A Siamese-based Approach to Improve Parkinson's Disease Detection and Severity Prediction from Speech Using X-Vector Embedding

Attila Zoltán Jenei[1], Réka Ágoston[1], and István Valálik[2]

Abstract—**Parkinson's disease is incurable and is considered one of the most common neurological diseases. It is a progressive disease, which highlights the importance of early detection. Machine learning-based diagnostic support is desirable since the diagnosis is based on history, visual inspection, and drug tests. Speech is presumed to be one of the promising biomarkers that can predict the state of the disease. Combining speech data with deep learning feature extraction in Siamese-based architecture may improve the detection compared with direct regression with acoustic and prosodic features. Read text-based speech samples were acquired from 98 patients with Parkinson's disease and 107 healthy participants. Feature vectors were extracted with pre- trained x-vector embedding and were used directly with a support vector regressor in a nested cross-validation setup (baseline approach). Furthermore, pairs were allocated, and difference vectors were calculated. These difference vectors were then used to train support vector regressor models in nested cross-validation (Siamese-based approach). Severity predictions and classification were performed with the outcomes. The Siamese-based setup outperformed the baseline approach both in regression and classification metrics. The relative improvement in root mean square error is 14.4%, and the Pearson correlation is 12.5% at best. After the classification, the relative improvement is 6.0% in sensitivity, 3.0% in specificity, and 4.5% in accuracy. Furthermore, comparing the test sample to not only one but multiple others decreases the average standard deviation of the predicted severity by 16.5% in relative value. Changing only the architecture of the traditional examination setup to a Siamese-based approach may increase the performance of the models.**

Index Terms—**Classification, Deep-learning, Parkinson's Disease, Siamese Network**

I. Introduction and Literature Study

Parkinson's disease (PD) is one of the most common neurological disorders, which also manifests in movement disabilities. PD affects mainly the aging population and has a prevalence of 1% after 60 years [1]. The importance of detecting the disease in an early stage is its progressive nature and because it is incurable, according to recent clinical knowledge. The development of the disease is characterized by the loss of dopamine-producing cells with the appearance of knowledge. The development of the disease is characterized by the loss of dopamine-producing cells with the appearance of Lewy protein aggregates [2].

The diagnosis of PD is based on the patient's history and examination. The cardinal symptoms are resting tremor, bradykinesia, and rigidity that started asymptotically [3]. The patient may have small handwriting, a masked face, and a soft voice. Vocal disorders are prominent symptoms as they manifest in 90% of PD cases [4]. These motor symptoms are less notable by visual observation at the early stage, which stresses the need for such a diagnostic support system.

The speech affected by PD may be monotonic, have less intensity, and can include sudden stops and starts. Tremors also can appear in the phonation. Speech perceptive analysis is part of the Unified Parkinson's Rating Scale (UPDRS), part III: Motor Examination, where the irregularity can be rated between 0 and 4 [5]. Next to the perceptual analysis, objective methods using descriptive features from the speech can also facilitate the diagnosis in clinical settings. Especially, speech can be acquired non-invasively and analyzed even on a mobile device [6].

According to [7], speech-related studies can be categorized into the following four aspects: a) phonatory, b) articulatory, c) prosodic, and d) cognitive-linguistic studies. Generally, all of them can be used to build a diagnostic support system; however, b) and c) are the most commonly applied in this area. The most common features are jitter, shimmer, noise ratio, pitch, intensity, articulation rate, or pauses. With these descriptors, the machine learning algorithms (such as support vector machine (SVM) or k-nearest neighbor (k-NN)) can reach up to 97% accuracy [8], [9], [10].

