sciendo
ISSN 1335-8871
MEASUREMENT SCIENCE REVIEW
Journal homepage: https://content.sciendo.com

# Evaluating the Performance of wav2vec Embedding for Parkinson's Disease Detection

Ondřej Klempíř[1*], David Příhoda[2], Radim Krupička[1]

[1]*Department of Biomedical Informatics, Faculty of Biomedical Engineering, Czech Technical University in Prague, nám. Sítná, 3105, 272 01, Kladno, Czech Republic,* klempond@gmail.com
[2]*Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology, Technicka, 5, 160 00, Prague, Czech Republic*

Abstract: Speech is one of the most serious manifestations of Parkinson's disease (PD). Sophisticated language/speech models have already demonstrated impressive performance on a variety of tasks, including classification. By analysing large amounts of data from a given setting, these models can identify patterns that would be difficult for clinicians to detect. We focus on evaluating the performance of a large self-supervised speech representation model, wav2vec, for PD classification. Based on the computed wav2vec embedding for each available speech signal, we calculated two sets of 512 derived features, wav2vec-sum and wav2vec-mean. Unlike traditional signal processing methods, this approach can learn a suitable representation of the signal directly from the data without requiring manual or hand-crafted feature extraction. Using an ensemble random forest classifier, we evaluated the embedding-based features on three different healthy vs. PD datasets (participants rhythmically repeat syllables /pa/, Italian dataset and English dataset). The obtained results showed that the wav2vec signal representation was accurate, with a minimum area under the receiver operating characteristic curve (AUROC) of 0.77 for the /pa/ task and the best AUROC of 0.98 for the Italian speech classification. The findings highlight the potential of the generalisability of the wav2vec features and the performance of these features in the cross-database scenarios.

Keywords: Classification, deep learning, features embedding, Parkinson's disease, wav2vec.

## 1. INTRODUCTION

Age is the greatest risk factor for the development and progression of neurodegenerative disorders, such as Parkinson's disease (PD) [1]. Neurodegenerative disorders usually occur because of neuronal death in the brain [2]. In the case of PD, this is the death of neurons in the basal ganglia, specifically in the substantia nigra area, which produces a dopamine neurotransmitter [3]. The lack of this chemical substance leads to a wide range of problems typical of PD, e.g. tremor, stiffness, slowing, gait disorders and speech disorders. The onset of PD occurs most frequently in people over the age of 60 [4]. As human life lengthens, the number of people affected by PD also increases.

The new generation of large language models is experiencing rapid development in medicine [5]. Pre-trained models can be tested in a wide variety of areas. One of the promising areas is audio processing. Speech, including articulation, is an important cue for motor function and is extremely sensitive to impairments in neurological diseases [6]. In PD, individuals often exhibit a range of speech disorders that significantly impair communication. These include distinctive patterns such as monoloudness, monopitch, reduced stress, imprecise articulation, variability in speech rate, a breathy and raspy voice, disfluency, voice tremor and other manifestations that can lead to an overall reduced intelligibility of speech [7]. Therefore, a speech disorder is also considered one of the most serious manifestations of PD [8]. Advances in speech assessment and an objective description of the changes that voice and speech undergo during neurodegenerative disease development may reveal future impairments. The importance of automatic speech analysis in PD is reflected in the extensive research on machine learning applications, with two essential tasks of interest − classification and regression. The fundamental and general role, not only in original research publications but also in international competitions and hackathons, is to classify a healthy control group (HC) vs. PD group. For clinical practise, an equally important task is to estimate the degree of disability by developing a regression model, usually indicated by the unified Parkinson's disease rating scale (UPDRS) [9], articulation rate [10], pronunciation intelligibility [11] or other socio-demographic descriptors such as age and gender [12]-[13].

There are studies based on a large sample of prodromal and early-stage participants that have demonstrated the effectiveness of automated speech analysis in identifying

DOI: 10.2478/msr-2023-0033
*Corresponding author: klempond@gmail.com (O. Klempir)*

early-stage PD and its prodromal stages, such as idiopathic rapid eye movement sleep behaviour disorder (iRBD). For example, Jeancolas et al. [14] presented accurate detection of early-stage PD and iRBD through a high-level feature extraction process involving aspects such as prosody, phonation, speech fluency and rhythm using specialised software called Praat [15]. The study by Rusz et al. [16] focused on early, untreated PD and used a comprehensive acoustic voice assessment across ten distinct speech dimensions, including phonation, articulation, prosody, and speech timing. Furthermore, progress was made in addressing challenges across multiple languages [17] and in categorising speech subtypes within de novo PD [18]. The latter involved the development of a fully automated acoustic quantitative assessment approach for the 7 distinctive patterns of hypokinetic dysarthria. While these studies have been very successful in detecting the progression of disease early, it is important to note that the specific features calculated may vary between studies and are often tailored to a specific PD detection task.

