Nöth, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1820–1828, Nov. 2015.

[45] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. 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," *J. Acoust. Soc. Amer.*, vol. 139, no. 1, pp. 481–500, Jan. 2016.

[46] J. R. Orozco-Arroyave, *Analysis of Speech of People With Parkinson's Disease*. Berlin, Germany: Logos Verlag, 2016.

[47] *Scikit-Learn*. Accessed: May 7, 2020. [Online]. Available: https://scikit-learn.org/stable/

[48] B. McFee, M. McVicar, C. Raffel, D. Liang, O. Nieto, J. Moore, D. Ellis, D. Repetto, P. Viktorin, and J. F. Santos. (2017). *Librosa: V0.5.0*. Accessed: Dec. 20, 2020, doi: 10.5281/zenodo.293021.

[49] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul, and H. Apaydin, "A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable Q-factor wavelet transform," *Appl. Soft Comput.*, vol. 74, pp. 255–263, Jan. 2019.

[50] (2019). *Autonomio Talos Computer Software*. Accessed: Oct. 5, 2020. [Online]. Available: http://github.com/autonomio/talos

[51] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classification," 2016, *arXiv:1603.06995*. [Online]. Available: http://arxiv.org/abs/1603.06995

[52] A. M. Little, P. E. Macsharry, E. J. Hunter, J. Sielman, and L. O. Raming, "Suitability of dysophonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[53] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a parkinson speech dataset with multiple types of sound recordings," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 4, pp. 828–834, Jul. 2013.

[54] M. Nilashi, O. Ibrahim, and A. Aha, "Accuracy improvement for predicting Parkinson's disease progression," *Sci. Rep.*, vol. 6, Sep. 2016, Art. no. 34181, doi: 10.1038/srep34181.

CHANGQIN QUAN received the M.E. degree from Central China Normal University, Wuhan, China, in 2005, and the Ph.D. degree from the University of Tokushima, Tokushima, Japan, in 2011. She is currently an Associate Professor with Kobe University. Her research interests include machine learning algorithms, natural language processing, human–computer interface, and medical bioinformatics.

KANG REN received the M.E. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2008. He is currently pursuing the Ph.D. degree with Kobe University. His research interests include system control, human–computer interface, and medical engineering.

ZHIWEI LUO received the M.E. and Ph.D. degrees from Nagoya University, Japan, in 1991 and 1992, respectively. He was an Assistant Professor with the Toyohashi University of Technology, in 1992, a Frontier Researcher of RIKEN, in 1994, and an Associate Professor with Yamagata University, in 1999. He was a Team Leader of RIKEN, in 2001, where he leaded the development of the world first human care robot RI-MAN. Since 2006, he has been a Professor with Kobe University. His research interests include system control, robotics, human–computer interface, and health engineering. He was honored as a Fellow of SICE, in 2016. He is currently a Board Member of SCI, Japan.