Phonological posteriors and GRU recurrent units to assess speech impairments of patients with Parkinson's disease

J. C. Vásquez-Correa[1,2*], N. Garcia-Ospina[1], J. R. Orozco-Arroyave[1,2], M. Cernak[3], and E. Nöth[2]

[1]Faculty of Engineering, University of Antioquia UdeA, Medellín, Colombia.
[2]University of Erlangen-Nüremberg, Germany.
[3]Logitech, Laussane, Switzerland.
jcamilo.vasquez@udea.edu.co

Abstract. Parkinson's disease is a neurodegenerative disorder characterized by a variety of motor symptoms, including several impairments in the speech production process. Recent studies show that deep learning models are highly accurate to assess the speech deficits of the patients; however most of the architectures consider static features computed from a complete utterance. Such an approach is not suitable to model the dynamics of the speech signal when the patients pronounce different sounds. Phonological features can be used to characterize the voice quality of the speech, which is highly impaired in patients suffering from Parkinson's disease. This study proposes a deep architecture based on recurrent neural networks with gated recurrent units combined with phonological posteriors to assess the speech deficits of Parkinson's patients. The aim is to model the time-dependence of consecutive phonological posteriors, which follow the sound patterns of English phonological model. The results show that the proposed approach is more accurate than a baseline based on standard acoustic features to assess the speech deficits of the patients.

Key words: Parkinson's disease, dysarthria assessment, phonological posteriors, gated recurrent units, recurrent neural network

1 Introduction

Parkinson's disease (PD) is a neurological disorder characterized by the progressive loss of dopaminergic neurons in the mid-brain, producing several motor and non-motor impairments [1]. The motor symptoms include different speech deficits including reduced loudness, monopitch, monoloudness, reduced stress, breathy, hoarse voice quality, and imprecise articulation. These impairments are grouped together and called *hypokinetic dysarthria* [2]. The disease progression in motor activities is currently evaluated with the third section of the movement disorder society, unified Parkinson's disease rating scale (MDS-UPDRS-III) [3].

Several studies in the literature have described the speech impairments developed by PD patients in terms of four different dimensions: phonation, articulation, prosody, and intelligibility [4, 5]. These feature extraction strategies have shown to be suitable to support the diagnosis process and to assess the neurological state of the patients. Although the success of these classical feature extraction approaches, in the recent years deep learning methods have shown to be highly accurate to assess the speech of PD patients [6, 7]. In [6] the authors proposed a deep learning model to assess the severity of dysarthria in speech. The model considers an intermediate interpretable hidden layer to assess four perceptual dimensions: nasality, vocal quality, articulatory precision, and prosody. The authors reported a Spearman's correlation of up to 0.82 between the output of the deep learning model and a perceptual score of the severity of dysarthria provided by speech and language therapists. In [7] the authors considered a convolutional neural network and time-frequency representations to model articulation impairments in the speech of PD patients [7]. The model classified PD patients vs. healthy control subjects with accuracies of up to 89%. To the best of our knowledge, most of the related studies that consider deep learning methods to assess pathological speech only consider static features computed from a complete utterance. Those methods are not able to model the dynamics of the speech signal properly when the patients pronounce different sounds in continuous speech such as sentences or monologues. From the different deep learning architectures, the recurrent neural networks (RNN) have been designed to process the time-dependence of sequential inputs such as text or speech, which makes it suitable to model the dynamics of speech features computed for different frames. On the other hand, voice quality of the speech can be characterized using phonological features [8], by computing phonological posterior features for modal and non-modal phonations on consecutive speech frames.

This study combines the phonological analysis and RNNs to model the dynamics of the features computed on consecutive frames to assess the speech deficits of PD patients. RNNs are formed with gate recurrent units (GRUs) [9], which have shown similar results than the standard long short-term memory (LSTM) units, but with less parameters to learn. The results indicate that the dynamics phonological posterior features are better to model the speech impairments of PD patients than the standard acoustic features based on Mel frequency Cepstral coefficients (MFCCs). In addition, the phonological posteriors provide interpretable results for the medical examiner to evaluate the speech state of the patients.

