DIAGNOSTYKA, 2024, Vol. 25, No. 1

e-ISSN 2449-5220
DOI: 10.29354/diag/176931

PARKINSON'S DISEASE DIAGNOSTICS USING AI AND NATURAL LANGUAGE KNOWLEDGE TRANSFER

Maurycy CHRONOWSKI [1], **Maciej KŁACZYŃSKI** [1,*], **Małgorzata DEC-ĆWIEK** [2], **Karolina PORĘBSKA** [2]

[1] AGH University of Science and Technology, Adama Mickiewicza 30, 30-059 Kraków, Poland,
[2] Jagiellonian University, Collegium Medicum, Jakubowskiego 2, 30-688, Kraków, Poland
[*] Corresponding author, e-mail: maciej.klaczynski@agh.edu.pl

Abstract

With global life expectancy rising every year, ageing-associated diseases are becoming an increasingly important problem. Very often, successful treatment relies on early diagnosis. In this work, the issue of Parkinson's disease (PD) diagnostics is tackled. It is particularly important, as there are no certain antemortem methods of diagnosing PD - meaning that the presence of the disease can only be confirmed after the patient's death. In our work, we propose a non-invasive approach for classification of raw speech recordings for PD recognition using deep learning models. The core of the method is an audio classifier using knowledge transfer from a pretrained natural language model, namely wav2vec 2.0. The model was tested on a group of 38 PD patients and 10 healthy persons above the age of 50. A dataset of speech recordings acquired using a smartphone recorder was constructed and the recordings were labelled as PD/non-PD with the severity of the disease additionally rated using Hoehn-Yahr scale. We then benchmarked the classification performance against baseline methods. Additionally, we show an assessment of human-level performance with neurology professionals.

Keywords: Parkinson's disease, digital diagnostics, artificial intelligence, speech processing

List of Symbols/Acronyms

PD – Parkinson's disease;
HP – Healthy population;
SNR – signal to noise ratio;
GRU – Gated Recurrent Units;
FFT – Fast Fourier Transform;
UPDRS – Unified Parkinson's Disease Rating Scale;

1. INTRODUCTION

Parkinson's disease (PD) is a progressive disorder of the nervous system that affects parts of the brain responsible for the motor functions. It is estimated that in industrialised societies PD affects about 1% of the population above the age of 60 [15]. Despite its commonness, there is still no antemortem test for PD. Therefore, the diagnosis relies on patient's history and physical examination. Novel approaches are examined in works such as [3, 5, 10 14].

Previous findings, including [1,8,12], have shown that Parkinson's disease can be accurately diagnosed using speech recordings and machine learning techniques. In authors' earlier research covering the presented PD dataset [4] the results indicated a significant signal in speech recordings acquired using a smartphone. In this work, we approach this topic using deep learning audio models. We propose an architecture based on wav2vec 2.0 [13] and we test it in a transfer learning setup. Ultimately, we discuss the possibility of implementing our approach as a remote diagnostics tool and we present a human-level performance assessment consulted with the medical experts in the neurology domain. Our goal was to determine if an audio model that was trained on a large-scale natural language dataset can be transferred and fine-tuned to a downstream task of medical diagnostics. Medical tasks usually suffer from insufficient amount of labelled training data, therefore it would be much beneficial to observe such knowledge transfer.

2. RELATED WORKS

2.1. Wav2vec2.0

As a backbone architecture for our experiments, we use pretrained parts of the wav2vec 2.0 model. It is a raw audio speech recognition transformer model published in [13]. The model is pretrained in an unsupervised manner and was shown to deliver state-of-the-art performance in speech recognition tasks using very limited fine-tuning. In this work, we utilise the pretrained convolutional layers of wav2vec 2.0.

Received 2023-10-04; Accepted 2023-12-13; Available online 2023-12-19

© 2023 by the Authors. Licensee Polish Society of Technical Diagnostics (Warsaw. Poland). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

2.2. Explainability in AI

Explainable AI (XAI) plays an important role in medical applications of artificial intelligence. It was shown in [11] that black-box models with wrong explanations encourage distrust in deep learning models, despite their good overall performance. It is therefore important to design models in a way that their predictions can be explained and understood by domain experts, who might not be familiar with machine learning at all. In this work, we discuss possible explanations of the audio models and we present a survey among neurology experts, aiming to assess the human-level performance of speech-based PD diagnostics.

