INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease

J. C. Vásquez-Correa^{1,2}*, J. R. Orozco-Arroyave*^{1,2}*, E. Nöth*^{2}

^{1}Faculty of Engineering, University of Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia.
^{2}Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

jcamilo.vasquez@udea.edu.co

Abstract

Speech impairments are one of the earliest manifestations in patients with Parkinson's disease. Particularly, articulation deficits related to the capability of the speaker to start/stop the vibration of the vocal folds have been observed in the patients. Those difficulties can be assessed by modeling the transitions between voiced and unvoiced segments from speech. A robust strategy to model the articulatory deficits related to the starting or stopping vibration of the vocal folds is proposed in this study. The transitions between voiced and unvoiced segments are modeled by a convolutional neural network that extracts suitable information from two time–frequency representations: the short time Fourier transform and the continuous wavelet transform. The proposed approach improves the results previously reported in the literature. Accuracies of up to 89% are obtained for the classification of Parkinson's patients vs. healthy speakers. This study is a step towards the robust modeling of the speech impairments in patients with neuro–degenerative disorders.

Index Terms: Parkinson's disease, Articulation, Convolutional neural network, Time–frequency representations, Wavelet transform.

1. Introduction

Parkinson's disease (PD) is a neurological disorder that alters the function of the basal ganglia in the midbrain, producing motor and non–motor deficits in the patients [1]. Speech impairments are an early and prominent manifestation that can contribute primarily to the diagnosis of PD [2]. The main symptoms of the impaired speech of PD patients include reduced loudness, monopitch, monoloudness, hypotonicity, breathy, hoarse voice quality, and imprecise articulation. These symptoms are typically grouped and called *hypokinetic dysarthria* [3].

Several studies in the literature have described the speech impairments of PD patients in terms of phonation, articulation, and prosody [4, 5, 6]. Phonation is related to the capability of the speaker to make the vocal folds vibrate to produce vocal sounds, articulation is related with the modification of the position, stress, and shape of several muscles to produce speech, and prosody reflects variation of loudness, pitch, and timing to produce natural speech. Articulation deficits in PD patients are mainly related to reduced amplitude and velocity of lip, tongue, and jaw movements [7]. Particularly, imprecise consonant articulation was perceptually found as one of the most deviant speech dimensions in PD [8].

In general, articulation impairments of PD patients have been analyzed in several studies both from the medical and engineering perspective. In [5] the authors evaluated possible correlations between vowel articulation, global motor performance, and the stage of the disease. A total of 68 patients and 32 healthy control (HC) speakers are considered. According to the results obtained in several statistical tests, the authors concluded that the vowel articulation index (VAI) is significantly reduced in PD speakers. In [9] six different articulatory deficits in PD were modeled: vowel quality, coordination of laryngeal and supra-laryngeal activity, precision of consonant articulation, tongue movement, occlusion weakening, and speech timing. The authors studied the rapid repetition of the syllables /pa-ta-ka/ pronounced by 24 Czech native speakers, and reported an accuracy of 88% discriminating between PD patients and HC. Articulation impairments have been also analyzed using time–frequency representations (TFR) [10], where three TFR were computed from continuous speech utterances with the aim of detecting changes in the low frequency components of the spectrum that could be associated to the presence of tremor in the speech. The TFR include modulation spectra, the wavelet packet transform, and the Wigner-Ville distribution. The authors extract features related to the energy content and spectral centroids in different frequency bands, and report an accuracy of up to 77% classifying PD patients and HC speakers using several classification strategies.

In [11] it was introduced a method to model difficulties observed in PD patients to start/stop the vibration of vocal folds. The method consists of detecting the transitions from voiced to unvoiced (v-uv), i.e. offset, and from unvoiced to voiced (uv-v), i.e. onset in the speech recording. Then the energy content in frequency bands separated according to the Bark scale is computed. In order to improve the method presented in [11], in the present study the onset and offset are modeled with a more robust strategy that considers both the temporal and frequency domains of the transitions. The onset and offset are modeled using two TFR: the short time Fourier transform (STFT) and the continuous wavelet transform (CWT). The TFRs are used to feed a convolutional neural network (CNN) that learns high–level representations from the low–level raw features from the TFR. The combination of TFRs and CNNs has been previously used in speech recognition and other speech processing tasks [12, 13, 14].

