On Estimating Age and Gender from Parkinson's Disease Diagnostic-Oriented Recordings Using Wav2Vec 2.0

Ondrej Klempir✉, Ales Tichopad, Radim Krupicka

Faculty of Biomedical Engineering, Czech Technical University in Prague, Sitna Square 3105, Kladno, Czech Republic

✉ *Corresponding author: klempond@fbmi.cvut.cz*

Abstract

Can self-supervised speech foundation models (SFMs) be used for automatic patient metadata extraction, even when no prior demographic information is available and speech is affected by pathology? SFMs show strong cross-task generalization, yet it remains unclear to what extent demographic attributes such as age and gender are intrinsically encoded, particularly in pathological speech. This study evaluated the capability of a pretrained SFM Wav2Vec 2.0 to estimate age and gender across healthy controls (HC), Parkinson's disease (PD) subjects, and related parkinsonian syndromes (multiple system atrophy, progressive supranuclear palsy), without exposing the model to any data from the evaluated datasets. A frozen, publicly available Wav2Vec 2.0 model was used to extract speech representations from three independent multilingual datasets. No machine learning model was trained or fine-tuned on the target data. The analysis solely assessed information already present in the pretrained embeddings. Multiple speech tasks (read text, diadochokinesis, sustained vowels) and diagnostic groups were evaluated using gender accuracy, correlations with true age, chi-square tests, and group-level analyses. Gender estimation achieved consistently high accuracy (min. 94%, up to 100%) across datasets and tasks. Age estimation showed significant correlations with true age for read speech, including PD speakers. Analyses of vowel data demonstrated preserved gender distributions but systematic age bias across all diagnostic groups. Without task-specific training, pretrained Wav2Vec 2.0 embeddings robustly encode gender and preserve age-related structure in connected, including pathological, speech, whereas age estimation from isolated vowel phonation remains unreliable. These findings highlight the demographic robustness of SFMs and the need to monitor age-related bias in clinical applications.

Keywords: age, gender, wav2vec 2.0, Parkinson's disease, speech, voice

1. Introduction

Self-supervised foundation models have been rapidly reshaping how healthcare can leverage unstructured data. In speech research, these models are capable of learning powerful acoustic and prosodic representations directly from large and heterogeneous audio collections, enabling downstream systems to detect clinical signals that previously required handcrafted features or task-specific modeling. Speech foundation models (SFMs), such as Wav2Vec 2.0 or HuBERT,

have shown strong performance in Parkinson's disease (PD) and other voice disorders detection and can outperform traditional acoustic feature sets [1, 2]. PD is the second most common neurodegenerative disorder after Alzheimer's disease (AD). It primarily affects individuals aged 65 years and older [3]. PD cases are expected to double by 2030 (since 2005) and continue a sharp growth trajectory into the following decade [4].

The recent discussion and corresponding study on the usefulness of speech data for developing classifiers for PD detection [5], demonstrated the benefits of balancing participants' gender and disease-status distributions. However, the corpora obtained either from a publicly or internally available source may lack reliable demographic metadata. This might pose a problem, as age and gender can affect baseline acoustic characteristics and may interact with disease-related changes. When such information is missing, models may theoretically exploit correlations with demographics instead of pathology, especially given that PD cohorts are typically older while healthy controls (HC) are often younger. Without addressing these imbalances, reported accuracy may reflect demographic confounds rather than true clinical discriminability. These concerns are amplified by the behavior of recent foundation models. Being trained on large, heterogeneous speech corpora, they naturally cover demographic structure. When applied to imbalanced datasets as feature extractors, downstream classifiers may rely on age or gender cues rather than PD-specific voice patterns. This shortcut learning can artificially inflate performance, particularly in designed binary contrasting (HC vs. PD). Therefore, in case when demographic variables are not available, at least coarse estimation or characterization might be of interest to identify potential confounders, support external validation, and enable transparent reporting.