In addition to the manual features, deep learning-based feature extraction is also applied to maximize the disease representation in the features. As medical data is hard to acquire in large amounts, transfer learning from another domain is a form of use. These models are initially trained on a large dataset for general purposes, such as speaker recognition. The x-vector and e-capa time-delay neural networks are based on speaker recognition and are still applied in PD detection [11], [12]. The transfer learning-based deep learning models can improve the detection and make the process robust (avoiding overfitting). In [13], Sztahó and his colleagues classified 85 PD and 85 healthy individuals using an x-vector approach with probabilistic linear

[1] Department of Telecommunications and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Budapest, Hungary (E-mail: jenei.attila.zoltan@vik.bme.hu, orcid id: 0000-0003-1007-9907, reka10007@gmail.com).

[2] Department of Neurosurgery, St. John's Hospital, Budapest, Hungary (E-mail: valalik@parkinson.hu).

DOI: 10.36244/ICJ.2025.1.9

discriminant analysis (PLDA) and achieved 84.1% accuracy. The severity of the disease was measured with the Hoehn and Yahr (H&Y) scale [14]. This test assigns a number to the patient between 0 and 5, where 0 means no deviation from normal functioning while 5 means the patient is in bed or a wheelchair due to the disease. This scale is not linear, so an H&Y score of 2 does not mean two times as severe symptoms as the H&Y score of 1. Using this scale, the patients had a mean score of 2.7 with a standard deviation of 1.1. The participants read the "The North Wind and the Sun" tale.

Another approach to overcoming limited medical data is to use distance-based solutions where the pairs of samples could be allocated with higher freedom. One of its applications is in the siamese networks. Bhati and his colleagues [15] used Long Short-Term Memory (LSTM)-based siamese networks to learn better PD representation. Then, they trained classifiers to detect PD with the features resulting from the siamese part. Shalaby and Belal [16] used Siamese networks to enhance the data clustering before classification. They improved the detection by 9.5% relative accuracy.

In the [17] study conducted in 2017, the authors used speech samples from 51 PD patients and 27 healthy participants in Hungarian language. The severity was measured with the H&Y scale. The mean severity was 2.58, with a standard deviation of 0.9. The speech samples included the read version of "The North Wind and the Sun" tale. Prosodic features were used with several regression and classification algorithms. The authors achieved 77.8% sensitivity and 83.6% accuracy with classification, 1.052 RMSE, and 0.73 Pearson correlation with SVM and support vector regression (SVR).

The authors repeated the study [18] two years later. They used samples from 55 PD and 33 healthy participants. The mean severity was almost the same as in the previous study. The prosodic features were extended with various acoustic ones. They reached 84.8% sensitivity, 81.8% specificity, and 83.5% accuracy with SVM. Regression was conducted with SVR, 1.071 RMSE, and 0.72 R^2 (the square of the Pearson correlation coefficient), which also resulted from that study.

We introduced sample pairs and extracted features with x-vector technology based on these studies. Using the difference in the feature vectors, we examined PD's classification and severity regression in the Hungarian language. Our goal is to examine how PD can be detected when we switch from individual samples (and feature vectors) to sample pairs (and feature vector differences) in the same process.

After the *Introduction*, we will present the applied materials and the examination methods in the *Methodology*. We will present the regression and classification results in the *Results* section. Finally, we will discuss the findings and conclude the work in the *Discussion* and *Conclusion* sections.

II. METHODOLOGY

The setup of the experiments can be seen in Fig. 1, where two approaches are detailed. The Baseline approach includes the x-vector feature extraction from the Hungarian Parkinson Speech Dataset (HPSD) and the severity prediction with regressor algorithms. The Siamese-inspired approach takes two samples from the dataset, extracts x-vector features, and calculates the vector difference. Then, this dissimilarity is used to predict severity. In the following subsections, we will detail all components of the approaches.

A. Hungarian Parkinson Speech Dataset

PD patients were recorded at the Semmelweis and Virányos Clinic, Budapest, alongside healthy participants as a control group. 42 females and 56 males were in the PD class with a 65.4 average age and 8.4 standard deviation. 70 females and 37 males were in the healthy control (HC) class with 45.8 average age and 17.7 standard deviation. The sex distribution and age between the two classes appeared significant, with a p-value lower than 0.05 (sex by chi-square test, age by Mann-Whitney U test). These may affect the results when compared to others in the literature. However, we propose a baseline to highlight our findings so the new technique will be compared with a traditional one with the same dataset.