The proposed model is tested in the classification of PD patients vs. HC subjects in three different languages: Spanish, German, and Czech. The results obtained are compared to a baseline computed with the strategy introduced in [11]. According to the results, the proposed approach improves the results relative to previous studies. Accuracies of up to 89% are obtained for the classification of PD patients vs. HC speakers. This study is a step towards the robust modeling of the speech impairments in patients with neuro–degenerative disorders

2. Methods

The proposed method is divided into three stages: (1) the detection of the onset and offset transitions, (2) the computation

http://dx.doi.org/10.21437/Interspeech.2017-1078

of the two TFRs (STFT and CWT), and (3) the feature learning and classification using the CNN. Each stage is described as follows.

2.1. Onset and Offset detection

The offset and onset are segmented according to the presence of the fundamental frequency F_0 using Praat. The borders are detected, and 80 ms of the signal are taken to the left and to the right of each border, forming "chunks" of signals with 160 ms length. Each one of those chunks is modeled using the STFT and the CWT.

2.2. Time–Frequency Representations

2.2.1. Short time Fourier transform

The STFT with 64 frequency bins is computed for each transition. Frames of 16 ms and time-shift of 10 ms are considered, forming and image of 65 × 25 pixels to feed the CNN. Figure 1 shows the difference between the onset for a PD patient (female, 73 years old) and a HC speaker (female, 73 years old). In both cases a female speaker with 73 years old is considered. Note that onset is more spread out for the PD patient than for the HC, where the starting point of the voiced segment is well defined. The figure also indicates breathy voice in the speech of the patient. Note that most of the energy is concentrated at the beginning of the frame for PD and at the end of the frame for HC, which may be caused by compensatory articulatory movements of the patient at the beginning of the vocal fold vibration.

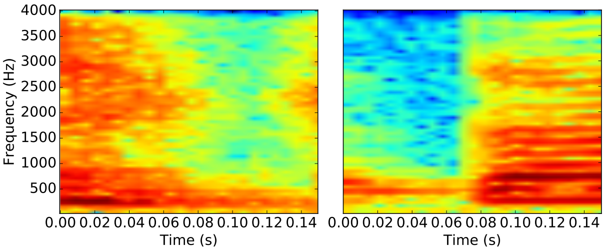

Figure 1: *STFT of the onset for a PD patient (left) and a HC speaker (right) when they pronounce the syllable /ta/*

2.2.2. Continuous wavelet transform

The CWT is introduced as an alternative to represent and decompose non–stationary signals. This TFR allows a time–frequency multi–resolution analysis based on the decomposition of the signal into time–variable length frames. The CWT is computed for each transition using the Morlet wavelet function, which is closely related to the human hearing perception and has been used in other speech processing tasks [15]. The signal is decomposed into 512 scales in steps of 16. As in the STFT, frames of 16ms length with a time–shift of 10ms are considered, forming an image of 34 × 26 pixels. Figure 2 shows the CWT obtained for the same speakers considered in Figure 1. Note the difficulty of the patient to start the vibration of the vocal folds. Note the same effect observed in the STFT: most of the energy is concentrated at the beginning of the utterance of the PD patient which may be caused by compensatory movements.

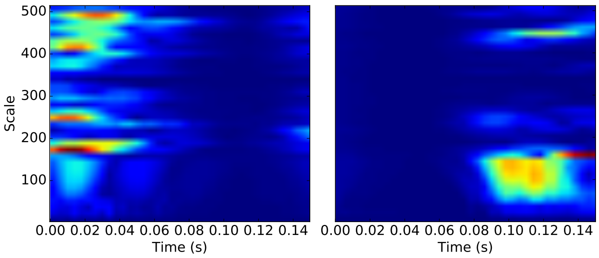

Figure 2: *CWT of the onset for a PD patient (left) and a HC speaker (right) when they pronounce the syllable /ta/*