Speech-based PD detection can be influenced by speaker gender, age, language, recording device, task characteristics and other factors. Evidence showed that sex (biological attributes) affects disease-related voice changes, with males often exhibiting greater impairments [6]. Recording conditions and the choice of diagnostic-oriented tasks can substantially alter the acoustic footprint and downstream performance. To address this partially, several robust, language-independent acoustic features relevant to PD detection have already been identified [7]. However, all of these factors (and not limited to) underscore the need to understand how they possibly interact with foundation model behavior, with age and gender standing out for considering possible demographics differences.

Foundation models can generally predict speaker gender and age with high accuracy in healthy speech [8], which is expected for clean, normative recordings. At least a few pre-trained, openly available models hosted on HuggingFace, e.g. [8, 9, 10], report near-perfect gender-prediction accuracy and also possibly support age estimation. These models require no additional training and are immediately usable, making them candidates for demographic characterization. However, for disordered speech evidence remains limited. Although many PD assessment protocols use standardized tasks (e.g. fixed reading passages, monologues, diadochokinetic tasks, sustained vowels), little is known about how well foundation models infer demographic attributes across these task types when pathology is present. Our goal is to evaluate whether such "ready-to-go," multipurpose models remain reliable when applied to PD-related speech.

In our prior work [11], we used Wav2Vec 1.0 embeddings to predict age in an Italian dataset that included PD participants during a standardized reading task. The model captured age structure with reasonable accuracy, effectively separating younger from older speakers despite pathological variation. While gender can often be inferred perceptually, foundation models offer an opportunity to objectivize this process, reduce listener bias, and support reproducible demographic characterization. To provide an example, recent studies have reported the gender distribution of the Mobile Device Voice Recordings at King's College London (MDVR-KCL) English corpus [12, 13], despite the dataset lacking official demographic metadata. To our knowledge, these studies did not describe how gender was inferred, and it is anticipated that classifications were made by listening. These assumptions highlight the need for systematic, model-based demographic estimation in PD speech datasets.

Existing relevant research found that demographic cues remain detectable even in pathological speech. The study by Gómez-García et al. [14], demonstrated that age classification (adult vs. older adult) is feasible using sustained vowels, showing that age structure can be recovered even in disordered voices and suggesting a path toward age-aware pathology models. For gender, the “Octopus” vowel-based PD and gender recognition study [15] reported an average accuracy of 93.8% across six classifiers using a vowel dataset (three vowels per person, 252 speakers). More recently, GeHirNet [16] proposed a two-stage, gender-aware pathology classifier trained on multiple pathological speech corpora, including the gender-balanced PC-GITA PD dataset. Although the method achieved high accuracy (97.6%) by modeling gender cues in Stage 1 and pathology in Stage 2, it requires end-to-end training, focuses on pathology detection rather than standalone demographic estimation, and is limited to sustained vowel /a/. To our knowledge, no study directly and comprehensively evaluated the capability to predict age and gender in PD multi-lingual corpora.

From a broader perspective, data-driven phenotyping in neurodegeneration often requires signals that are not easily recoverable from conventional sources such as billing records, medication history, or diagnostic coding alone. Many clinically relevant attributes, disease severity, cognitive status, motor symptoms, comorbid conditions, age-related decline, or progression trajectories, are either under-reported, delayed in documentation, or not represented at all in administrative datasets. This motivates the use of alternative modalities such as speech, which may provide auxiliary structured information and capture subtle but biologically meaningful variation beyond what is typically available in routinely collected clinical data.

Herein, we investigate how accurately a state-of-the-art foundation speech model can estimate gender and age from voice and speech diagnostic-oriented recordings. We evaluate the performance across three corpora that differ in language, task design, and the presence of PD, parkinsonism or related disorders.

<u>What is already known on this topic and What this communication adds</u>

While gender and age prediction works reliably in healthy speech, it remains unclear whether this capability extends to various types of pathological recordings. Clinical assessment of

speech disorders typically involves standardized tasks, yet many recordings or entire datasets lack demographic annotation. Our aim is therefore to assess whether a single, ready-to-use foundation model can provide consistent demographic estimates across heterogeneous PD-related speech material, supporting clearer dataset characterization and more reliable downstream model evaluation.