Machine learning for speech processing has seen many improvements in recent years. To highlight a relevant example [19]: Speech emotion recognition with deep convolutional neural networks (CNN) is an article by Issa et al. This work provides a framework for speech emotion recognition using CNN when fed a set of mel-frequency cepstral coefficients (MFCC). As for specialised pipelines for machine learning of speech in PD, many authors have tested the performance of a variety of models and pre-processing phases. Tuncer et al. proposed a minimum-average-maximum tree and singular value decomposition to extract a novel feature signal, subsequently processed by the k-nearest neighbour classifier [20]. Another novel feature introduced in the article by Karan et al. is an intrinsic mode function cepstral coefficient, which should lead to higher classification accuracy compared to standard MFCCs [21]. A non-linear dynamic complexity measure, a discrete wavelet transform, measures of fundamental frequency variation (jitter) and measures of amplitude variation (shimmer) are common baseline features that describe input speech recordings. When combined with other advanced features, such as tunable Q-factor wavelet transform features, and appropriate subsequent feature subset selection, high values for accuracy (94.7%), sensitivity (98.4%), specificity (92.7%) and precision (97.2%) were observed [22]. Neural networks must certainly not be missing from the examples of machine learning models for PD voice recordings classification [23].

As documented in the previous sections, computer-aided methods used in recent years to determine speech parameters in PD are very accurate. However, the processing pipelines, including feature extraction, contain many specific steps and these approaches are not general enough to describe PD speech and different types of tasks. For these reasons, it would be promising to find an automatic feature extraction approach that has generalised applicability to different types of PD speech classification tasks and could simplify highly specialised automated speech recognition (ASR) systems with a comparable degree of accuracy.

Innovative approaches to deep learning (DL) for representing numerical audio vectors are increasingly coming into the limelight. One of the most recent is Facebook's wav2vec embedding AI. wav2vec (including version 2.0 with transformer encoder) is a very promising approach, showing powerful speech recognition in languages for which there are no large datasets to train [24]. wav2vec is trained on a corpus with large amounts of unlabelled audio data. Unlike the traditional signal processing methods, this method can learn a suitable representation of the audio signal directly from the data without requiring manual feature extraction. According to the latest findings, wav2vec can be used across languages [25]. The principle of the method is similar to common types of embedding in numeric vectors, such as word2vec in natural language processing [26] or pfam2vec in bioinformatics [27]. The effectiveness of self-supervised pre-training for ASR has already been demonstrated for non-medical applications [28]. Recently, Bayerl et al. compared x-vectors, ECAPA-TDNN, and wav2vec 2.0 embeddings in a corpus of academic spoken English to detect vocal fatigue [29].

In the biomedical field, wav2vec 2.0 is a speech recognition system that has been used to evaluate cognitive disorders [30]. ASR systems have also been used in the field of dysarthric speech recognition [31]. In both cases, the focus of the evaluation was on the word error rate (WER), which is a common measure of the accuracy of speech recognition systems. Recently, the wav2vec 2.0 representations of speech were found more effective in distinguishing between PD and HC subjects compared to language representations including word-embedding models [32].

The aim of this article is to evaluate wav2vec on three different PD datasets and demonstrate its applicability in achieving high performance for supervised classification tasks without requiring manual or hand-crafted feature extraction. We aim to study the generalisability of wav2vec features and the performance of these features in cross-database scenarios.

## 2. SPEECH DATABASES & METHODS

### *A. Dataset-1: Participants rhythmically repeat syllables /pa/*

As a first dataset, we used the data analysed in our previous study [33], in which the training signals of 30 male PD and 30 male age-matched HC underwent an extraction of relatively well-discriminating features in terms of energy and spectral speech properties. The data consisted of audio signals in wav format with a sampling frequency of 48 kHz. We focus here primarily on the evaluation of the "pa" recordings, which are considered as a standardised speech examination in PD, regardless of the speaker's language. For each participant 2 recordings were available, in this analysis we trained the model primarily with the first of them. In this setup, we wanted to test whether it was sufficient to record only one 30-second recording of the "pa" task.

### *B. Dataset-2: Italian Parkinson's speech*

The second dataset analysed in this study was related to the paper by Dimauro et al. (2017) [11]. This report included data from the Università degli Studi di Bari, Dipartimento di Informatica, Italy. The original study assessed speech intelligibility in PD using a proprietary Speech-to-Text API

powered by Google. For our experiments and model validation, we used a dataset of 50 available subjects (HC = 22, PD = 28). For HC, persons aged 60-77 years were included, 10 men and 12 women. None of the persons reported specific speech or language disorders. For PD, patients aged 40-80 years were included, 19 men and 9 women. None of the patients reported speech or language disorders unrelated to their PD symptoms. All patients were receiving antiparkinsonian treatment. As PD mainly affects older people, a young HC group was not considered for our wav2vec model training. The data consisted of audio signals in wav format with a sampling frequency of 44.1 kHz. The content of the recordings were 2 copies of reading a phonemically balanced text as well as recordings of a repeated execution of syllables. We focused primarily on the reading part of speech. The measurements of the text readings are available for each individual with the corresponding class information.

### *C. Dataset-3: HC and PD voice recordings at King's College London (KCL)*

We chose this English dataset [34] for validation and to demonstrate the suitability of the presented methods across the languages of the subjects (HC = 21, PD = 16), and to project these data onto the Italian data source (Dataset-2). To our knowledge, no information is available on gender and age balance. They asked participants to read aloud "The North Wind and the Sun". The measurements (with a sampling frequency of 44.1 kHz) of the text readings are available for each individual with the corresponding class information.