2.3. Convolutional Neural network

A CNN can be defined as a variant of the standard neural networks. Instead of using fully connected hidden layers the CNN introduces a structure that consists of alternating convolution and pooling layers. The CNN receives as inputs a matrix $\mathrm{V} \in \mathbb{R}^{\mathrm{t}\times\mathrm{f}}$, where t and f are the time and frequency indexes of the TFR. A weight matrix $\mathrm{W} \in \mathbb{R}^{\mathrm{m}\times\mathrm{m}\times\mathrm{d}}$ is convolved with the input matrix, where m is the order of the convolution filter (kernel size), and d is the number of hidden units of the layer i.e., feature maps. The weight matrix transforms the input image into d small local TF patches of size $(\mathrm{t}-\mathrm{m}+1)\times(\mathrm{f}-\mathrm{m}+1)$. After performing the convolution, a max–pooling layer is applied to remove variability in the time–frequency plane that appears due to the speaking style, channel distortion, or other external factors. The pooling layer performs a sub–sampling operation to reduce the time–frequency space. The last layer of a CNN corresponds to a fully connected layer with h hidden units followed by a sigmoid activation function to make the final decision of whether the TFR corresponds to a PD patient or a HC speaker. Figure 3 shows the architecture of the CNN used in this study, which is formed with two convolutional and max–pooling layers followed by a fully connected multi–layer perceptron (MLP).

The CNN is trained using the stochastic gradient descent (SGD) algorithm with a defined batch size, using cross–entropy as the loss function. TensorFlow [16] is used for the implementation of the CNN. Rectifier linear (ReLu) activation functions are used in the convolutional layers, and dropout is included in the training stage with the aim of avoiding over–fitting. Dropout consists of setting to zero the output of each hidden neuron with probability 0.5. The neurons which are "dropped out" in this way do not contribute to the forward pass and do not participate in back-propagation [17]. The meta–parameters used to train the CNN are detailed in Table 1. A 10–fold speaker independent cross–validation strategy is performed to train the CNN.

3. Experimental framework

3.1. Databases

3.1.1. Spanish

The PC-GITA database [18] is used in this study. The data contain speech utterances from 50 PD and 50 HC Colombian native speakers balanced in age and gender. The participants pronounce several speech tasks 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

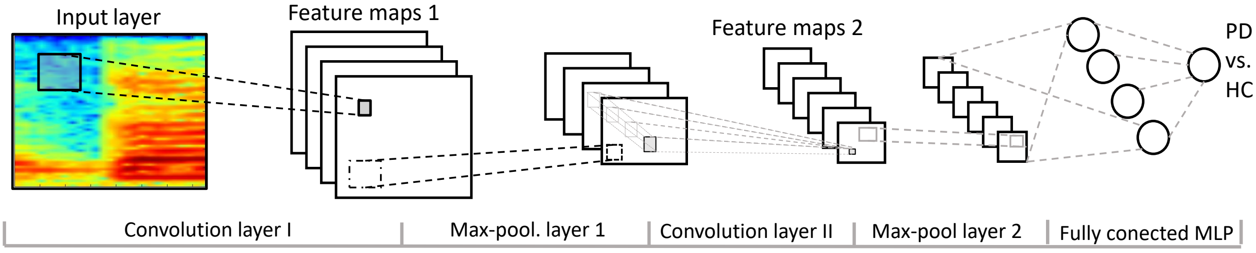

Figure 3: *Architecture of the convolutional neural network implemmented for this study*

Table 1: *Parameters used to train the CNN*

Meta-parameter	Values
Batch size	64
kernel size conv. layer I	$\{4, 6, 8\}$
kernel size conv. layer II	$\{5, 7, 9\}$
max–pool size layer I	2
max–pool size layer II	2
depth of convolutional layers	$\{16, 32, 64\}$
hidden units in fully connected MLP	$\{16, 32, 64, 256\}$
training rate	0.05
number of iterations	40000

expert.

3.1.2. German

The German data contain recordings from 88 PD patients and 88 HC subjects. The speakers perform several speech tasks, including the repetition of /pa-ta-ka/, isolated sentences, a read text, and a monologue [6].

3.1.3. Czech

The Czech data are formed with recordings from 20 PD patients and 15 HC subjects. The patients were newly diagnosed with PD, and none of them had been medicated before or during the recording session. The speech tasks performed by the speakers include the repetition of /pa-ta-ka/, a read text and a monologue [6].

3.2. Experiments

The STFT and the CWT are used to train CNNs individually with the aim of discriminate whether the transitions were uttered by PD patients or HC speakers. The CNN is trained using the onset and offset transitions obtained from read texts, monologues, isolated sentences, and the rapid repetitions of syllables /pa-ta-ka/, /pa-ka-ta/, /pe-ta-ka/, /pa/, /ta/, and /ka/. The classification is performed in the three languages, i.e., Spanish, German, and Czech. Cross–language experiments are also performed, i.e., train with utterances from one language, and test with the utterances from the other two languages.