2. Methods

2.1. Datasets

PC-GITA Dataset-1. The PC-GITA dataset [17] consists of speech recordings from 50 HC subjects and 50 individuals with PD, all Spanish speakers from Colombia. The dataset is fully balanced with respect to age and gender. For our analysis, we used the recordings of read text, the diadochokinetic task “pataka”, and the sustained vowel /a/. Monologue recordings were excluded mostly due to its close similarity with read text.

Italian Dataset-2. The Italian dataset [18] includes recordings from three groups: Young HC (YHC), Elderly HC (EHC), and PD patients. The original dataset contains 15 YHC participants, 22 EHC participants, and 28 PD participants. For our analysis, we focused on the read text recordings. For one EHC subject, the read-text sample could not be located (neither the B1-prefixed nor the B2-prefixed files were available); therefore, we used that subject’s word-list recording (prefix PR) instead. For an additional four EHC subjects, we were unable to reliably match their recordings to the corresponding age and gender metadata, and these samples were consequently excluded from the analysis.

PD and parkinsonism Dataset-3. To demonstrate possible generalization of the presented solution beyond PD, we employed the dataset described in [19]. This dataset contains synthesized replicas of sustained vowels produced by 22 patients with PD, 21 patients with multiple system atrophy (MSA), 18 patients with progressive supranuclear palsy (PSP), and 22 HC. For consistency, we restricted our analysis to a single copy of the prolonged vowel /A/.

2.2. Pre-trained speech model

To estimate speaker age and gender, we employed the 24-layer Wav2Vec 2.0 model for age and gender recognition [8]. The model takes raw audio waveforms as input and produces continuous age predictions (normalized between 0 and 1, approximately corresponding to 0–100 years) and gender probabilities across three classes: child, female, and male.

This model was obtained by fine-tuning Wav2Vec2-Large-Robust, originally pretrained on 16 kHz speech audio, using several publicly available corpora, including aGender, Mozilla Common Voice, TIMIT, and VoxCeleb2. We accessed the implementation via Hugging Face [20]. Given the composition of the pretraining corpora, it is reasonable to assume that the model was not exposed to any neurological-disorder speech datasets during training. For gender inference, we selected the class with the highest predicted probability. All analyses were

performed offline with a reasonable CPU runtime, requiring approximately one hour to process a single corpus and no specialized GPU hardware.

2.3. Age and gender prediction

Each speech recording was paired with the participant's chronological age "true age" and the model-estimated age "predicted age", with corresponding gender attributes following the same principle. To facilitate comparison of the overall age distributions, both true and predicted ages were discretized into 10-year bins: <30, 30–39, 40–49, 50–59, 60–69, 70–79, and ≥80 years. The prediction performance was evaluated using label agreement between the participant's true value and the model-predicted value. Because no predictive model was trained (i.e., predictions were obtained directly from the pretrained model without additional optimization step), no train/test data split was required.

2.4. Statistical methods

Beyond reporting simple accuracy metrics, the primary objective of the statistical analysis was to determine whether the age-prediction performance in individuals with PD was comparable to that observed in HC. Demonstrating equivalence or non-inferiority of predictive performance in PD would support the applicability of the model in clinical or research settings involving this population. For all statistical tests, a two-sided p-value < 0.05 was considered indicative of statistical significance.

For each defined group, we compared the predicted and true age distributions across the predefined 10-year bins. The Chi-square statistic quantified the deviation between observed (true) and model-estimated bin frequencies, treating the true age distribution as the reference. This test evaluated whether the model's predictions preserved the overall age structure within each group. In addition, to assess rank-order agreement and overall predictive validity at the individual level, we computed Spearman's rank correlation coefficient with its corresponding p-value between true and predicted continuous ages.

Agreement between true and predicted age was further evaluated using Bland–Altman plots, reporting mean bias and 95% limits of agreement, more details can be found in Supplementary Methods.