2 Methods

2.1 Phonological Features extraction

Phonological features were extracted using the deep learning approach from [10]. This process involves the following steps: (1) the speech signal is segmented in short-time frames, (2) 13 MFCCs and their derivatives are computed for every frame of the speech signal, and (3) a set of 15 pre-trained DNNs infers the

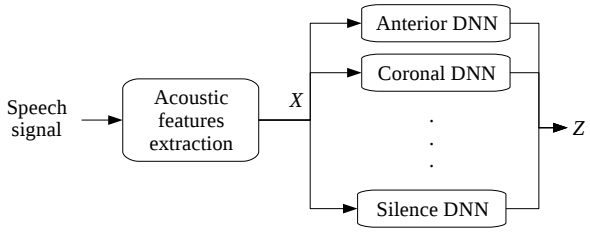

Fig. 1. Phonological feature extraction process

phonological posteriors from the acoustic feature vector. These posteriors are concatenated in a phonological feature vector z_t. The process is summarized in Figure 1, where X is the set of acoustic features and Z is the set of phonological features. A total of 14 phonological features are computed. Table 1 shows details of each phonological feature.

Table 1. List of phonological features

Feature	Brief description
Vocalic	Refers to the vocal folds vibration without constriction in the vocal tract.
Consonantal	Indicates sounds where there is an obstruction of the vocal tract.
High	The body of the tongue is above its neutral position.
Back	The body of the tongue is retracted from its neutral position.
Low	The body of the tongue is below its neutral position.
Anterior	Indicates an obstruction located in front of the palato-alveolar region of the mouth.
Coronal	The blade of the tongue is raised from its neutral position.
Round	Refers to narrowed lips.
Rising	Differentiates diphthongs from monophthongs.
Tense	Indicates stressed vowels.
Voice	Indicates voiced sounds.
Continuant	Differentiates plosives from non-plosives.
Nasal	Indicates a lowered velum, where the air to escape through the nose.
Strident	Refers to sounds with more energy in high frequency components.
Silence	Tells that there is no speech in the frame.

2.2 Recurrent neural network and GRU units

The RNNs process the sequence one element at a time, a state vector in their hidden units, which contains information about the history of all the past elements of the sequence [11]. The RNNs can be formed with different recurrent units including the conventional recurrent units, the long short-term memory (LSTM) units, or the gated recurrent units (GRUs). The conventional recurrent units can be seen as very deep networks where all the layers share the same weights. Although their main purpose is to learn long-term dependencies, there

is evidence that shows difficulties to learn very long sequences [12]. This problem may be fixed with the LSTM units, which have a memory cell to model the long-term time-dependency. The GRU were proposed as a modification of the LSTM replacing the separate input and forget gates with a reset gait to control the input information to the network. GRUs and LSTMs have provided similar results for several tasks including speech and language modeling [13]; however, the GRUs are faster to train and require less parameters [14], which make these units more suitable to be used when less train data is available.

Figure 2 shows the architecture used in this study to process the phonological posteriors sequences and the MFCC feature vectors. The phonological features are processed individually by two GRU layers. On the other path, classical MFCC features are modeled by other two GRU units. The ouput of the two paths are merged with two fully connected layers (h1, and h2), followed by the output layer to make the final decision. Three architectures are considered in this study: (1) a network to process only the phonological posteriors, (2) a network to process only the acoustic features, and (3) a network to combine the phonological and acoustic features (see Figure 2).

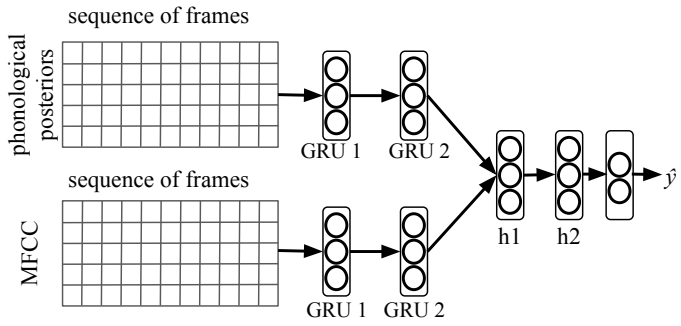

Fig. 2. Deep architecture to assess speech impairments of PD patients using phonological posteriors and GRU units

2.3 Validation

The experiments are validated using 80% of the data for training, 10% to optimize the hyper-parameters, i.e., development set, and the remaining 10% to test. The process is repeated 10 times with different partitions to produce different and independent test sets. The hyper-parameter tuning is performed with a Bayesian optimization approach [15] due to the large number of hyper-parameters to optimize. The tuning is performed based on an optimization problem, where the hyper-parameters that maximize the performance of the model on the development set are found. The range of the hyper-parameters to be optimized is shown in Table 2. A batch-size of 128 samples and a total of 100 epochs are considered with an early stopping strategy.

Table 2. Range of the hyper-parameters used to train the RNN.