The diagnosis and the severity estimation were done by the neurosurgeon doctor using the H&Y scale. These were made using history, visual examination, and drug tests. The mean severity is 2.8 H&Y with a 0.9 standard deviation. The distribution of the

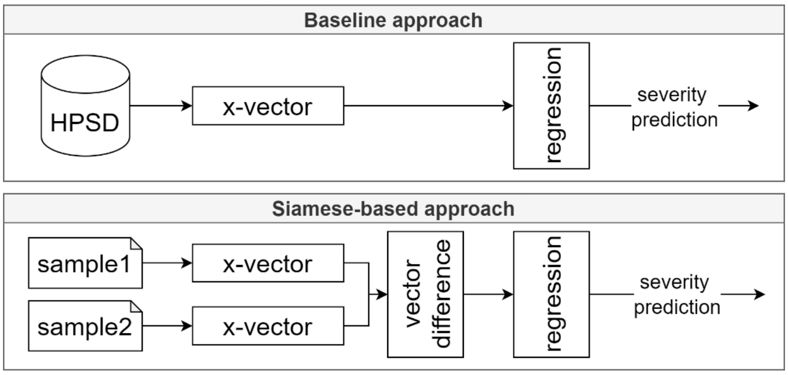


Fig. 1. The Baseline and Siamese-based examination process. HPSD is the dataset, sample1 and sample2 are paired subjects from the dataset.

H&Y severity in the dataset can be seen in Table 1. The HC samples were in the 0 stages of the H&Y scale.

TABLE I
DISTRIBUTION OF THE PATIENTS ACCORDING TO THEIR H&Y SEVERITY SCORE.

	1	**1.5**	**2**	**2.5**	**3**	**4**
PD	10	1	18	12	31	26

The speech task was to read the "*The North Wind and the Sun*" tale, which resulted in around one-minute-long recording per participant. All participants were native Hungarian speakers, and the recordings were done in Hungarian language. The samples were acquired with a clip-on microphone on 16 kHz sampling frequency and 16-bit quantization.

All participants were informed in advance about the research and the use of the samples and metadata. Data acquisition did not mean any harm or risk to the participants. Informed consent was collected from the people involved, and they had (and have) the option to change their minds anytime.

B. Preprocessing and Feature Extraction

The speech samples were normalized to peak value before feature extraction. After that, feature extraction was done with x-vector embedding originating from the speaker recognition applications [19]. The x-vector is a time-delay neural network (its architecture is in Fig. 2) that observes not only the speech frame at time point *t* but its surroundings in the frame level layers. The frame refers to the (sliding) time window, where features like filterbanks can be defined. These layers learn the local and speaker-specific patterns.

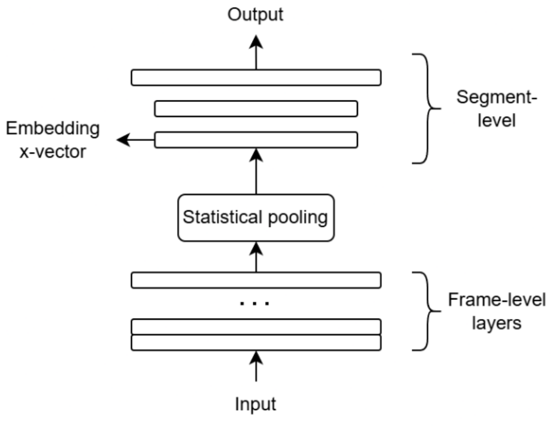


Fig. 2. The x-vector embedding architecture, which parts from the input layer are the frame level layers, statistical pooling, and segment level layers before the output.

Following the frame-level layers, the statistical pooling calculates the mean and standard deviation after aggregating the information by frames. This creates a fixed-length representation of the entire utterance Then, the speaker discriminative features are learned in the segment level from where the x-vector can be extracted with a dimension of 512.