3. Results

3.1. Dataset-1

The achieved results of gender and age estimation for the Spanish PC-GITA dataset are summarized in Table 1. Gender classification accuracy reached consistently high values (min. 94%) across all speech tasks (read text, pataka, sustained vowel /a/) and for both studied groups, i.e., HC and PD. Specifically, for the read text task, perfect classification accuracy was achieved for the HC group (100%), while accuracy reached 98% for the PD group.

Regarding age estimation, the principal findings indicate statistically significant correlations between the predicted and true age for the read text task in both groups. For HC, a moderate positive correlation was observed (Spearman's $\rho = 0.52$, $p < 0.001$), while for PD a slightly weaker but still significant correlation was obtained ($\rho = 0.44$, $p = 0.001$). These results demonstrate a statistically significant ability of the foundation model to estimate age, even in speakers affected by PD. In addition, chi-square goodness-of-fit tests were applied to compare predicted and true age distributions. In the majority of cases, the tests indicated statistically significant differences between the distributions. Only one non-significant result was observed, namely for the read text task in the PD group ($\chi^2 = 5.58$, $p = 0.13$), suggesting no substantial violation of the estimated age distribution in this group.

To further assess the agreement between predicted and true ages at both the individual and category levels, scatter plots (Fig. 1) and boxplots (Fig. 2) were generated for the HC group. Corresponding visualizations for the PD group are provided in Fig. 3 and Fig. 4. These figures reveal that while overall age-related trends were reasonably well captured for the read text and pataka tasks, the sustained vowel task failed to reflect the true age distribution. In this task, a pronounced underestimation of age was observed for both HC and PD groups, indicating a clear failure of the model in capturing age-related information from vowel phonation alone. Age-prediction agreement in the HC and PD cohorts was further examined using Bland–Altman analysis (Supplementary Figs. S1 and S2).

Table 1. Gender classification accuracy and age estimation results for the Spanish PC-GITA dataset across speech tasks and participant groups. Age was discretized into categories {<50, 50–59, 60–69, 70+}. Since the <30, 80–89, and 90+ age bins contained no observed samples, they were merged with adjacent categories to avoid zero expected cell counts in the chi-square tests.

Group	Task	Accuracy for Gender Estimation	Correlation with True Age (p-value)	Chi² Goodness of Fit (p-value)	MAD* years
HC (N = 50)	read text	1.00	0.52 **(p < 0.001)**	14.12 **(p = 0.003)**	5.0
	pataka	0.96	0.36 **(p = 0.01)**	16.60 **(p < 0.001)**	9.0
	/a/ vowel	0.94	0.05 (p = 0.72)	575.00 **(p < 0.001)**	27.0
PD (N = 50)	read text	0.98	0.44 **(p = 0.001)**	5.58 (p = 0.13)	6.0
	pataka	0.94	0.20 (p = 0.17)	75.63 **(p < 0.001)**	9.5
	/a/ vowel	0.98	0.23 (p = 0.10)	392.30 **(p < 0.001)**	25.0