3. PROPOSED METHOD

3.1. Data acquisition

The data was acquired according to the previous research presented in [4,7,9]. The dataset consisted of phonetic test recordings gathered using a mid-range Android smartphone. PD patients were labelled with Hoehn-Yahr ratings by the neurologists at the clinic and the clinical hospital. Healthy persons were recruited from participants above the age of 50, as majority of the PD patients are among the elderly. This helped to mitigate the potential age-related bias. The patients were asked to read out loud a set of vowels (including sustained phonation), syllables, and sentences in Polish language:

- vowels \a, \e, \i, \u pronounced normally (3x);
- sustained phonation of vowels \a, \e, \i, \u (3x);
- words {ala, as, ula, ela, igła} (3x);
- sentences (each 3x):
  - Dziś jest ładna pogoda.
  - Jacek mył kota.
  - Lola lubi bal.
  - Rysiek narysował bar.
  - Marysia namalowała dym.

Full recordings were later manually segmented into audio samples containing fragments of speech described above. The total length of the segmented speech samples was approximately 38 minutes in 2141.wav files, giving on average 43 recordings per subject. Exact numbers vary between patients due to the manual quality check process which ruled out incomprehensible and noisy samples.

3.2. Preprocessing and data augmentation

Before entering the pipeline, two-channel smartphone recordings were subtracted from each other for noise cancellation, as described in previous work [4]. The recordings were also peak-normalised to common gain. Taking into account the relatively small number of available audio samples in the dataset (2141) and a need for broad domain generalisation stemming from the usage of a smartphone recorder, it was necessary to strongly augment the dataset. So-called "audiomentations" [6] were used, including: addition of random background noise, addition of random coloured noise, random shift in time domain, random polarity inversion (Figure 1). The augmentations were prepared by addition of background noise. Two noise recordings were used to be randomly sampled into the training set:

- Recording of a busy street with people talking unintelligibly and objects rattling. Duration: 2:21 minutes.
- Recording of street traffic with cars passing by at different speeds. Duration: 2:00 minutes.

Random fragments of background noise were sampled at every iteration and added to the training samples.

The augmentations were prepared by addition of coloured noise. Parameters drawn randomly from:

- signal-to-noise ratio (SNR) [dB] in range [3,30]
- f_{decay} in range [-2, 2]

The augmentations were prepared by time shift of audio signals. Temporal shift was applied in range of ±10% difference without rollover.

The augmentations were prepared by polarity inversion of audio signals too. Applied to the whole training sample. Each of the augmentations was applied with 50% probability, drawn at every iteration for every augmentation separately. The samples were randomly augmented during each iteration of the training and were turned off during testing.

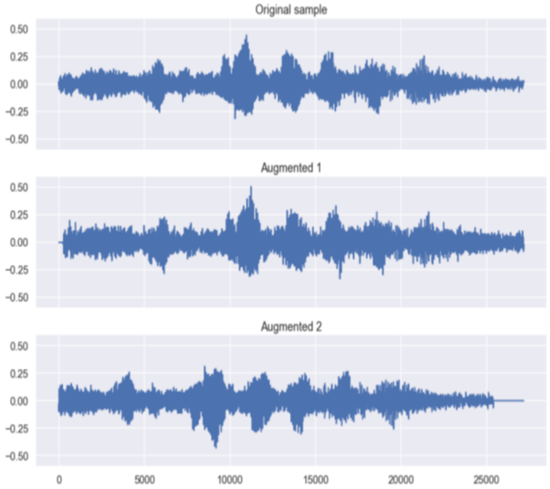


Fig. 1. Visualisation of the signal waveform before (first row) and after augmentations (bottom two). Both of the augmented signals are still clearly intelligible to the human ear.

3.3. Model architecture

The model architecture was designed in a sequence-to-one manner (Figure 2). The input to the model was expected to be a single-channel raw audio waveform that was then internally processed into a vector representation and classified into a class label. Wav2vec 2.0 model is by design a sequence-to-sequence transformer, therefore, sequence aggregation had to be performed after the representation was obtained.

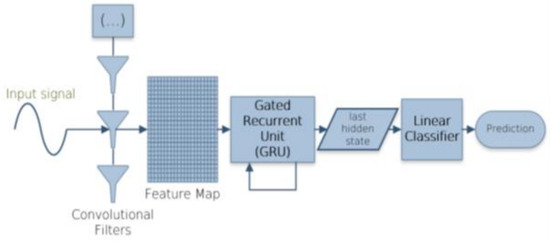


Fig. 2. Proposed model architecture. The convolutional layer was taken from a pretrained wav2vec 2.0 model.

