Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification

Marek Wodzinski[1], Andrzej Skalski[1], Daria Hemmerling[1], Juan Rafael Orozco-Arroyave[2], Elmar Nöth[3]

Abstract— **This study presents an approach to Parkinson's disease detection using vowels with sustained phonation and a ResNet architecture dedicated originally to image classification. We calculated spectrum of the audio recordings and used them as an image input to the ResNet architecture pre-trained using the ImageNet and SVD databases. To prevent overfitting the dataset was strongly augmented in the time domain. The Parkinson's dataset (from PC-GITA database) consists of 100 patients (50 were healthy / 50 were diagnosed with Parkinson's disease). Each patient was recorded 3 times. The obtained accuracy on the validation set is above 90% which is comparable to the current state-of-the-art methods. The results are promising because it turned out that features learned on natural images are able to transfer the knowledge to artificial images representing the spectrogram of the voice signal. What is more, we showed that it is possible to perform a successful detection of Parkinson's disease using only frequency-based features. A spectrogram enables visual representation of frequencies spectrum of a signal. It allows to follow the frequencies changes of a signal in time.**

I. Introduction

Speech disorders affect up to 89% of all patients with Parkinson's disease (PD, Parkinson's disease) [1]. James Parkinson described in his publication [2] the formation of speech disorders due to PD. Speech impairments are mainly caused by: laryngeal function deficit, impaired mimicry, reduced lung life capacity and decreased speech force. Such changes lead to appearance of numerous deficits in voice and speech such as: reduction of loudness, lowering the tone of the voice, limited modulation (monotonous speech), difficulties with changes in loudness, voltage reduction of vocal folds, rough and hoarse tone, as well as improper articulation (speech becomes indistinct) and change in speech pace [3], [4].

Diagnosis of Parkinson's disease is done mostly based on movement analysis and is extremely difficult in the disease early stages. PD develops over a long period of time and is clinically silent for many years. Subtle symptoms are often interpreted as a symptom of aging or misdiagnosed as another neurological disorder. Movement slowdown can occur in all forms of Parkinsonism, resting tremor can occur in drug-induced Parkinsonism and postural instability may indicate the development of atypical Parkinsonism. For this reason, motor symptoms are well suited to assess the progression of the disease over time.

Related work: Acoustic analysis of the speech of Parkinson's disease is investigated by many researchers and physicians and is a promising non-invasive, objective tool for PD detection. The acoustic analysis may be calculated by various features applied to classification task: to distinguish between voice samples representing healthy control (HC) or PD cases. Those studies include application of different machine learning methods in diagnostic system such as Deep Neural Networks, Support Vector Machines, Naive Bayes, Decision Trees, Rotation Forest and Regression [5]–[9]. The authors of [6] apply a classificator such as Parallel Distributed Neural Network. The detection accuracy of Parkinson's disease was up to 90%. The paper [7] presents different classificators to detect PD such as: Neural Networks, Decision Trees, Regression and DMneural algorithm. The best classification rate was achieved for Neural Network which is 92.9%. The study presented in [8] describes the usage of correlation-based feature selection to reduce the feature dimensionality (from 23 to 11) and its influence on Rotation Forest classifier. As the results, the authors show the increase of 2.7% using 11 features. The authors of [9] created the system for classification and the severity assessment of PD. They implement Support Vector Machines with Neural Networks for the classification. The results show 97.64% detection accuracy. The size of database used in the experiments is the major problem to determine proper classification performance. Usually, the authors implement very small amount of data, which is less than 60 voice recordings with various success. The paper [10] presents 98,6% classification accuracy using vowel /a/ from 33 patients suffering from PD and 10 healthy people. Authors of [11] obtained 92% of detection accuracy based on vowel /e/ from 20 people with PD and 20 healthy people. 71,6% classification accuracy was achieved and presented in the paper [12] based on vowel /a/ using 50 people diagnosed with PD and respect number of healthy persons. The corpus set up from 3 languages: 170 German speakers, 100 Spanish speakers and 35 Czech speakers, was applied for the classification process in [13]. The authors used the

*This work was partially funded by COLCIENCIAS project 111556933858 and NAWA-PROM, grant no. PPI/PRO/2018/1/00026/U/001.