***MAD (Median Absolute Deviation)**

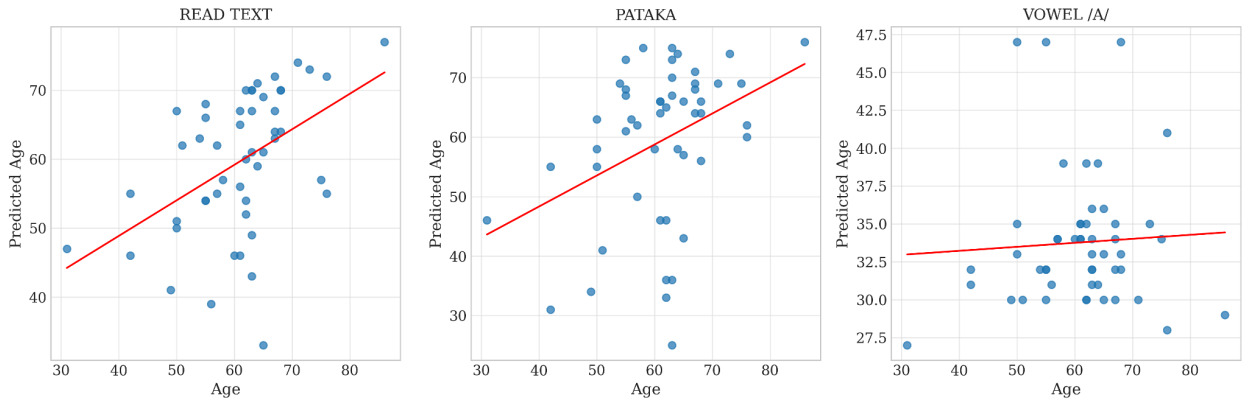

Fig. 1. Scatter plots showing the relationship between true age and predicted age for the HC group across the three speech tasks: read text, pataka, and sustained vowel (/a/).

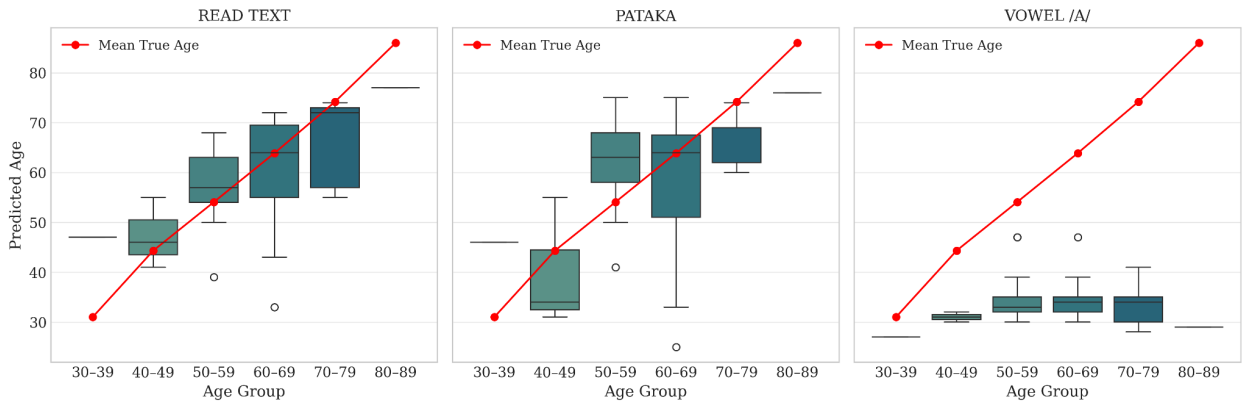

Fig. 2. Boxplots of predicted age across true age groups for the HC group. The mean true age for each group is overlaid as a reference marker.

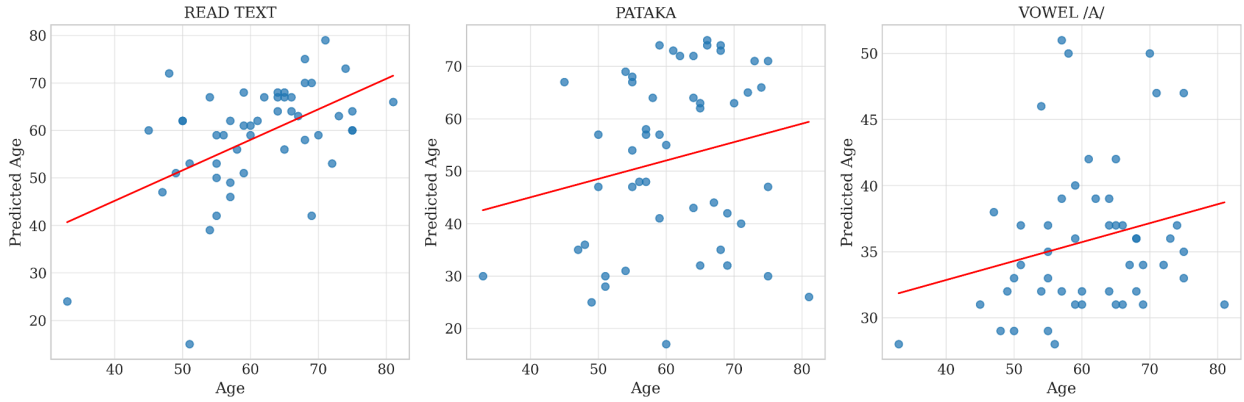

Fig. 3. Scatter plots showing the relationship between true age and predicted age for the PD group across the three speech tasks: read text, pataka, and sustained vowel (/a/).