Among tested configurations, the best-performing one was a GRU that was using a convolutional feature map from wav2vec as the input. We tested also a full transformer setup, but it failed to converge in every experimental run. Training logs from both of described approaches are shown in Figure 3 and Figure 4. The last hidden state of the GRU layer was passed on to a linear classifier that generated per-sample predictions.

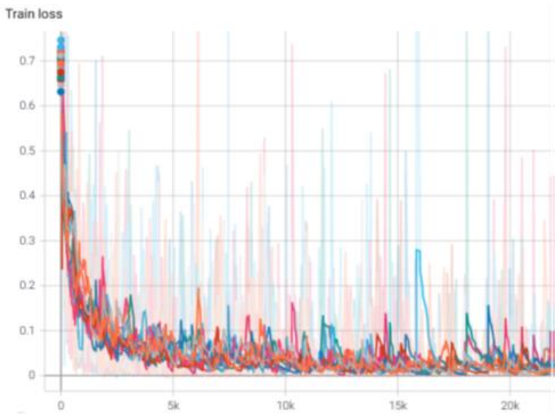


Fig. 3. Training loss of several runs using the simplified model shown in Figure 2.

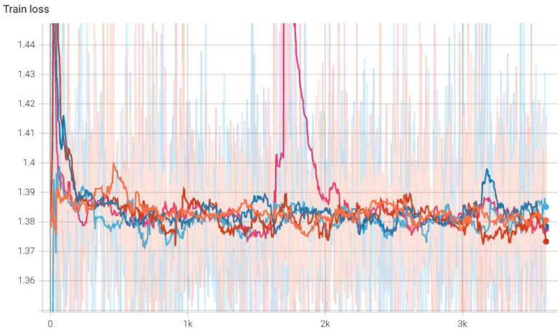


Fig. 4. Training loss of several runs using the full wav2vec transformer. In all tested setups, the transformer model failed to converge.

3.4. Voting inference

The models were trained to classify segmented audio samples. However, the final prediction needs to aggregate all of the single-sample predictions for a given patient. Using an end-to-end Multiple Instance Learning setup [2] was restricted due to hardware limitations. Voting inference was proposed to counteract this obstacle. After the models were trained to classify single samples, their predictions were aggregated for each patient. The final output label was the mode value of single-sample predictions. In the results, we report both the single-sample and aggregated voting performance.

3.5. Experimental setup

In our dataset, we gathered 38 PD patients at different stages of the disease's development (a detailed Hoehn-Yahr table is presented Table 1) and 10 healthy persons (HP) above the age of 50. After segmentation, the dataset consisted of a total of 2141 audio samples ranging from vowels to full sentences. Audio had to be resampled from original 44.1 kHz sampling rate to 16 kHz, which is the sampling frequency using which the wav2vec backbone was trained [13]. To verify the hypothesis that knowledge from pretrained natural language audio models can be transferred to medical tasks, we trained our models in 3 configurations:

- baseline model with pretrained and frozen convolutional layers (frozen conv)
- baseline model with pretrained convolutional layers and full fine-tuning (full + pretrained)
- baseline model with randomly initialised layers and full training (full + not pretrained)

The pretrained model that we used was Wav2Vec 2.0 base with no fine-tuning. The GRU was a bidirectional unit with 1 hidden layer and hidden size 256. Classifier head consisted of 2 hidden layers with hidden size equal to 128. Each of the configurations was trained in a 5-fold cross-validation setup. The folds were stratified in terms of Hoehn-Yahr score, meaning that each fold contained patients at different stages of PD. The reported metrics were averaged across the folds. Models were trained for 400 epochs with batch size 32 and Adam optimizer with 10e-4 learning rate and betas equal to (0.9, 0.999) on a Nvidia Tesla K40 XL GPU.

Table 1. Value counts in target subgroups

Group	Hoehn-Yahr grade	Count	Count (group)
PD	5	1	38
	4	11	
	3	13	
	2	11	
	1	2	
HP		10	

4. RESULTS

The results below are presented for a setup described in 3.5, unless otherwise noted. We report averaged 5-fold cross-validated test metrics.

In Figure 5 we present the voting inference metric. The measure is equivalent to the fraction of single-sample predictions that were predicted as PD-positive in a given patient. HP is healthy population. Dotted line is the 0.5 votes threshold between positive and negative grading. An important observation is that the only misclassified subject is a false negative, which very undesired in a medical classification system. Additional metrics, including false negative rate, are shown in Table 3.