Hyper-parameter	Values
GRU units in all layers	$\{8, 16, 32, 64\}$
Hidden units in fully connected layers	$\{16, 32, 64, 128\}$
Learning rate	$\{0.0001, \cdots, 0.01\}$
Dropout rate	$\{0.1, 0.2 \cdots 0.7\}$
Recurrent dropout rate	$\{0.1, 0.2 \cdots 0.7\}$

3 Data

3.1 m-FDA scale

The evaluation of PD patients according to the MDS-UPDRS-III scale has shown to be suitable to assess general motor impairments of PD patients; however, the deterioration of the communication skills of the PD patients is not properly evaluated because such a scale only considers speech impairments in one of its items. A modified version of the Frenchay dysarthria assessment scale (m-FDA), which can be administered based on speech recordings was recently developed [4, 8]. The scale includes several aspects of speech: respiration, lips movement, palate/velum movement, larynx, tongue, monotonicity, and intelligibility. The scale has a total of 13 items and each of them ranges from 0 (normal or completely healthy) to 4 (very impaired), thus the total score of the scale ranges from 0 to 52. The labeling process of the recordings was performed by three phoniatricians who agreed in the first ten speakers. Afterwards, each phoniatrician evaluated the remaining recordings independently. The inter-rater reliability among the labelers is 0.75.

3.2 Participants

We consider the PC-GITA database [16]. The data contain speech utterances from 50 PD and 50 HC Colombian Spanish native speakers balanced in age and gender. The participants pronounce several utterances including the rapid repetition of the syllables /pa-ta-ka/, /pa-ka-ta/, /pe-ta-ka/, /pa/, /ta/, /ka/, isolated sentences, a read text, and a monologue. All patients were recorded in ON state, i.e., no more than three hours after their morning medication, and were evaluated by a neurologist expert. Additional information from the participants is shown in Table 3. In addition, Figure 3 shows the distribution of the clinical scores for the patients. We divided the patients in three groups according to their level of the total MDS-UPDRS-III and to the speech item of the MDS-UPDRS-III scores. For the m-FDA score, the subjects are divided in four groups because that scale was applied also to HC subjects (white bars). The division consider the same number of subjects in each group.

Table 3. Demographic information of the participants from this study

	PD patients		HC subjects	
	male	**female**	**male**	**female**
Number of subjects	25	25	25	25
Age ($\mu \pm \sigma$)	61.3±11.4	60.7±7.3	60.5±11.6	61.4±7.0
Range of age	33-81	49-75	31-86	49-76
Duration of the disease ($\mu \pm \sigma$)	8.7±5.8	12.6±11.6	-	-
MDS-UDRS-III ($\mu \pm \sigma$)	37.8±22.1	37.6±14.1	-	-
MDS-UDRS-III speech ($\mu \pm \sigma$)	1.4±0.9	1.3±0.7	-	-
Total m-FDA ($\mu \pm \sigma$)	29.8±8.6	28.2±9.0	7.6±9.2	5.1±7.3

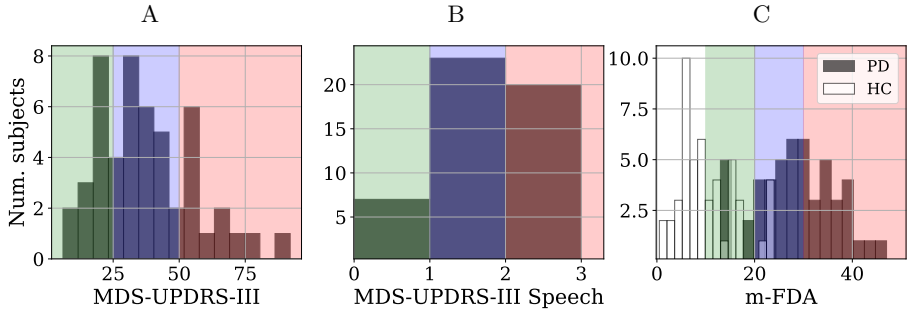

Fig. 3. Distribution of the clinical scores for the participants of this study. Figure includes the distribution of the total MDS-UPDRS-III score (A), the speech item of the MDS-UPDRS-III score (B), and the m-FDA score (C). The scores for the PD patients are grouped into three classes: low (green), intermediate (blue), and severe (red) according to the severity of the disease. The scores for the m-FDA scale also include HC subjects, represented with the white bars.

4 Experiments and results

Three experiments are performed: (1) classification of PD vs. HC subjects, (2) classification of HC vs. PD patients in three stages of the disease divided according to the speech item of the MDS-UPDRS-III score (see Figure 3 B), and (3) classification of HC and PD patients divided into four groups according to the total m-FDA scale (see Figure 3 C). The results are shown in Table 4.