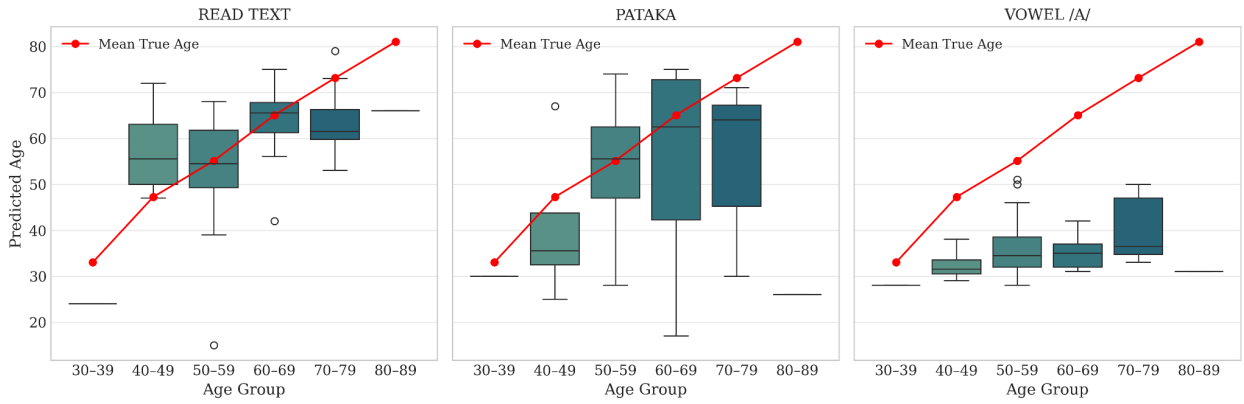

Fig. 4. Boxplots of predicted age across true age groups for the PD group. The mean true age for each group is overlaid as a reference marker.

3.2. Dataset-2

The results for the Italian dataset are summarized in Table 2. Compared to the Spanish PC-GITA dataset, this dataset is characterized by a smaller sample size and a different group structure, comprising YHC, EHC, and PD. The analysis was limited to the read text task. Gender estimation achieved perfect classification accuracy (100%) across all three groups.

With respect to age estimation, moderate positive correlations between predicted and true age were observed, although statistical significance varied across groups. For the EHC group, the correlation reached statistical significance (Spearman's $\rho = 0.49$, $p = 0.04$), whereas for YHC ($\rho = 0.39$, $p = 0.15$) and PD ($\rho = 0.28$, $p = 0.15$) the correlations did not reach significance. Chi-square goodness-of-fit testing was performed only for the PD group, where a statistically significant difference between predicted and true age distributions was observed ($\chi^2 = 26.27$, $p < 0.001$). For the YHC and EHC groups, chi-square tests were not reported due to insufficient cell counts required to satisfy test assumptions. To visually assess model performance, scatter plots illustrating the relationship between true and predicted age for the read text task are presented in Fig. 5 for YHC, EHC, and PD groups. Complementary boxplots are shown in Fig. 6. Age-prediction bias across YHC, EHC, and PD groups was assessed using Bland–Altman analysis (Supplementary Fig. S3).

Table 2. Gender classification accuracy and age estimation results for the Italian dataset (read text task only) across YHC, EHC, and patients with PD.