The results with the proposed approach are compared to a baseline computed with the method introduced in [11], where onset and offset are modeled with the energy content distributed according to the Bark scale, and using a radial basis support vector machine to perform the classification.

4. Results

4.1. Classification in the same language

Table 2 contains the results for the classification of PD patients vs. HC speakers in the three languages separately, i.e., training and testing on the same language. Note that the classification with the proposed method highly improves the results obtained with the baseline in Spanish and Czech languages. In German the results obtained with the baseline are slightly better than those obtained with the proposed approach for the separately classification of onset and offset; however the result with the proposed approach improves the baseline when onset and offset are combined. Onset and offset are combined in order to increase the amount of information and to analyze whether the information from the different transitions could be complementary. Note that such a combination slightly improves the results relative to the separate analyses in all scenarios. This fact indicates that the features extracted from onset and offset could provide complementary information that can be merged together to analyze the articulation impairments of PD patients.

Table 2: *Accuracies (%) for classification of PD patients vs. HC speakers in three different languages*

TFR	onset	offset	onset+offset
Spanish			
STFT	85.3	81.6	85.9
CWT	84.2	81.8	85.2
Baseline	69.3	69.6	71.6
German			
STFT	70.3	68.0	75.0
CWT	68.0	66.9	70.5
Baseline	72.7	70.9	74.0
Czech			
STFT	77.9	80.4	84.4
CWT	89.2	87.7	89.4
Baseline	75.3	74.4	78.8

Figure 4 displays the output of the CNN after the last max-pooling layer for the same speakers from Figure 1: an onset for a PD patient (left) and for a HC speaker (right). Note that the image for the HC speaker shows more energy content in the voiced region than the observed in the image from the patient, where the transition from unvoiced to voiced is not observed, which was also observed in the STFT representation.

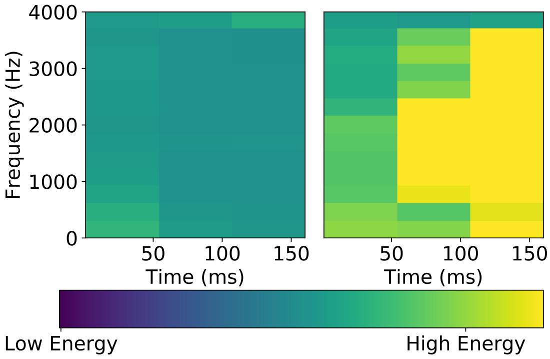

Figure 4: *Output of the CNN after the last max–pool layer for a PD patient (left) and a HC speaker (right) when they pronounce the syllable /ta/*

Table 3 shows the accuracies obtained with the transitions from read text, monologue, and the rapid repetition of /pa-ta-ka/ in Spanish, German, and Czech. The reported results are obtained with the TFR that provides the highest accuracy in each language, i.e., STFT for Spanish and German, and CWT for Czech. Note that there is no high differences among the results obtained for each task, which makes the approach proposed in this study independent from the speech task uttered by the speaker.

Table 3: *Individual accuracies (%) for monologues, read texts, and the repetition of /pa-ta-ka/ in the three different languages*

Speech Task	Spanish	German	Czech
read text	85.0	70.3	88.5
monologue	85.6	70.3	89.1
/pa-ta-ka/	85.4	70.7	89.2

4.2. Cross-language classification

The results when the languages used for train and test are different are shown in Table 4 with the aim to analyze the language independence of the proposed approach. In general, none of the proposed methods or the baseline are able to classify correctly the PD patients and the HC speakers. More experiments should be performed to obtain a robust strategy able to model the articulation impairments of PD patients independently of the language spoken by the subjects. The incremental insertion of speakers from the target language could be a good strategy as in [6].

5. Conclusions

A robust strategy to model the articulation impairments of Parkinson's patients is proposed in this study. The method focuses on analyzing the time–frequency components of the speech signal in the transitions from voiced to unvoiced and from unvoiced to voiced segments. The analysis is performed with the aim of evaluating the capabilities of the speaker to start/stop the vibration of the vocal folds. A convolutional neural network is used to extract features from the time–frequency representations and to make the final decision of whether the speech segments are from patients or healthy subjects.