*All of the participants signed an informed consent previously approved by the Ethical Committee of the Noel Clinic in Medellin, Colombia.

[1]AGH University of Science and Technology, Department of Measurement and Electronics, Al. Mickiewicza 30, 30-059 Krakow, Poland {wodzinski,skalski,hemmer}@agh.edu.pl

[2]Universidad de Antioquia UdeA, Department of Electronics and Telecommunications Engineering, Calle 70 No. 52-21 Medellin, Colombia rafael.orozco@udea.edu.co

[3]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg Erlangen, Germany elmar.noeth@fau.de

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

text read and calculated energy content in the transitions between voiced and unvoiced what segments what enabled achieving detection accuracy being between 91% and 98% depending on the languages. Up to our knowledge, deep learning architectures were not investigated in the context of automatic feature learning for Parkinson's disease diagnosis. The research about automatic features extraction [14], [15] were presented for other speech disorders which are much easily distinguishable from healthy person voice.

Contribution: In this preliminary work we propose usage of a modified ResNet [16] architecture to detect Parkinson's disease based on audio recordings of sustained vowels. We convert the voice recordings to an image-based representation and use a pre-trained network to perform the classification. The results are surprising because the whole classification is based only on frequency features, yet it is able to distinguish between the classes. The work is a solid foundation for further investigation into deep learning architectures with automatic or semi-automatic feature extraction to diagnose Parkinson's disease using the voice recordings.

II. Methods

The audio recordings are example of sequential data. A natural approach to classify such data using neural network is use of the long short-term memory (LSTM) layers or other recurrent neural architectures. Such an approach was presented in [15] to classify speech disorders using raw audio signals. The problem with this approach is related to the database size. The state-of-the-art deep learning methods require huge databases which do not exist for the problem being discussed. The openly available SVD [17] or MEEI [18] databases are not large enough. In this work we try to address this problem using network dedicated to the computer vision classification, for which huge databases for pre-training are openly available.

In this work we employed a slightly modified ResNet architecture with 18 layers (four convolutional layers per single bottleneck + the input and output pooling layers) in the autoencoder. However, we replaced the last linear layer with a three layer dense network with the PReLU as the activation function and dropout with probability equal to 0.5 after the first two layers. The reason was motivated by really low dataset size in a deep learning context, and a great need for a successful regularization. We experimentally verified that adding a relatively shallow but strongly regularized network after the autoencoder improved the results. What is important, the ResNet architecture is dedicated to the image classification problem, not for pathology detection using speech signals. However, it turned out that some features of the audio signal can be represented as an image.

A crucial factor for a successful training of deep learning is a proper data augmentation and pre-processing. In the case discussed it was not correct to use the traditionally applied affine transform, horizontal or vertical flipping, contrast changes or corners cropping. These transformations are meaningless for the spectrogram because it does not preserve properties which are true for natural images. Therefore, we decided to augment the data before the spectrogram calculation. Firstly, the input was randomly rolled to prohibit the network from fitting to the time localization of a given frequency component. It was possible because the signal represented a sustained vowel. Secondly, a band-pass filter was applied with a randomly chosen order and low/high pass frequencies. These two simple techniques were the crucial factors to prevent the overfitting. The network was unable to directly fit to the training data, while without them the network overfitted with 100% accuracy to the training set after just few epochs, without any generalization ability. The augmentation was applied only during training.