### *D. Signal processing using wav2vec embedding*

wav2vec generates its own features it has learned with a large data set. It uses a multi-layer CNN architecture to encode past context. Representations are learned by predicting the future in latent space under a contrastive binary binary classification task [24]. It implements a model consisting of a DL hierarchy built on multiple layers of CNNs. Our implementation was mainly based on the python libraries pytorch and fairseq (fairseq.models.wav2vec) [35]. We downloaded a publicly available wav2vec-large model trained with the LibriSpeech training corpus, which contains 960 hours of 16 kHz English speech [36]. A collection of wav2vec models has recently been extended with a Czech corpus [37].

The full high-level pipeline of the proposed method is shown in Fig. 1. Inspired by wav2vec success in language modelling, we build on its architecture to train the models and use these models in various classification tasks. The pipeline can be applied directly to recordings in raw wav format. To facilitate wav2vec feature computation, we have prepared and tested an executable command-line tool that connects python wav2vec to a Windows version of MATLAB 2018b. This solution might be of interest to non-python users/scientists and for using the method in a production environment. We used PyInstaller to package the python solution and include all necessary dependencies [38]. The main solution is included in the Makefile.

In this project, the audio data was pre-processed using the wav2vec-large model, which required resampling all signals to 16 kHz. We calculated the mean and sum of the wav2vec embedding to create a 512-dimensional feature vector for each audio signal. To be more specific, we used the wav2vec-mean or wav2vec-sum technique to extract an audio representation from a matrix representation, resulting in a form of 512 features per signal. This method is insensitive to the duration of the audio signal. All calculations were performed on CPU to ensure functionality outside of the hardware we used.

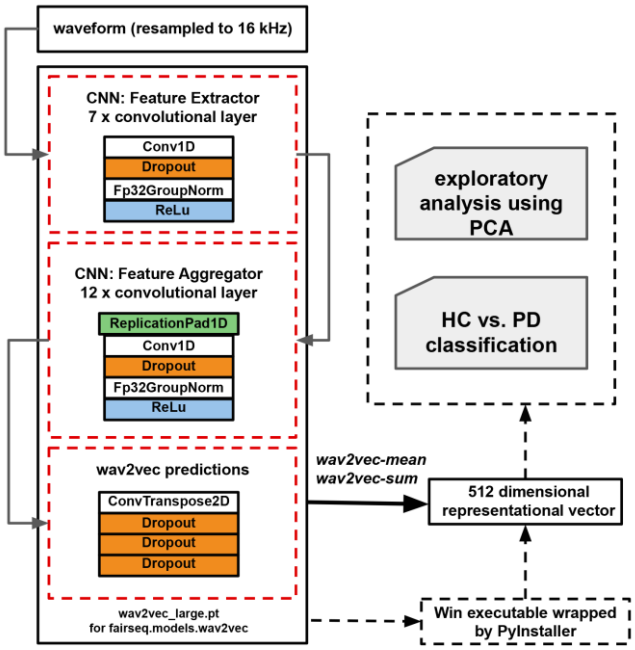


Fig. 1. High level overview of the proposed wav2vec methodology and the corresponding experiments.

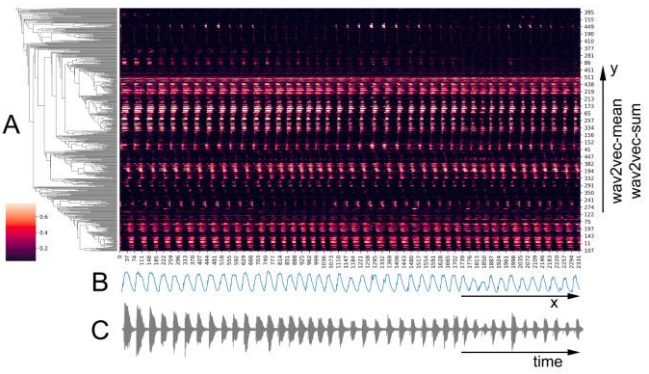


Fig. 2. (A) Two-dimensional clustermap of a full wav2vec embedding for one audio signal from the Dataset-1. Computed wav2vec waveform (B) and its correspondence with a raw signal (C).

A full wav2vec embedding (512 channels x receptive field of the context network-proportional to signal length) for one audio signal from the Dataset-1 is shown in Fig. 2. The x-axis corresponds to the resampled waveform. Aggregation of this matrix (sum or mean) across the x-axis will result in the computation of 512 features. In Fig. 2(A), hierarchical clustering of 512 wav2vec channels was applied to a given

audio signal (using cosine similarity). Individual time-dependent channels clustered according to the similarity of their waveforms are shown on the y-axis. To enable aligning wav2vec embedding with the raw waveform (Fig. 2(C)), the mean operation was computed over the y-axis of the wav2vec embedding (Fig. 2(B)).