Group	Task	Accuracy for Gender Estimation	Correlation with True Age (p-value)	Chi² Goodness of Fit (p-value)	MAD* years
YHC (N = 15)	read text	1.00	0.39 (p = 0.15)	Not reported	13.0
EHC (N = 18)	read text	1.00	0.49 **(p = 0.04)**	Not reported	6.0
PD (N = 28)	read text	1.00	0.28 (p = 0.15)	26.27 **(p < 0.001)**	5.0

***MAD (Median Absolute Deviation)**

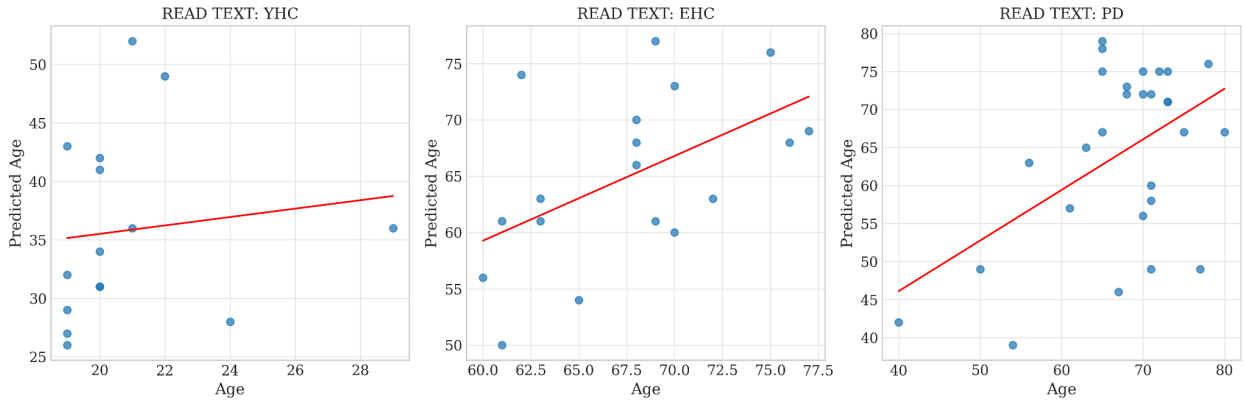

Fig. 5. Scatter plots illustrating the relationship between true age and predicted age for the read text task in the Italian dataset.

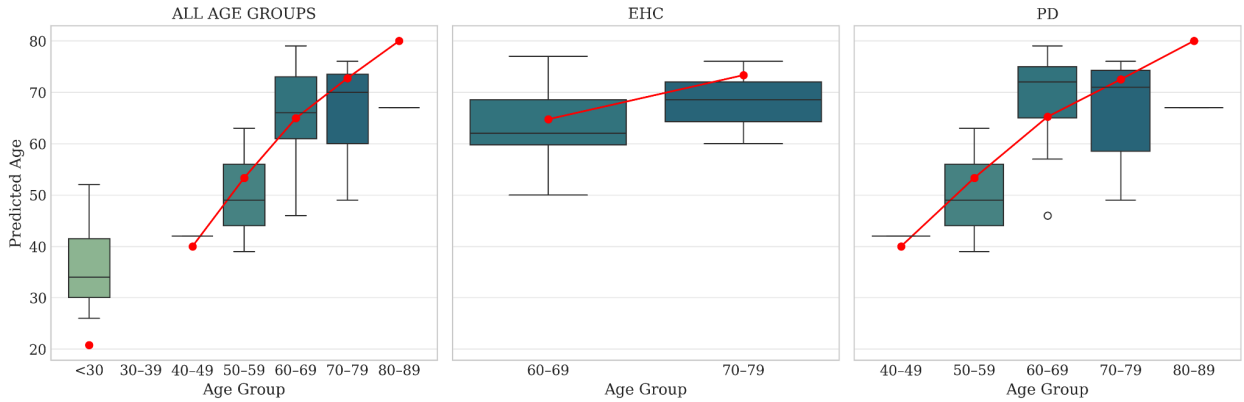

Fig. 6. Boxplots of predicted age for the Italian dataset based on the read text task. The leftmost plot shows results across all age groups, followed by separate visualizations for the EHC and PD. The YHC group is not shown due to its narrow age range. Notably, the 50–79 age interval was well captured by the model (typical for PD onset).