The embedding was implemented through the SpeechBrain toolkit [20], which stores the pre-trained models on the Huggingface website. We used this pre-trained version of the model without further training. The initial training was done on the training data of Voxceleb1 and Voxceleb2 datasets (in English). From the resulting feature vector, the top 65 features were selected with the Random Forest algorithm.

Next, pairs were allocated for the Siamese-based approach to use the features directly with the predictive algorithms (Baseline approach). HC-HC (similar) and HC-PD (dissimilar) pairs were created randomly without repeating already allocated pairs. As a result, 205 pairs for the similar category and 205 for the dissimilar category were examined. 200 pairs were of the same sex, while 210 pairs were of the opposite sex. In this approach, age was not considered during the pair allocation.

C. Severity Prediction and Classification

To estimate the severity of PD in the H&Y scale, we employed a regression method using the Support Vector Regression (SVR) algorithm [21]. The decision was based on the study [18] where the SVR was the prominent algorithm to predict PD severity in the Hungarian language.

The *C*, *gamma*, and *epsilon* parameters were optimized for nested cross-validation. The optimization includes an outer 10-fold cross-validation setup where 10% (one fold) was separated as an independent test set, and the rest was used in the inner 5-fold cross-validation loop. In this inner loop, 80% of the remaining data was used for training and 20% for optimization.

The target severity was normalized between 0 and 1 before inputting them into the predictive models. As a result, the output was generated between 0 and 1, which was up-scaled to the original H&Y scale. With these outputs and the original scores, a linear regression was performed, measuring the mean absolute error (*eq. 1*) and the root mean square error (*eq. 2*). In these metrics, y_i is the original severity of the *i*-th sample, $\widehat{y_i}$ is the predicted score of the *i*-th sample, and *n* is the number of samples.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \widehat{y_i}|}{n} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}} \tag{2}$$

The R^2 value shows how linearly the estimated score fits the original severity score (*eq. 3*). Its value is between 0 and 1, within which a value close to 1 shows a better fit than a value close to 0. The $\overline{y_i}$ in the denominator is the mean of the predicted scores.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{3}$$

We also created classifications with the output score by drawing a decision threshold at H&Y stage 1. Below the score 1, the samples were classified as HC and above as PD. The classification is correct in True Positive (TP) and True Negative (TN) cases and not correct in False Positive (FP) and False Negative (FN) cases.

From these, metrics like sensitivity (*sens – eq. 4*), specificity (*spec – eq. 5*), and accuracy (*acc – eq. 6*) were derived.

$$sens = \frac{TP}{FN + TP} \tag{4}$$

$$spec = \frac{TN}{TN + FP} \tag{5}$$

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

D. Experimental Cases

In this study, we created a Baseline approach, where x-vector embeddings were extracted from the speech samples, and then in a nested cross-validation setup, SVR models were trained and tested. With the output, regression, and classification were performed.

After that, a Siamese-based approach was made by pairing the samples. The x-vector embeddings were extracted from both samples in the pairs similarly, and then the difference between the two vectors was calculated. This vector was the input to the SVR models for regression and classification. Within this approach, we had two cases: 1) create only ONE pair with the selected test sample, and 2) create TEN pairs with the selected test sample and average the predicted scores. Since the train/test split was done after the pair allocation, one pair was only in the train, validation, or test sets. One speaker could be in more sets but not in the same pair.

The *Results* will be divided according to these three cases: a) baseline, b) Siamese approach with one pair, and with ten pairs.

III. Results

A. Baseline approach

The results of the predicted and the original severity scores are presented in the left plot (a part) of Fig. 3 with blue dots for the Baseline approach. The red line represents the perfect mapping of the original scores. Table 2 includes the metrics presented in the next section for all three cases.