### *E. Modelling and validation*

In this study, a series of analytical methods, starting with principal component analysis (PCA), were used to explore either wav2vec-sum or wav2vec-mean embedding in a reduced feature space. The modelling phase then consisted of binary classification using an ensemble random forest classifier (imported from python scikit-learn). We used supervised classification to classify speech signals into two categories based on wav2vec features. The length of the recordings and the design of the individual studies used varied. The language (English vs. Italian) of the participants also varied. For this reason, and to verify the validity of the models, judge the model's applicability domain, and, most importantly, avoid overfitting, we performed 5-fold cross-validation with 5 repeated fits to evaluate the performance of our classification model. The cross-validation process ensured that our model did not overfit to the training data and that the results could be generalised to new data. Standard metrics such as area under the receiver operating characteristic curve (AUROC), precision-recall or accuracy were calculated to demonstrate the obtained results. We implemented a custom python library for plotting all results, including scatter plots or the confusion matrix, to get scores for the true positive rate (TPR) or false positive rate (FPR).

## 3. RESULTS

### *A. Dataset-1: Participants rhythmically repeat syllables /pa/*

Table 1 summarises the unweighted fold average results of our experiments. We also plotted the average AUROC curve and the average precision-recall curve for wav2vec-sum in Fig. 3(A) and Fig. 3(B), respectively. The classification results showed that the random forest algorithm achieved a higher average AUROC score of 0.77 for wav2vec-sum compared with wav2vec-mean (AUROC = 0.66). In addition, the wav2vec-sum model achieved an AUROC of 0.81, an accuracy of 0.71, a precision of 0.73 and a recall of 0.73 in an unweighted fold average case. We also calculated the cumulative confusion matrix to further analyse the performance of our method. Fig. 3(C) shows the confusion matrix for Dataset-1. The plot shows cumulative wav2vec-sum results for individual folds and repeats.

Table 1. Dataset-1: Unweighted fold average AUROC, accuracy, precision, and recall for wav2vec-sum and wav2vec-mean.

| Metric | wav2vec-sum | wav2vec-mean |
|---|---|---|
| AUROC | 0.81 | 0.69 |
| Accuracy | 0.71 | 0.61 |
| Precision | 0.73 | 0.62 |
| Recall | 0.73 | 0.68 |

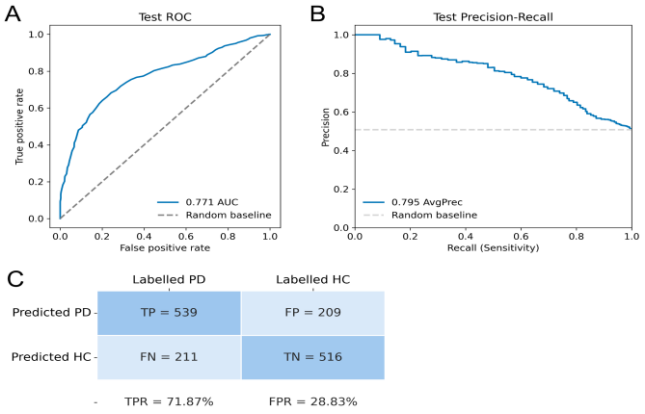


Fig. 3. Test wav2vec-sum AUROC curve (A) and test wav2vec-sum precision-recall curve (B) for Dataset-1. (C) Confusion matrix for Dataset-1.

### *B. Italian Parkinson's speech*

The reported results have a similar structure to Dataset-1. Table 2 summarises the unweighted fold average results of our experiments. Both the wav2vec-mean and wav2vec-sum models performed extremely well. For the wav2vec-mean with an AUROC of 0.98, precision was 0.94, recall was 0.87 and accuracy was 0.95. The average wav2vec-mean AUROC curve for Dataset-2 is shown in Fig. 4(A), and the precision-recall curve is shown in Fig. 4(B). The confusion matrix for Dataset-2 is shown in Fig. 4(C). The plot shows cumulative wav2vec-mean results for individual folds and repeats.

Table 2. Dataset-2: Unweighted fold average AUROC, accuracy, precision and recall for wav2vec-sum and wav2vec-mean.

| Metric | wav2vec-sum | wav2vec-mean |
|---|---|---|
| AUROC | 0.97 | 0.98 |
| Accuracy | 0.93 | 0.95 |
| Precision | 0.92 | 0.94 |
| Recall | 0.95 | 0.97 |

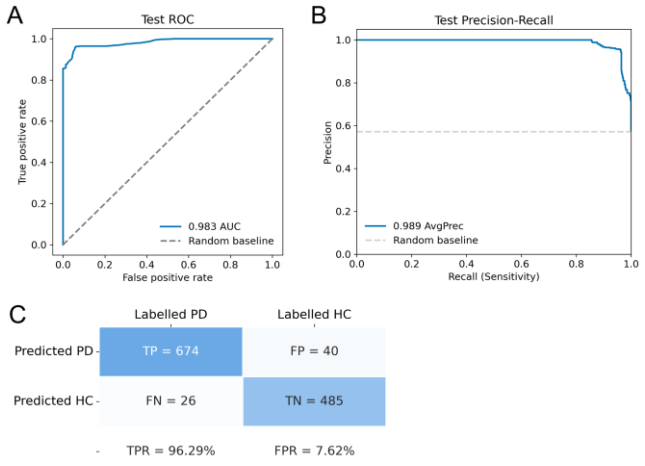


Fig. 4. Test wav2vec-mean AUROC curve (A) and test wav2vec-mean precision-recall curve (B) for Dataset-2. (C) Confusion matrix for Dataset-2.