Table 2. Comparison of models' scores between different training schemes

Model Single-sample	accuracy Inferred	voting accuracy	Inferred voting ROC AUC
frozen conv	89.92%	97.92%	0.99
+ pretrained	82.07%	83.33%	0.84
full + not pretrained	81.11%	83.33%	0.80

Table 3. Comparison of models' sensitivity and specificity

Model Inferred voting	Sensitivity (true positive rate)	Inferred voting specificity (true negative rate)
frozen conv	0.97	1
full + pretrained	1.00	0.20
full + not pretrained	1.00	0.20

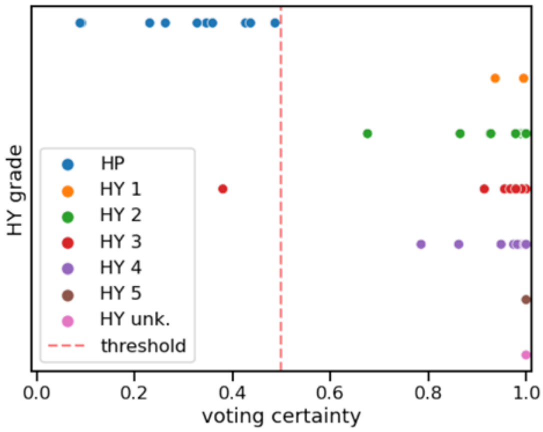


Fig. 5. Plot of the voting certainty at different stages of Hoehn-Yahr scale.

In Table 2, we compare single-sample accuracy to inferred voting accuracy across 3 training setups. Two observations can be drawn from the table:

1. pretraining improves the classification performance;
2. fine-tuning the convolutional part degrades the classification performance.

5. HUMAN-LEVEL PERFORMANCE ASSESSMENT AND INTERPRETABILITY

The aim the performed assessment was to determine if human experts can also pick up some signal in speech recordings solely. A survey was conducted among experts in neurology who did not examine the patients otherwise. In a provided questionnaire, the experts were provided with the recordings sampled from different parts of the phonetic test.

The subsets in the questionnaire consisted of:

- all parts of the phonetic test;
- only full sentences;
- only words and syllables;
- only vowels and sustained phonation.

Six experts took part in the survey. They were asked to label each set of recordings (per patient) with one of the following: no symptoms; mild PD symptoms; advanced PD symptoms; symptoms of a disease other than PD. We provide a detailed table of collected answers in Table Y. Averaged accuracy of the experts predictions on a binary task of PD scores up to 75% when using mode value (similar to the proposed voting inference). We can therefore draw a conclusion that: 1) our model outperforms the human experts in speech classification; 2) there is a significant signal that can be distilled from speech only. This encourages further examination of the model's explainability, which could provide experts with reliable diagnostic input and promote trust in the proposed AI-based diagnostic tool. We also approach the model in terms of interpretability. We wanted to observe if the feature map created by

Table 4. Summary of the answers to the questionnaire. Options in the questionnaire were: 1 - no symptoms, 2 -symptoms other than PD, 3 - early-stage PD, 4 - advanced-stage PD. 'Hit' means that at least one of the experts provided correct answer.

Set 1: all types of recordings				
Subject	True H-Y	Mode(s)	Average	Hit
1 (PD)	1	1	1.3	YES
2 (HP)	-	1, 3	2.0	YES
3 (PD)	3	3	3.0	YES
4 (HP)	-	1	1.0	YES
5 (PD)	5	2, 3	2.5	YES
6 (HP)	-	1	1.7	YES
Set 2: only full sentences				
Subject	True H-Y	Mode(s)	Average	Hit
7 (PD)	2	1	1.3	YES
8 (PD)	2	3	2.3	YES
9 (PD)	4	4	3.3	YES
10 (PD)	4	4	3.3	YES
11 (HP)	-	1	1.0	YES
12 (HP)	-	3	2.7	YES
Set 3: words and syllables				
13 (HP)	-	3	3.2	NO
14 (PD)	2	1	2.3	YES
15 (HP)	-	3	2.7	YES
16 (HP)	-	1	1.5	YES
17 (HP)	-	1	1.3	YES
18 (PD)	4	4	3.2	YES
Set 4: vowels and sustained phonation				
19 (PD)	2	4	3.2	YES
20 (HP)	-	1	1.5	YES
21 (PD)	3	4	3.7	NO
22 (PD)	4	2, 4	3.0	YES
23 (PD)	4	4	4.0	YES
24 (PD)	2	4	4.0	YES

wav2vec convolutional layers can be interpreted in time-feature domain, similar to how spectrograms are analysed in time-frequency domain. We present a sample comparison in figure 6. The spectrogram was calculated with FFT length 1024 and 1/8 window overlap so that the output frequency resolution matched with the number of features in the internal wav2vec representation (512). We observe that there is no interpretable pattern in the feature map, however, the representation is much more evenly distributed than in case of the spectrogram, meaning that it likely yields more information.