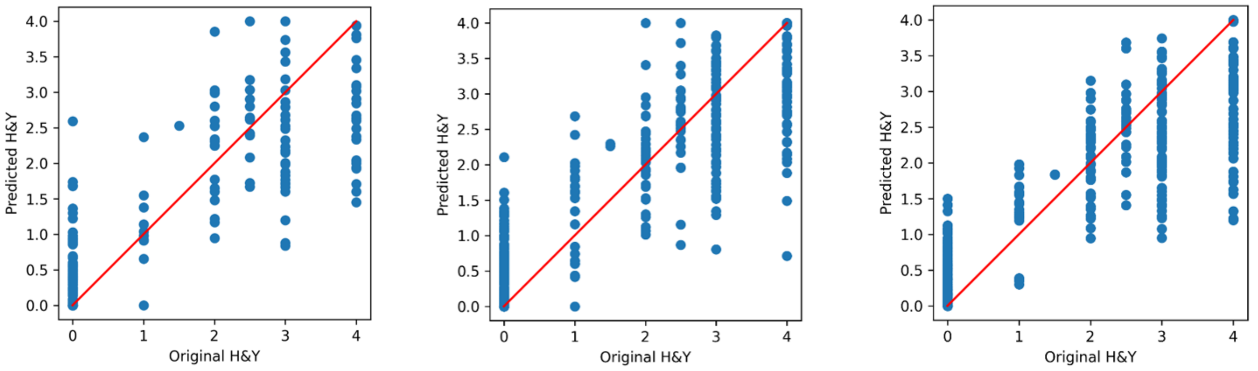
The MAE and RMSE resulted from the predicted and original scores of 0.59 and 0.85. The R^2, which describes the fit of the predicted data to the original, is 0.69. That means that the original score accounts for 69% of the change in the estimated score. Checking the standard deviation of the predicted scores at different H&Y stages, stage 0 had the lowest value at 0.46, and stage 3 had the highest at 0.79. The standard deviation for the other stages is 0.58 (stage 1), 0.76 (stage 2), and 0.73 (stage 4). This deviation can be seen in the figure where the points at stage 3 are scattered farther from the red line. The model predicts stages 1, 2, and 2.5 symmetrically while stage 0 is over-, and stage 4 is underestimated. Also, stage 3 has a slight bias toward lower scores.

After transforming these scores to labels (performing the classification), 91.6% sensitivity, 92.4% specificity, and 92.0% accuracy scores were achieved. The ratio between the positive and negative classes also remains in the predictive scores, as sensitivity and specificity have almost the same values.

B. Siamese-based approach

The results of the Siamese-based approach can be seen in Fig. 3, where the middle plot (b part) presents the one-pair case, and the right side (c part) presents the average of the ten-pair case.

Comparing the test sample with only one other sample (using one pair) resulted in 0.51 MAE, 0.73 RMSE, and 0.78 R^2 metrics. The standard deviations are 0.41, 0.70, 0.62, 0.69, and 0.70 at the stages of the severity scale from 0 to 4, respectively. Symmetrical distribution can be seen at stages 1, 2, and 2.5.



a) Baseline approach

b) Siamese approach with ONE pair

c) Siamese approach with TEN pairs

Fig. 3. Original and predicted severity scores
with the Baseline approach (a) and with Siamese approach (b – where one pair, c – where ten pairs were allocated.).

Stage 0 is over-, and stages 3 and 4 are underestimated by the models. After calculating classification metrics, 95.1% sensitivity, 91.7% specificity, and 93.4% accuracy were achieved.

When we compared one test sample to ten other test samples and averaged the predicted scores, 0.55 MAE, 0.77 RMSE, and 0.75 R^2 were the results. The standard deviations were 0.31, 0.50, 0.55, 0.69, and 0.72 according to the stages from 0 to 4. Here, stages 2 and 2.5 seem symmetrical, while stages 0,1 are over-, and 3 and 4 are underestimated. After the classification, 97.1% sensitivity, 95.1% specificity, and 96.1% accuracy were observed.