### *C. Binary classification leave-group-out on combined Dataset-2 and Dataset-3*

In this section, we elaborate on the leave-group-out classification, where each group-class (Italian vs. English) was used as a test set, while the rest was used as a training set. First, we used 2D PCA to visualise the distribution of individual recordings in a lower-dimensional space (Fig. 5(A)). This allowed us to more clearly identify individual clusters in the data. As for the obtained classification leave-group-out results, the AUROC curve is shown in Fig. 5(B) and the precision-recall curve is shown in Fig. 5(C). Table 3 summarises the unweighted fold average results of the leave-group-out experiments. The wav2vec-mean method achieved an average AUROC of 0.78 and a precision-recall of 0.77. The ability to differentiate between the healthy and control groups for individual languages in a leave-group-out scenario is also shown in Fig. 5(B). The confusion matrix is shown in Fig. 5(D). The plot shows cumulative wav2vec-mean results for individual folds and repeats.

Table 3. Leave-group-out on combined Dataset-2 and Dataset-3: Unweighted fold average AUROC, accuracy, precision and recall for wav2vec-sum and wav2vec-mean.

| Metric | wav2vec-sum | wav2vec-mean |
|---|---|---|
| AUROC | 0.67 | 0.81 |
| Accuracy | 0.49 | 0.68 |
| Precision | 0.49 | 0.63 |
| Recall | 0.98 | 0.93 |

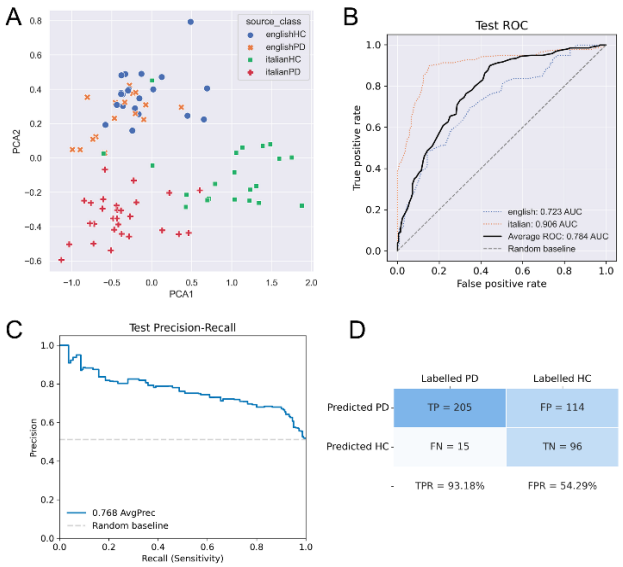


Fig. 5. (A) Principal component analysis for combined Dataset-2 and Dataset-3 using wav2vec-mean features. Test wav2vec-mean AUROC curve (B) and test wav2vec-mean precision-recall curve (C) for combined Dataset-2 and Dataset-3. (D) Confusion matrix for combined Dataset-2 and Dataset-3.

## 4. DISCUSSION

Based on experiments and obtained classification metrics (e.g. average AUROC > 0.77 in all cases), the results showed that wav2vec embedding was effective in detecting the target class in both individual datasets and the leave-group-out classification scenario. Interestingly, although the leave-group-out model had a relatively high FPR and tended to label HC as diseased more often, it outperformed the random baseline for the leave-group-out scenario, further indicating the robustness of the presented methods in different evaluation settings and languages. To this extent, the multicentre study [17], conducted in different languages including Czech, English, German, French and Italian, found no evidence that language-related variations affect clinical parkinsonian phenotypes. This further supports the notion that the observed speech impairments in PD are consistent across diverse linguistic contexts.

It turned out that it was not possible to clearly determine whether the wav2vec-mean or the wav2vec-sum was better for the representation of 512 features. For Dataset-1, we observed better results using wav2vec-sum; for Dataset-2 and Dataset-3, wav2vec-mean worked better. Our explanation is that the measurement character of the signal is related to different behaviour. The Dataset-1 protocol follows a deterministic speech task, while for Dataset-2 and Dataset-3 it was free speech and read text. In addition to reporting the performance of the models, we presented a full pipeline that is easy to use and could help in creating speech recognition models for different languages and using them in a production environment.

The presented methods and the full pipeline were originally implemented for the purposes of the Biosignal Challenge 2020 and were awarded first place. The Biosignal Challenge 2020 was an international competition for students from the European Union and the United States of America, organised jointly by the Faculty of Electrical Engineering of the Czech Technical University in Prague and Humusoft (MathWorks). The aim of the Biosignal Challenge 2020 was to use the MATLAB computing environment to develop an algorithm for articulation rate estimation in human speech signals by detecting the number of syllables and measuring the duration of fluent speech, excluding all types of pauses in each utterance. 100 utterances from children were used as the training dataset. The best performing LASSO regression model tested with unseen data achieved a Pearson correlation coefficient = 0.95, mean absolute error = 3.8 and r2 = 0.91. Regression is yet another promising application area where we observed high accuracy of wav2vec.