Last thing to mention, but perhaps the most important, is the applied transfer learning. The main reason why we decided to use ResNet is the availability of ImageNet database [19] which allowed us to use a pre-trained network. Surprisingly, it turned out that features learned from natural images improved the results significantly. In fact, without the transfer learning from ImageNet, due to really low dataset size, the network was not able to learn anything. After pre-training on the ImageNet and changing the last layer to the dense network, the whole network was pre-trained again using the SVD database. Only after that, the network was finally trained on the Parkinson's disease database giving stable and meaningful results. The whole system work-flow is presented in Figure 1.

III. Results

A. Participants and data collection

The PC-GITA database [22] was used to create and evaluate the models proposed in this paper. The corpus includes recordings of 50 PD patients and 50 HC subjects, all of them are Colombian Spanish speakers. Two recording tasks were considered in this study, [22] the participants were asked to produced the sustained vowel /ah/ as long as possible in one breath. All of the participants signed an informed consent previously approved by the Ethical Committee of the Noel Clinic in Medellin, Colombia. The speech signals were recorded in a soundproof booth using a Shure SM63L microphone and a professional audio card. The audios were recorded at 44,1kHz with a resolution of 16 bits. Each group of participants contain 25 male and 25 female speakers. The corpus is also balanced in age (t-test for independent samples, p=0,77). All of the patients were diagnosed by an expert neurologist. In this study we used only the sustained vowels.

B. Experimental Setup

The set with 100 speakers was divided into 10 folds, with the 90/10% (training/validation) proportion. Each patient was in the validation set exactly once but with three separate recordings. As a result, the total support is equal to 300. We decided not to use a separate test set due to a low database size and the fact that the network architecture was not created from scratch since we employed a predefined

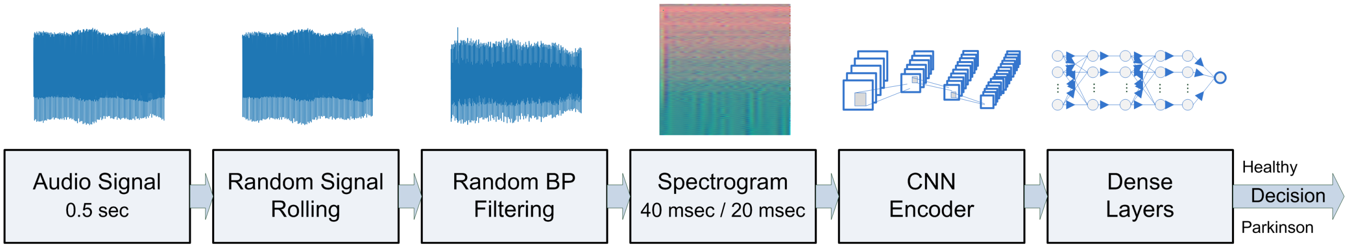

Fig. 1. The work-flow of the proposed deep learning based Parkinson's disease classification system.

ResNet encoder. Therefore the model parameters were not really tuned using the validation set. As a result the validation set can be considered as test set. Based on experimental results we decided to use only the sustained vowel /a/. We used the SGD with learning rate equal to 0.0005 for all layers except the dense network for which the learning rate was set to 0.008. The momentum was equal to 0.95, the mini-batch size was equal to 64 for the SVD database and 16 for the Parkinson's database. The network was trained for 200 epochs for each fold using the Parkinson's database. The cost function used was the cross-entropy loss. The pre-training using the SVD database was performed for 2000 epochs once. The whole system was implemented using PyTorch [20].

C. Classification Results

The accumulated confusion matrix, for all 10 folds, is shown in Table I. The Table II presents the precision, recall, F1-score globally and for each class separately. We show results only for the vowel /a/. Results for other vowels turned out to be worse, as well as using all vowels at once.