Comparing the three experiment cases, *no significant difference* can be noted between the predicted severity scores or the classification results with Mann-Whitney tests next to 5% significant level. Furthermore, the difference was calculated between the predicted and the original severity score separately for the male-male/female-female and the female-male pairs. The mean difference was 0.04 for the pairs from the same gender and 0.009 for the pairs from the opposite gender. The standard deviation of the differences was around the same.

IV. Discussion

The results of the Baseline and Siamese-based approaches are summarized in Table 2, along with the metrics described above. Baseline refers to the Baseline approach, ONE pair to the one-pair version, and TEN pairs to the ten-pair version of the Siamese approach. Bold style marks the best values for each metric.

Comparing the regression metrics to the different cases, it can be seen that the Siamese-based approach decreased the error in the predictions as MAE and RMSE decreased. Both one-pair and ten-pair cases were better in metrics than the Baseline approach. However, the ten-pair version of the Siamese-based approach did not yield better than the one-pair version examining the regression metrics. A similar tendency can be seen with the R^2 as it is higher with the Siamese-based approach than the Baseline. However, the ten-pair version is not better than the one-pair version.

Comparing the classification results, the improvement is continuous. Sensitivity increased from 91.6% to 97.1%, while the specificity was slightly lower with one pair but increased with ten pairs. The accuracy improved from 92.0% to 96.1%.

TABLE II

Summary table of the Baseline and Siamese-based (one and ten-pair versions) approaches.

	MAE	RMSE	R2	sens	spec	acc
Baseline	0.59	0.85	0.69	91.6%	92.4%	92.0%
ONE pair	**0.51**	**0.73**	**0.78**	95.1%	91.7%	93.4%
TEN pair	0.55	0.77	0.75	**97.1%**	**95.1%**	**96.1%**

Comparing our results with similar studies [17], [18], it can be seen that the results achieved in this study outperformed them. Even the Baseline method reached lower error and higher classification metrics. This could be due to the extended database and the nested cross-validation setup for the optimization.

The results could improve compared to the Baseline method by facilitating the Siamese-inspired approach. This result highlights the possible benefit of increasing the number of input data by creating pairs. However, comparing someone to only one other person is not always reliable. We compared one test sample to ten others to overcome this limitation and averaged the predicted scores. This technique seemingly did not improve the regression metrics but did improve the classification. This is because the average standard deviation decreased from 0.62 to 0.55 by averaging the pair's predictions. With this, the samples that were misclassified with only one pair were classified correctly by the other 9 pairs. 188 TN and 195 TP classification were done with the one-pair case, while 195 TN and 199 TP with the ten-pair case.

Comparing our results to the existing solutions in the international literature is cumbersome due to the different datasets with different languages. However, we believe that our results fit into this international level. Furthermore, the x-vector was used here without further training; it was initially trained with English samples. The usage of such model is promising since we could use it in the Hungarian language without the language's specifications.

The present study concentrated on the the x-vector algorithm. Nevertheless, the employment of alternative algorithms in speech technology (like transformers) has the potential to enhance the findings of this study.

VI. Conclusion

PD is one of the most common neurological disorders that is not curable according to recent clinical knowledge. Early detection is important to maintain the quality of life and slow disease progression. Speech is one of the promising biomarkers that can capture the pre-motor symptoms of the disease. Many studies used acoustic and prosodic features to describe the speech and used them with statistical methods or machine learning algorithms to distinguish between PD and other groups. Next to the manual features deep learning-based feature extraction and detection became prevalent due to the more detailed representation. However, these solutions require a huge amount of data.

In this study, we explored the possibilities of the Siamese-based architecture with PD patients and HC participants using read-text-based speech. The results indicated that the Siamese-based approach outperforms the Baseline in regression and classification. Pairing the sample with more than one other sample decreases the standard deviation of predictions and improves the classification further. The results fit the literature since they outperform studies with similar datasets and are also comparable with similar studies that use different languages or databases.

It should be noted that this study used the x-vector trained on English utterances while the samples in scope were in Hungarian. However, the results show high performance without specifying the embedding to Hungarian. This is also promising for language-independent solutions. Additional analysis may be required to discover possible influencing effects like age or sex (even though the results showed no significant influence of these effects).