Table 4: *Accuracies (%) for classification of PD patients vs. HC speakers in three different languages when the train and the target languages are different*

Test Lang.	TFR	onset	offset	onset+offset
Train with Spanish				
German	STFT	51.7	50.2	54.7
German	CWT	50.8	50.3	50.6
German	Baseline	53.7	55.0	54.1
Czech	STFT	53.0	55.0	51.7
Czech	CWT	55.2	55.4	57.9
Czech	Baseline	60.3	57.4	60.4
Train with German				
Spanish	STFT	58.0	55.7	55.8
Spanish	CWT	51.5	51.3	50.8
Spanish	Baseline	53.5	53.5	53.6
Czech	STFT	53.0	52.4	53.0
Czech	CWT	53.1	51.7	52.5
Czech	Baseline	50.9	51.7	52.6
train with Czech				
Spanish	STFT	55.1	51.2	50.5
Spanish	CWT	53.8	56.3	56.7
Spanish	Baseline	53.4	51.6	52.4
German	STFT	54.0	51.8	54.0
German	CWT	50.8	50.2	50.6
German	Baseline	51.2	51.0	50.7

The proposed method is able to discriminate between patients and healthy subjects when the language used for train and test is the same. The proposed strategy improves the results relative to a baseline where the articulation impairments of the patients in the transitions between voiced and unvoiced segments are evaluated.

Additional experiments and methods need to be proposed to improve the results when the language used for train and test is different. The main aim would be to find a language–independent model to discriminate between Parkinson's patients and healthy speakers. One possible alternative would be to incrementally add speakers of the target language into the train set as it was presented in [6]; however, this strategy assums to have access to speakers of the target language, which is not always possible. Models based on recurrent neural networks (RNN) which models the time dependence between consecutive voiced–unvoiced transitions should be also considered to assess co-articulation. This approach could contribute to understand difficulties of PD speakers to produce stops, plosives and other consonant in continuous speech.

6. Acknowledgements

The work reported here was carried out during the 2016 Jelinek Memorial Summer Workshop on Speech and Language Technologies, which was supported by Johns Hopkins University via DARPA LORELEI Contract No HR0011-15-2-0027, and gifts from Microsoft, Amazon, Google, and Facebook. Thanks also to CODI from University of Antioquia by the grant Number 2015–7683, and to COLCIENCIAS by the grant Number 111556933858.

7. References

[1] O. Hornykiewicz, “Biochemical aspects of Parkinson’s disease,” *Neurology*, vol. 51, no. 2 Suppl 2, pp. S2–S9, 1998.

[2] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, “Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinsons disease,” *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[3] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, “Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients,” *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[4] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, “Objective automatic assessment of rehabilitative speech treatment in Parkinson’s disease,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 181–190, 2014.

[5] S. Skodda, W. Visser, and U. Schlegel, “Vowel articulation in parkinson’s disease,” *Journal of Voice*, vol. 25, no. 4, pp. 467–472, 2011.

[6] 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,” *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 481–500, 2016.

[7] H. Ackermann and W. Ziegler, “Articulatory deficits in Parkinsonian dysarthria: an acoustic analysis.” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 54, no. 12, pp. 1093–1098, 1991.

[8] F. L. Darley, A. E. Aronson, and J. R. Brown, “Differential diagnostic patterns of dysarthria,” *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 246–269, 1969.

[9] M. Novotnỳ, J. Rusz, R. Čmejla, and E. Růžička, “Automatic evaluation of articulatory disorders in Parkinson’s disease,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 9, pp. 1366–1378, 2014.

[10] T. Villa-Cañas, J. D. Arias-Londoño, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, and E. Nöth, “Low-frequency components analysis in running speech for the automatic detection of Parkinson’s disease.” in *16th International Conference of the Speech and Communication Association (INTERSPEECH)*, 2015, pp. 100–104.

[11] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, “Towards an automatic monitoring of the neurological state of the Parkinson’s patients from speech,” in *41st International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2016.

[12] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[13] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. Mohamed, G. Dahl, and B. Ramabhadran, “Deep convolutional neural networks for large-scale speech tasks,” *Neural Networks*, vol. 64, pp. 39–48, 2015.

[14] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[15] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, and E. Nöth, “Wavelet–based time–frequency representations for automatic recognition of emotions from speech,” in *12 ITG Symposium on Speech Communication*, 2016, pp. 1–5.

[16] A. Martín and et. al., “Tensorflow: A system for large-scale machine learning,” in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283.

[17] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, “New Spanish speech corpus database for the analysis of people suffering from Parkinson’s disease,” in *Language Resources and Evaluation Conference, (LREC)*, 2014, pp. 342–347.