TABLE I
ACCUMULATED CONFUSION MATRIX FOR CLASSIFICATION RESULTS FOR THE 10-FOLD VALIDATION SET

		Predicted	
		HC	PD
Actual	HC	136	14
	PD	11	139

TABLE II
CLASSIFICATION SUMMARY FOR THE ACCUMULATED CONFUSION MATRIX CALCULATED FOR THE VALIDATION SET

	Precision	Recall	F1-score	Support
HC	0.93	0.91	0.92	150
PD	0.91	0.93	0.92	150
Overall	0.92	0.92	0.92	300
Accuracy:		**0.917**		

The proposed method was tested using 10-fold validation. Based on the results presented in table I and table II it is seen that the precision calculated for healthy control group is higher than the precision for people diagnosed with Parkinson's. The opposite trend is visible or recall - it is higher for PD group than healthy control patients. It means that the system predicted more PD patients as being healthy, rather than the healthy. Out of 10-fold validation set, the system classified properly 275 out of 300 voice recordings. The overall classification accuracy is 91.7%.

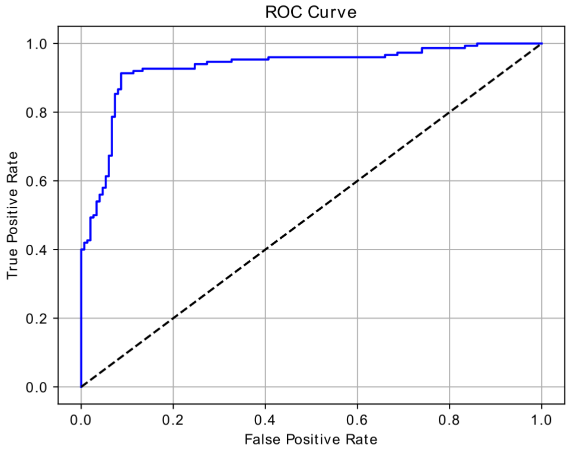

Fig. 2. Figure presenting the ROC. Positive class is represented by the Parkinson's recordings. The AUC=0.93.

IV. Discussion and Conclusion

The results show accuracy for the validation set above 90%, which is comparable to the state-of-the-art. The results are even better when the reader takes into account that only the frequency-based features from spectrograms were used to classify the voice recordings. Another advantage of presented approach is the usage of the sustained vowel /a/. Pronouncing the vowel is an easy task to be performed by the patient, it does not require long instructions and any previous exercises. This task is also fast, it takes only a few seconds. The proposed methodology might help physicians in the diagnosis process of Parkinson's disease.

In the future work we plan to combine the presented approach with deep, sequential networks using the LSTM or Attention layers [21] which can both take advantage of the data sequential nature and autoencoder benefits. We hope that using more features than the localization of frequencies in time will improve the results significantly. Extending the vocal signal description based on acoustic features, enables to make the system more specific to detect more voice impairments caused by Parkinson's disease. The usage of bigger data set, which includes more patient's recordings would be also valuable for teaching the algorithm.

To conclude, the paper describes the usage of modified ResNet algorithm to detect Parkinson's disease based on audio recordings of sustained vowels /a/. The data set of voice recordings to test the algorithm included in total 100 patients, each recorded 3 times. We convert the voice recordings to an image-based representation describing only frequency features and use a pre-trained network to perform the classification. The algorithm is able to classify at the accuracy above 90% two classes based on voice signal: healthy control group and group of people diagnosed with Parkinson's disease. The work is a solid foundation for further investigation into deep learning architectures with automatic or semi-automatic feature extraction to diagnose Parkinson's disease using the voice recordings.

References

[1] M. Trail, C. Fox, L.O. Ramig, S. Sapir, J Howard, E.C. Lai, Speech treatment for Parkinson's disease. NeuroRehabilitation, vol. 20, no.3, pp.205-221, 2005.

[2] J. Parkinson, An essay on the shaking palsy, The Journal of Neuropsychiatry and Clinical Neurosciences, vol. 14.2, pp. 223?236, 2002.