The phonological features provide higher accuracies than those obtained with MFCCs to discriminate between PD patients and HC subjects. On the other hand, note that when we consider the multi-class experiments e.g., the classification of the UPDRS-speech item and the m-FDA scores, the highest accuracies are obtained with the fusion of MFCC and phonological features, which indicate that these two feature sets provide complementary information to assess the speech of PD patients in several stages of the disease. Further experiments with other deep architectures are required to improve the results for multi-class assessment of the patients.

Table 4. Results of the proposed approach to classify PD patients vs. HC subjects, and to assess the speech deficits of the patients following the speech item of the MDS-UPDRS-III and the m-FDA scores. **ACC:** accuracy in the test set, **AUC:** Area under receiving operating characteristic curve for the two-class experiments.

Features	Classification Task	Num. Classes	ACC.	AUC
Phonological	PD vs HC	2	76.0±5.8	0.78
Phonological	UPDRS-speech	3	57.0±4.0	-
Phonological	m-FDA	4	30.8±1.4	-
MFCC	PD vs HC	2	65.0±4.7	0.66
MFCC	UPDRS-speech	3	59.4±6.9	-
MFCC	m-FDA	4	33.7±2.0	-
Phonological+MFCC	PD vs HC	2	64.0±5.6	0.69
Phonological+MFCC	UPDRS-speech	3	59.4±6.9	-
Phonological+MFCC	m-FDA	4	39.5±11.3	-

5 Conclusion

This study considers phonological posterior features and recurrent neural networks based on GRU units to assess speech impairments of PD patients. A total of 15 phonological features are computed based on the sound pattern of English to model several aspects of the speech production system. The phonological posteriors used in this study can be interpreted by medical experts, which may support the evaluation of the speech state of the patients.

The results obtained with phonological features are compared with those obtained with standard acoustic features based on MFCCs. The phonological features are better to model the speech impairments of PD patients than the standard acoustic features, specially to discriminate between PD patients and HC subjects, however, the combination of acoustic features with the phonological posteriors shows to be suitable to assess the speech deficits of the patients in several stages of the disease. Further experiments are required with other deep architectures to assess the neurological state and the dysarthria level of the patients. Additionally, other feature sets based on phonation, articulation, or prosody analyses could be considered.

Acknowledgments

The work reported here was financed by CODI from University of Antioquia by grants Number 2015–7683. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 766287.

References

1. Hornykiewicz, O.: Biochemical aspects of Parkinson's disease. Neurology **51**(2 Suppl 2) (1998) S2–S9

2. Logemann, J.A., Fisher, H.B., Boshes, B., Blonsky, E.R.: Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. Journal of Speech and Hearing Disorders **43**(1) (1978) 47–57
3. C. G. Goetz et al.: Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Movement Disorders **23**(15) (2008) 2129–2170
4. Orozco-Arroyave, J.R., Vásquez-Correa, J.C., et al.: Neurospeech: An open-source software for Parkinson's speech analysis. Digital Signal Processing (In press) (2017)
5. Hlavnicka, J., Cmejla, R., Tykalova, T., Sonka, K., Ruzicka, E., Rusz, J.: Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder. Nature Scientific Reports **7**(12) (2017) 1–13
6. Tu, M., Berisha, V., Liss, J.: Interpretable objective assessment of dysarthric speech based on deep neural networks. In: Proceedings of INTERSPEECH. (2017) 1849–1853
7. Vásquez-Correa, J.C., Orozco-Arroyave, J.R., Nöth, E.: Convolutional neural network to model articulation impairments in patients with Parkinson's disease. In: Proceedings of INTERSPEECH. (2017) 314–318
8. Cernak, M., et al.: Characterisation of voice quality of Parkinsons disease using differential phonological posterior features. Computer Speech & Language **46** (2017) 96–208
9. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). (2014) 1724–1734
10. Cernak, M., Potard, B., Garner, P.N.: Phonological vocoding using artificial neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2015) 4844–4848
11. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553) (2015) 436–444
12. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks **5**(2) (1994) 157–166
13. Irie, K., Tüske, Z., Alkhouli, T., Schlüter, R., Ney, H.: Lstm, gru, highway and a bit of attention: An empirical overview for language modeling in speech recognition. In: Proceedings of INTERSPEECH. (2016) 3519–3523
14. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Deep Learning and Representation Learning Workshop. (2014)
15. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: Advances in neural information processing systems (NIPS). (2012) 2951–2959
16. Orozco-Arroyave, J.R., et al.: New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In: Language Resources and Evaluation Conference, (LREC). (2014) 342–347