Acknowledgments

The work was funded by the National Research, Development and Innovation Office – NKFIH, project K143075.

References

[1] O.-B. Tysnes and A. Storstein, 'Epidemiology of Parkinson's disease', *J Neural Transm*, vol. 124, no. 8, pp. 901–905, Aug. 2017, **DOI**: 10.1007/s00702-017-1686-y.

[2] E. Tolosa, A. Garrido, S. W. Scholz, and W. Poewe, 'Challenges in the diagnosis of Parkinson's disease', *The Lancet Neurology*, vol. 20, no. 5, pp. 385–397, May 2021, **DOI**: 10.1016/S1474-4422(21)00030-2.

[3] P. Rizek, N. Kumar, and M. S. Jog, 'An update on the diagnosis and treatment of Parkinson disease', *CMAJ*, vol. 188, no. 16, pp. 1157–1165, Nov. 2016, **DOI**: 10.1503/cmaj.151179.

[4] G. Solana-Lavalle and R. Rosas-Romero, 'Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation', *Biomedical Signal Processing and Control*, vol. 66, p. 102 415, Apr. 2021, **DOI**: 10.1016/j.bspc.2021.102415.

[5] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, 'Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation', *J Neural Transm*, vol. 124, no. 3, pp. 303–334, Mar. 2017, **DOI**: 10.1007/s00702-017-1676-0.

[6] T. J. Wroge, Y. Ozkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H. Ghomi, 'Parkinson's Disease Diagnosis Using Machine Learning and Voice', in *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Philadelphia, PA: IEEE, Dec. 2018, pp. 1–7. **DOI**: 10.1109/SPMB.2018.8615607.

[7] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, 'Advances in Parkinson's Disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects', *Biomedical Signal Processing and Control*, vol. 66, p. 102 418, Apr. 2021, **DOI**: 10.1016/j.bspc.2021.102418.

[8] O. Yaman, F. Ertam, and T. Tuncer, 'Automated Parkinson's disease recognition based on statistical pooling method using acoustic features', *Medical Hypotheses*, vol. 135, p. 109 483, Feb. 2020, **DOI**: 10.1016/j.mehy.2019.109483.

[9] A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad, 'Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease', in *2015 International Conference on Electrical and Information Technologies (ICEIT)*, Marrakech, Morocco: IEEE, Mar. 2015, pp. 300–304. **DOI**: 10.1109/EITech.2015.7163000.

[10] K. Wu, D. Zhang, G. Lu, and Z. Guo, 'Learning acoustic features to detect Parkinson's disease', *Neurocomputing*, vol. 318, pp. 102–108, Nov. 2018, **DOI**: 10.1016/j.neucom.2018.08.036.

[11] L. Moro-Velazquez, J. Villalba, and N. Dehak, 'Using X-Vectors to Automatically Detect Parkinson's Disease from Speech', in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, May 2020, pp. 1155–1159. **DOI**: 10.1109/ICASSP40776.2020.9053770.

[12] L. Jeancolas *et al.*, 'X-Vectors: New Quantitative Biomarkers for Early Parkinson's Disease Detection From Speech', *Front. Neuroinform.*, vol. 15, p. 578 369, Feb. 2021, **DOI**: 10.3389/fninf.2021.578369.

[13] D. Sztahó, A. Z. Jenei, I. Valálik, and K. Vicsi, 'The Effect of Speech Fragmentation and Audio Encodings on Automatic Parkinson's Disease Recognition', *JBiSE*, vol. 15, no. 01, pp. 6–25, 2022, **DOI**: 10.4236/jbise.2022.151002.

[14] R. Bhidayasiri and D. Tarsy, 'Parkinson's Disease: Hoehn and Yahr Scale', in *Movement Disorders: A Video Atlas*, in Current Clinical Neurology., Totowa, NJ: Humana Press, 2012, pp. 4–5. **DOI**: 10.1007/978-1-60327-426-5_2.