[3] L.O. Ramig, C. Fox, S. Sapir, Speech disorders in Parkinson's disease and the effects of pharmacological, surgical and speech treatment with emphasis on Lee Silverman voice treatment (LSVT). Handbook of clinical neurology, vol. 83, pp. 385-399, 2007.

[4] J.A. Logemann, H.B. Fisher, B. Boshes, 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.

[5] A. Bourouhou, A. Jilbab, C. Nacir, A. Hammouch, Comparison of classification methods to detect the Parkinson disease. In Electrical and Information Technologies (ICEIT), 2016 International Conference on, pp. 421-424, IEEE, 2016..

[6] M. Can, Neural networks to diagnose the Parkinson?s disease. Southeast Europe Journal of Soft Computing, vol. 2, no. 1, 2013.

[7] Parkinson's Disease-symptoms, Stages of Life Expectancy, Pathology, Lecturio, https://www.lecturio.com/magazine/parkinsons-disease/

[8] A. Ozcift, A. Gulten, Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Computer methods and programs in biomedicine, vol. 104, no. 3, 443-451, 2011.

[9] I. Mandal, N. Sairam, Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system. International journal of medical informatics, vol. 82, no. 5, pp. 359-, 2013.

[10] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE transactions on biomedical engineering, vol. 59, no. 5, pp. 1264-1271, 2012

[11] E.A. Belalcazar-Bolaos, J.R. Orozco-Arroyave, J.F. Vargas-Bonilla, J.D. Arias-Londoo, C.G.Castellanos-Domnguez, E. Nth, New cues in low-frequency of speech for automatic detection of Parkinson?s disease. In International Work-Conference on the Interplay Between Natural and Artificial Computation, pp. 283-292, Springer, Berlin, Heidelberg, 2013

[12] T. Villa-Caas, J.R. Orozco-Arroyave, J.F. RVargas-Bonilla, J.D. Arias-Londoo, Modulation spectra for automatic detection of Parkinson's disease. In Image, Signal Processing and Artificial Vision (STSIVA), 2014 XIX Symposium on, pp. 1-5, IEEE, 2014

[13] J.R. Orozco-Arroyave, F. Hnig, J.D. Arias-Londoo, J.F. Vargas-Bonilla, S. Skodda, J. Rusz, E. Nth, Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease, In Sixteenth Annual Conference of the International Speech Communication Association, 2015

[14] B. Woldert-Jokisz, "Saarbruecken voice database", 2007.

[15] Massachusetts Eye and Ear Infirmary, Elemetrics Disordered Voice Database (Version 1.03), Voice Speech Lab, Boston, MA, USA, 1994. Available: http://www.kayelemetrics.com/

[16] K. He *et al.*, Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp.770-778, 2016

[17] M. Alhussein and G. Muhammad, Voice Pathology Detection Using Deep Learning on Mobile Healthcare Framework. IEEE Access, vol. 6, pp. 41034-41041, 2018

[18] P. Harar *et al.*, Voice Pathology Detection Using Deep Learning: a Preliminary Study. 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), 2017

[19] O. Russakovsky *et al.*, ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, pp.211-252, 2015

[20] A. Paszke *et al.*, Automatic differentiation in PyTorch. NIPS-W, 2017

[21] V. Mnih *et al.*, Recurrent Models of Visual Attention. Advances in Neural Information Processing Systems 27 (NIPS-27), 2014

[22] J.R. Orozco-Arroyave, J.D. Arias-Londoo, J.F. Vargas-Bonilla, M.C. Gonzalez-Rtiva, E. Nth, New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. Language resources and evaluation conference (LREC), pp. 342?347, 2014

[23] C.G. Goetz, B.C. Tilley, S.R. Shaftman, G.T. Stebbins, S. Fahn, P. Martinez-Martin, N. LaPelle, Movement Disorder Society-sponsored revision of the Unified Parkinson?s Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Movement Disorders, vol. 23, no. 15, pp. 2129?2170, 2008