One of the biggest advantages of wav2vec is that it is already pre-trained and can be applied directly to a small group of patients without any need to manually handcraft and engineer features. On the other hand, despite its high generalisability across speech recognition tasks and given that its corpus is not related to PD, wav2vec has been shown to be successfully applied to specific tasks – classification of HC vs. PD. The obtained results recommend further testing in clinical practise, e.g. wav2vec feature mapping and UPDRS. As can be seen from the clustering method in Fig. 2, wav2vec embedding can also be represented as an image, which can lead to an evaluation of its informative content and can be compared with other signal-image representations, such as the wavelet transform [39].

In quantitatively comparing the results presented with a relevant, previously published comparative study of speech analysis methods for predicting PD [40], Toye and Kompalli used the same datasets (Dataset-2 and Dataset-3). They computed MFCC and other acoustic feature sets and trained seven classification models (k-nearest neighbour, decision trees, support vector machines, naive Bayes, logistic regression, gradient boosting, random forests). The highest obtained results achieved an accuracy of 98% for the Italian Parkinson's Voice and Speech Database (Dataset-2) and > 80% accuracy for the KCL database (Dataset-3). In general, we can say that the current methods of speech recognition are very accurate, and we have shown that wav2vec achieves comparable accuracy (accuracy = 0.95 for Dataset-2 and 0.72 AUROC for Dataset-3 in the leave-group-out scenario). In addition, wav2vec has the presented unique benefits that make wav2vec stand out - i.e. seamless integration and generalisability. In comparing Dataset-1 with our previously published study [33], our main focus in this article has been on whether it is sufficient to use only one recording from the set of two recordings for each participant. Although we obtained slightly worse results (AUROC = 0.77 vs. avg AUROC = 0.88), we performed a more robust validation. Using the same validation method as in [33] and using two recordings, we achieved an average AUROC = 0.90, which is comparable to the best model presented there (AUROC = 0.92 for weighted k-NN and manually crafted features).

To address the influence of gender and age on speech performance in PD, it is important to consider the demographic composition of the datasets used in our study. Two recent phenotypic studies clearly showed a significant influence of gender [16] and age [41] on speech performance in PD. In addition, a finding from the article [14] revealed that speech impairments in early-stage PD were more pronounced in men than in women. In our study, Dataset-1 includes only data from male individuals with PD and age-matched male controls. For Dataset-2, it should be noted that the PD group shows an imbalance in terms of gender representation, consisting of 19 men and only 9 women. For Dataset-3, there is limited information available on gender and age demographics. Despite these considerations, we have made an effort to incorporate multiple datasets to ensure the generalisability of our findings.

The project has some limitations. We are aware that the quality of the recordings plays an essential role for studies with wav2vec and speech classification. In this case, the audio files used for training our models were free of noise. The significantly different results observed for the Italian dataset (Dataset-2) compared to the others can be attributed to a combination of factors. Firstly, variations in recording equipment, environmental settings and conditions across datasets can introduce variability in the acoustic signals measured. Secondly, differences in disease progression and severity among individuals in Dataset-2 compared to the other datasets may lead to distinct speech patterns more easily recognised by the classification algorithm. It is worth noting that, as mentioned earlier, the observed results are consistent with the results of a separate published study [40]. To increase the robustness of the models, it would be worth trying to augment the recordings with audio-specific augmentation techniques such as noise addition or volume control to perturb the models with more data and to test wav2vec, including its version 2.0, more thoroughly. Although the presented methods can significantly simplify speech classification pipelines for PD detection, a certain disadvantage is that the wav2vec features are not easy to interpret. It is also important to note that we did not perform any hyperparameter optimisation or other fine-tuning on our models. It should also be acknowledged that access to additional clinical data, such as UPDRS scores (available only for Dataset-2 and Dataset-3), The Montreal Cognitive Assessment (MoCA) assessments, disease duration and detailed medication status data were largely unavailable for the reported datasets, which is another limitation of this study.

## 5. Conclusion

In clinical medicine, a very common problem is the limited amount of annotated data. In this article, we implemented a wav2vec-based pipeline to obtain embeddings from raw waveforms and evaluate them on a classification task. The main advantage of the presented method thus confirmed its ability to use a large language corpus (English audio recordings) for application to a related problem in a specific domain, i.e. the analysis of speech signals in the biomedical domain for the detection of Parkinson's disease. The high AUROC scores and other performance metrics show that our models were able to accurately classify speech signals into the two categories (healthy vs. diseased). In this paper, we have shown the generalisability of the wav2vec features and the performance of these features in a cross-database scenario. We believe that our results can easily be extended to a wider range of neurological applications and other machine learning speech applications in biomedicine.

## Acknowledgment

Supported by the project of the National Institute for Neurological Research (Programme EXCELES, ID Project No. LX22NPO5107) - funded by the European Union – Next Generation EU. We also would like to acknowledge and thank T. Krajca for contributing to this work and the Grant no. SGS23/089/OHK4/1T/17 by the Grant Agency of the Czech Technical University in Prague, for supporting this work.