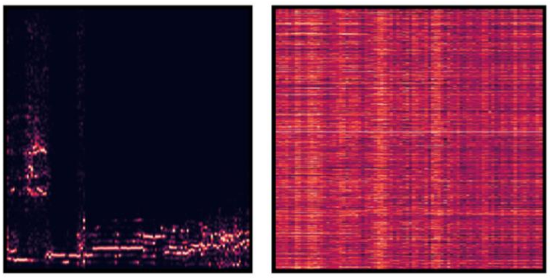


Fig. 6. Comparison of a spectrogram (left) with a wav2vec feature map (right) for a sentence "dziś jest ładna pogoda" ("the weather is nice today"). Spectrogram visualises the acoustic input in time-frequency domain, while extracted feature map does so in a trainable time-features domain

6. DISCUSSION

In our experiments, we have shown that it is possible to use an audio model trained on natural language to improve the performance on a downstream medical task. Our novel contribution is the construction of a machine learning framework for medical audio classification that takes the advantage of existing speech processing models. We have shown that our implementation obtains very good performance on downstream tasks, scoring up to 97.92% accuracy. In needs to be noted, however, that our approaches towards obtaining accurate multi-class predictions for different stages of the disease were so far unsuccessful, most probably due to insufficient representation of each Hoehn-Yahr subset in the training data. Further studies should focus on constructing a model that would differentiate the subjects in terms of the stage of disease's development. Probably, a different grading scale could be used, such as UPDRS. Our classifier should also be used with a given uncertainty margin, especially when considering an implementation of a downstream diagnostic tool. Our method can be used efficiently to separate healthy population from PD patients, but false negative rate has to be taken into account to avoid missing disease-impaired subjects. Another study should also check how the classifier performs in the presence of other diseases, most importantly ones impairing the human speech in any way. Having addressed all these uncertainties, it might be possible to develop a remote diagnostic tool for supporting the traditional clinical PD diagnostic process, based on the proposed method.

Ethical supervision: *All of the patient's data was collected and processed according to the decision of The Bioethics Committee of the Jagiellonian University (decision no. 1072.6120.271.2019 from 21 November 2019).*

Acknowledgements: *Model training was executed on ACK Cyfronet Prometheus cluster, using PLGrid infrastructure.*

The publication has a preprint:
https://arxiv.org/pdf/2204.12559.pdf

Source of funding: *The work was created as part of research project of Department of Mechanics and Vibroacoustics no. 16.16.130.942 AGH and CMUJ in Kraków.*

Author contributions: *research concept and design, M.C., M.K.; Collection and/or assembly of data, M.C., M.D.-Ć., K.P.; Data analysis and interpretation, M.C.; Writing the article, M.C., M.K.; Critical revision of the article, M.K., M.D.-Ć.; Final approval of the article, M.C., M.K., M.D.-Ć., K.P.*

Declaration of competing interest: *The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.*