[15] S. Bhati, L. M. Velazquez, J. Villalba, and N. Dehak, 'LSTM Siamese Network for Parkinson's Disease Detection from Speech', in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Ottawa, ON, Canada: IEEE, Nov. 2019, pp. 1–5. **DOI**: 10.1109/GlobalSIP45357.2019.8969430.

[16] M. Shalaby, N. A. Belal, and Y. Omar, 'Data Clustering Improves Siamese Neural Networks Classification of Parkinson's Disease', *Complexity*, vol. 2021, pp. 1–9, Jun. 2021, **DOI**: 10.1155/2021/3112771.

[17] D. Sztaho, M. G. Tulics, K. Vicsi, and I. Valalik, 'Automatic estimation of severity of Parkinson's disease based on speech rhythm related features', in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Debrecen: IEEE, Sep. 2017, pp. 000 011–000 016. **DOI**: 10.1109/CogInfoCom.2017.8268208.

[18] D. Sztaho, I. Valalik, and K. Vicsi, 'Parkinson's Disease Severity Estimation on Hungarian Speech Using Various Speech Tasks', in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Timisoara, Romania: IEEE, Oct. 2019, pp. 1–6. **DOI**: 10.1109/SPED.2019.8906277.

[19] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, 'Deep Neural Network Embeddings for Text-Independent Speaker Verification', *Interspeech 2017*, Aug. 2017, **DOI**: 10.21437/interspeech.2017-620

[20] M. Ravanelli *et al.*, 'SpeechBrain: A General-Purpose Speech Toolkit'. arXiv, Jun. 08, 2021. Accessed: Jun. 24, 2024. [Online]. Available: http://arxiv.org/abs/2106.04624

[21] M. Awad and R. Khanna, 'Support Vector Regression', in *Efficient Learning Machines*, Berkeley, CA: Apress, 2015, pp. 67–80. **DOI**: 10.1007/978-1-4302-5990-9_4.

Attila Zoltán Jenei was born in Debrecen, Hungary in 1995. He graduated from the Budapest University of Technology and Economics as a Biomedical Engineer (Master's Degree, 2020). Since January 2020, he has been a department engineer and Ph.D. student at the Laboratory of Speech Acoustics, Department of Telecommunications and Artificial Intelligence, Faculty of Electrical Engineering and Informatics. His research focuses on diagnostic support for Parkinson's disease with non-invasive medical data. He participated in the Student Research Societies of Budapest University of Technology and Economics and was awarded in 2017 and 2019. From 2021, he was the Vice President, and from 2023 to 2024, he was the President of the Department of Engineering Sciences in the National Association of Doctoral Students.

Réka Ágoston was born in Dunaújváros in 1999. She earned a bachelor's degree in biochemical engineering from Pannon University. She continued her studies at the Budapest University of Technology and Economics (BME), where she pursued a master's degree in biomedical engineering. For her master's thesis, she used machine learning to analyze speech data for Parkinson's disease severity detection, combining her passion for biomedical data and machine learning.

István Valálik MD., Ph.D., MSc. neurosurgeon and head physician of the Department of Neurosurgery at St. John's Hospital, Budapest, honorary associate professor at the University of Debrecen. In 2011 he defended PhD thesis "CT-guided stereotactic thermolesion and deep brain stimulation in the treatment of Parkinson's disease", in 2015 MSc in Health Services Management. His scientific interest focused on movement disorders, MR-tractography-based surgical planning, psychiatric surgery, acoustic and motion analysis. He developed a planning software for stereotactic brain surgery and portable neuro-navigation system. Since 2010 he is acting in the Executive Committee of the European Society for Stereotactic and Functional Neurosurgery (www.essfn.org). In 2013 he was awarded by the Hungarian Academy of Sciences for the book "Stereotactic and Functional Neurosurgery". In 2019 he participated in mission of successful surgical separation of Bangladeshi craniopagus twins.