## References

[1] Hindle, J. V. (2010). Ageing, neurodegeneration and Parkinson's disease. *Age and Ageing*, 39 (2), 156-161. https://doi.org/10.1093/ageing/afp223

[2] Gorman, A. M. (2008). Neuronal cell death in neurodegenerative diseases: Recurring themes around protein handling. *Journal of Cellular and Molecular Medicine*, 12 (6a), 2263-2280. https://doi.org/10.1111%2Fj.1582-4934.2008.00402.x

[3] Damier, P., Hirsch, E. C., Agid, Y., Graybiel, A. M. (1999). The substantia nigra of the human brain. II. Patterns of loss of dopamine-containing neurons in Parkinson's disease. *Brain*, 122 (8), 1437-1448. https://doi.org/10.1093/brain/122.8.1437

[4] Reeve, A., Simcox, E., Turnbull, D. (2014). Ageing and Parkinson's disease: Why is advancing age the biggest risk factor? *Ageing Research Reviews*, 14, 19-30. https://doi.org/10.1016/j.arr.2014.01.004

[5] Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616 (7956), 259-265. https://doi.org/10.1038/s41586-023-05881-4

[6] Rusz, J., Hlavnička, J., Tykalová, T., Bušková, J., Ulmanová, O., Růžička, E., Šonka, K. (2016). Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder. *Sleep Medicine*, 19, 141-147. https://doi.org/10.1016/j.sleep.2015.07.030

[7] Tykalova, T., Rusz, J., Cmejla, R., Ruzickova, H., Ruzicka, E. (2014). Acoustic investigation of stress patterns in Parkinson's disease. *Journal of Voice*, 28 (1), 129.e1-129.e8. https://doi.org/10.1016/j.jvoice.2013.07.001

[8] Tjaden, K. (2008). Speech and swallowing in Parkinson's disease. *Topics in Geriatric Rehabilitation*, 24 (2), 115-126. https://doi.org/10.1097%2F01.TGR.0000318899.87690.44

[9] Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L., Farahmand, M. (2018). A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. *Biocybernetics and Biomedical Engineering*, 38 (1), 1-15. https://doi.org/10.1016/j.bbe.2017.09.002

[10] Novotny, M., Rusz, J., Cmejla, R., Ruzicka, E. (2014). Automatic evaluation of articulatory disorders in Parkinson's disease. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22 (9), 1366-1378. https://doi.org/10.1109/TASLP.2014.2329734

[11] Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D., Girardi, F. (2017). Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system. *IEEE Access*, 5, 22199-22208. https://doi.org/10.1109/ACCESS.2017.2762475

[12] Kwasny, D., Hemmerling, D. (2021). Gender and age estimation methods based on speech using deep neural networks. *Sensors*, 21 (14), 4785. https://doi.org/10.3390/s21144785

[13] Sánchez-Hevia, H. A., Gil-Pita, R., Utrilla-Manso, M., Rosa-Zurera, M. (2022). Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools and Applications*, 81 (3), 3535-3552. https://doi.org/10.1007/s11042-021-11614-4

[14] Jeancolas, L., Mangone, G., Petrovska-Delacrétaz, D., Benali, H., Benkelfat, B.-E., Arnulf, I., Corvol, J.-C., Vidailhet, M., Lehéricy, S. (2022). Voice characteristics from isolated rapid eye movement sleep behavior disorder to early Parkinson's disease. *Parkinsonism & Related Disorders*, 95, 86-91. https://doi.org/10.1016/j.parkreldis.2022.01.003

[15] Boersma, P., Weenink, D. *Praat: Doing phonetics by computer*. https://www.fon.hum.uva.nl/praat/

[16] Rusz, J., Tykalová, T., Novotný, M., Zogala, D., Růžička, E., Dušek, P. (2022). Automated speech analysis in early untreated Parkinson's disease: Relation to gender and dopaminergic transporter imaging. *European Journal of Neurology*, 29 (1), 81-90. https://doi.org/10.1111/ene.15099

[17] Rusz, J., Hlavnička, J., Novotný, M., Tykalová, T., Pelletier, A., Montplaisir, J., Gagnon, J.-F., Dušek, P., Galbiati, A., Marelli, S., Timm, P. C., Teigen, L. N., Janzen, A., Habibi, M., Stefani, A., Holzknecht, E., Seppi, K., Evangelista, E., Rassu, A. L., Dauvilliers, Y., Högl, B., Oertel, W., St. Louis, E. K., Ferini-Strambi, L., Růžička, E., Postuma, R. B., Šonka, K. (2021). Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson disease. *Annals of Neurology*, 90 (1), 62-75. https://doi.org/10.1002/ana.26085

[18] Rusz, J., Tykalova, T., Novotny, M., Zogala, D., Sonka, K., Ruzicka, E., Dusek, P. (2021). Defining speech subtypes in de novo Parkinson disease. *Neurology*, 97 (21), e2124-e2135. https://doi.org/10.1212/WNL.0000000000012878