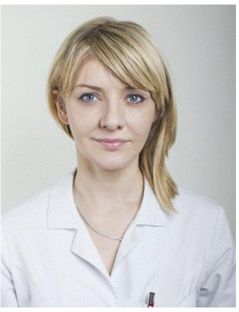
REFERENCES

1. Almeida JS, Rebouças Filho PP, Carneiro T, Wei W. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters 2019; 125: 55–62. https://doi.org/10.1016/j.patrec.2019.04.005.
2. Babenko B. Multiple instance learning: Algorithms and applications. 01 2008.
3. Boualoulou N, Mounia M, Nsiri B, Behoussine Drissi T. A novel Parkinson's disease detection algorithm combined EMD, BFCC, and SVM classifier. Diagnostyka. 2023; 24(4): 2023404. https://doi.org/10.29354/diag/171712.
4. Chronowski M, Kłaczyński M, Dec-Ćwiek M. Speech and tremor tester - monitoring of neurodegenerative diseases using smartphone technology. Diagnostyka 2020;21(2):31–39. https://doi.org/10.29354/diag/122335
5. Han Z, Tian R, Ren P, Zhou W, Wang P, Luo M. Parkinson's disease and Alzheimer's disease: a Mendelian randomization study. BMC Med Genet 2018;19. https://doi.org/10.1186/s12881-018-0721-7.
6. Jordal I, Nishi K, Bredin H. asteroid-team/torchaudiomentations:v0.10.1, 2022. https://doi.org/10.5281/zenodo.6381721.
7. Kłaczyński M. Vibroacoustic methods in diagnosis of selected laryngeal diseases. Journal of Vibroengineering 2015; 17(4): 2089-2098.
8. Liaqat A, Ce Z, Mingyi Z. Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection. Expert Systems with Applications 2019; 137: 22–28. https://doi.org/10.1016/j.eswa.2019.06.052.
9. Mąka J. The Polish linguistic test review in the assessment of Central Auditory Processing Disorders. Investigationes Linguisticae 2009; 18: 55. https://doi.org/10.14746/il.2009.18.4.
10. Reich SG. Does this patient have parkinson disease or essential tremor? Clinics in Geriatric Medicine 2020; 36(1): 25–34. https://doi.org/10.1186/s12881-018-0721-7

11. Ribeiro MT, Singh S, Guestrin C. "Why should trust you?": Explaining the predictions of any classifier. 2016. https://doi.org/10.48550/ARXIV.1602.04938.
12. Sakar B, Isenkul M, Sakar CO. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. Biomedical and Health Informatics, IEEE Journal of 2013; 17: 828–834. https://doi.org/10.1109/JBHI.2013.2245674.
13. Schneider S, Baevski A, Collobert R, Auli M. wav2vec: Unsupervised pre-training for speech recognition. arXiv:1904.05862 [cs], September 2019. https://doi.org/10.48550/arXiv.1904.05862
14. Signaevsky M, Marami B, Prastawa M. Antemortem detection of Parkinson's disease pathology in peripheral biopsies using artificial intelligence. acta neuropathol commun 2022; 10(21). https://doi.org/10.1186/s40478-022-01318-7
15. Sveinbjornsdottir S, Marami B, Prastawa M. The clinical symptoms of Parkinson's disease. Journal of Neurochemistry 2016; 139(S1): 318–324. https://doi.org/10.1111/jnc.13691. 1.

Maurycy CHRONOWSKI - Data Scientist, M.Sc. in Computer Science (specialisation in systems modelling and intelligent data analysis) and B.Eng. in Acoustic Engineering; both degrees graduated from the AGH University of Science and Technology. An enthusiast of modern technologies, passionate about applications of computer intelligence in medicine. Professionally works on developing AI solutions for small molecule drug discovery.
e-mail: moryc.chronowski@gmail.com

Maciej KŁACZYŃSKI - Ph.D. D.Sc. Eng. Prof. AGH, works at Department of Mechanics and Vibroacoustics in AGH University of Science and Technology in Krakow. His research is focused on measurement, signal processing and pattern recognition methods of vibroacoustic signals applied in medicine, technology and environmental monitoring. Author of over one hundred seventy scientific publications and conferences papers. Member of European Acoustics Association (EAA), Polish Acoustical Society (PTA) and Polish Society of Technical Diagnostics (PTDT).
e-mail: mklaczyn@agh.edu.pl



Małgorzata DEC-ĆWIEK - MD, PhD, graduated from Medical School at the Jagiellonian University in Krakow. She has a position of a consultant neurologist in the Department of Neurology at the Jagiellonian University Medical College. She specializes in movement disorders. Specifically, she has been involved in neuromodulation procedures (deep brain stimulation, spinal cord stimulation) for several years. Her research is focused on Parkinson's disease, other parkinsonisms and dystonia. Member of Polish Neurological Society and International Movement Disorder Society.
e-mail: malgorzata.dec-cwiek@uj.edu.pl

Karolina PORĘBSKA MD Ph.D., neurologist, works in Neurology Department of University Hospital and in Jagiellonian University Collegium Medicum in Kraków. During studies she had practice in Jerusalem, Arkhangelsk and was Erasmus student in Rome. Her research is focused on neurodegenerative diseases particularly on Parkinson's disease, dystonia and Alzheimer's disease. She is a member of movement disorders team and a specialist in treating advanced Parkinson's disease with Deep Brain Stimulation (DBS) and infusion therapies (Duodopa and Apomorphine). She has also experience in treating dystonia and spasticity with injections of botulinum toxin under ultrasound guidance.
e-mail: karolina.porebska@uj.edu.pl