[19] Issa, D., Fatih Demirci, M., Yazici, A. (2020) Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59, 101894. https://doi.org/10.1016/j.bspc.2020.101894

[20] Tuncer, T., Dogan, S., Acharya, U. R. (2020). Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels. *Biocybernetics and Biomedical Engineering*, 40 (1), 211-220. https://doi.org/10.1016/j.bbe.2019.05.006

[21] Karan, B., Sahu, S. S., Mahto, K. (2020). Parkinson disease prediction using intrinsic mode function based features from speech signal. *Biocybernetics and Biomedical Engineering*, 40 (1), 249-264. https://doi.org/10.1016/j.bbe.2019.05.005

[22] Solana-Lavalle, G., Galán-Hernández, J.-C., Rosas-Romero, R. (2020). Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. *Biocybernetics and Biomedical Engineering*, 40 (1), 505-516. https://doi.org/10.1016/j.bbe.2020.01.003

[23] Castro, C., Vargas-Viveros, E., Sánchez, A., Gutiérrez-López, E., Flores, D.-L. (2020). Parkinson's disease classification using artificial neural networks. In *VIII Latin American Conference on Biomedical Engineering and XLII National Conference on Biomedical Engineering*. Springer, IFMBE Proceedings 75, 1060-1065. https://doi.org/10.1007/978-3-030-30648-9_137

[24] Schneider, S., Baevski, A., Collobert, R., Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv*, https://arxiv.org/abs/1904.05862.

[25] Riviere, M., Joulin, A., Mazare, P.-E., Dupoux, E. (2020). Unsupervised pretraining transfers well across languages. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7414-7418. https://doi.org/10.1109/ICASSP40776.2020.9054548

[26] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*, https://arxiv.org/abs/1301.3781.

[27] Hannigan, G. D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Temesi, G., Hazuda, D. J., Woelk, C. H., Bitton, D. A. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, 47 (18), e110. https://doi.org/10.1093/nar/gkz654

[28] Baevski, A., Mohamed, A. (2020). Effectiveness of self-supervised pre-training for ASR. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7694-7698. https://doi.org/10.1109/ICASSP40776.2020.9054224

[29] Bayerl, S. P., Wagner, D., Baumann, I., Bocklet, T., Riedhammer, K. (2023). Detecting vocal fatigue with neural embeddings. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2023.01.012

[30] Švec, J., Polák, F., Bartoš, A., Zapletalová, M., Víta, M. (2022). Evaluation of Wav2Vec speech recognition for speakers with cognitive disorders. In *Text, Speech, and Dialogue: 25th International Conference (TSD 2022)*. Springer, LNAI 13502, 501-512. https://doi.org/10.1007/978-3-031-16270-1_41

[31] Hernandez, A., Pérez-Toro, P. A., Nöth, E., Orozco-Arroyave, J. R., Maier, A., Yang, S. H. (2022). Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. *arXiv*, https://arxiv.org/abs/2204.01670.

[32] Escobar-Grisales, D., Ríos-Urrego, C. D., Orozco-Arroyave, J. R. (2023). Deep learning and artificial intelligence applied to model speech and language in Parkinson's disease. *Diagnostics*, 13 (13), 2163. https://doi.org/10.3390/diagnostics13132163

[33] Klempir, O., Krupicka, R. (2018). Machine learning using speech utterances for Parkinson disease detection. *Lékař a Technika / Clinician and Technology*, 48 (2), 66-71. https://ojs.cvut.cz/ojs/index.php/CTJ/article/view/4881

[34] Jaeger, H., Trivedi, D., Stadtschnitzer, M. (2019). Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls. *Zenodo*, https://zenodo.org/doi/10.5281/zenodo.2867215.

[35] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng. N., Grangier, D., Fairseq, M. A. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*. https://github.com/facebookresearch/fairseq

[36] *wav2vec large.pt* https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_large.pt

[37] Lehecka, J., Svec, J., Prazak, A., Psutka, J. V. (2022). Exploring capabilities of monolingual audio transformers using large datasets in automatic speech recognition of Czech. *arXiv*, https://arxiv.org/abs/2206.07627.

[38] Klempíř, O. (2023). Evaluating the performance of wav2vec embedding for Parkinson's disease detection. *GitLab*, https://gitlab.fel.cvut.cz/klempond/wav2vec-embedding-for-pd-detection/.

[39] Klempíř, O., Krupička, R., Bakštein, E., Jech, R. (2019). Identification of microrecording artifacts with wavelet analysis and convolutional neural network: An image recognition approach. *Measurement Science Review*, 19 (5), 222-231. https://doi.org/10.2478/msr-2019-0029

[40] Toye, A. A., Kompalli, S. (2021). Comparative study of speech analysis methods to predict Parkinson's disease. *arXiv*, https://arxiv.org/abs/2111.10207.

[41] Rusz, J., Tykalová, T., Novotný, M., Růžička, E., Dušek, P. (2021). Distinct patterns of speech disorder in early-onset and late-onset de-novo Parkinson's disease. *npj Parkinson's Disease* 7, 98. https://doi.org/10.1038/s41531-021-00243-1

Received June 1, 2023
Accepted October 17, 2023