3.3. Dataset-3

To assess the potential generalization of the proposed approach beyond PD, we evaluated its performance on Dataset-3, which included synthesized sustained vowel productions from patients with PD, MSA, PSP and HC. Unlike Datasets 1 and 2, individual-level ground-truth annotations were not available for this dataset. Therefore, the evaluation was performed at the group level, and the reported metrics followed a different format.

Gender and age estimation performances across diagnostic groups are summarized in Table 3. A chi-square goodness-of-fit test comparing true and predicted gender distributions across all groups revealed no statistically significant differences ($\chi^2 = 2.2$, df = 7, $p = 0.95$), indicating that the predicted gender proportions matched the true distributions at the group level. In contrast, age estimation showed a pronounced systematic underestimation across all diagnostic categories. While the true average ages ranged between approximately 61 and 67 years, the

predicted average ages were substantially lower, spanning from 29.8 to 36.1 years. This underestimation was consistent across HC, PD, MSA, and PSP groups.

Table 3. Group-level comparison of true and predicted gender counts and average age estimates, including HC, PD, MSA, and PSP. Age values are reported as group means. Due to the absence of individual-level ground-truth annotations, evaluation was performed at the group level only.

Estimated	F (true/predicted)	M (true/predicted)	avg. age (true/predicted)
HC (N = 22)	11 / 11	11 / 11	63.6 / 29.8
MSA (N = 21)	12 / 16	9 / 5	61.0 / 36.1
PD (N = 22)	12 / 14	10 / 8	64.4 / 34.2
PSP (N = 18)	6 / 5	12 / 13	66.7 / 32.2

4. Discussion

True value of the foundational models lies in understanding how existing pretrained models behave when applied to clinically relevant problems where prior evidence is limited. In this context, the present study should be viewed primarily as a validation study in a domain with possibly missing or limited knowledge, rather than as the development of a novel algorithm.

In this study, we systematically evaluated the pretrained speech foundation model, Wav2Vec 2.0, across multiple datasets, speech tasks, and neurological conditions, including PD and related parkinsonian syndromes. This design enabled a realistic assessment of what can (and cannot) be expected from such a model when deployed without task-specific training. Our findings demonstrated near-perfect performance in gender estimation across all datasets and speech tasks, confirming that gender-related information is robustly encoded in pretrained speech representations. While age estimation proved more challenging, the model consistently captured meaningful age-related structure, allowing reliable characterization of datasets and disease-specific age distributions in both HC and patients.

Importantly, unlike our previous work on age prediction using Wav2Vec-based models [11], no training or fine-tuning was performed in this study. We deliberately limited comparisons with task-specific methods for age and gender estimation, as our objective was not to maximize predictive performance but to assess the intrinsic capability of a general pretrained model. This constraint can be viewed as a methodological strength, i.e. by preventing any model from “seeing” the evaluated data during training, we reduce the risk of overfitting and provide a clearer picture of true model generalization. Moreover, using the model documented in [20], both age and gender can be estimated in a single execution, eliminating the need to deploy separate task-specific models. According to the model card documentation, it is unlikely that the evaluated foundation model was exposed to dysarthric speech during pretraining, yet it still achieved reasonable accuracy within the PD subgroup. This observation suggests that

task-specific fine-tuning may not be strictly required for reliable gender estimation in pathological speech. Based on the results observed for Datasets 1 and 2, a broader representation of older speakers (typically 80+ years; undetected) in pretraining data may further contribute to improved age prediction performance.

Quantitatively, analyses on Datasets 1 and 2 did not reveal a systematically higher age prediction error in the patient subgroups. In PD speakers from the PC-GITA dataset, the MAD was 6.0 years, compared to 5.0 years in HC. Similarly, in the Italian dataset, the MAD was 6.0 years for EHC and 5.0 years for PD subjects. MAD is a commonly used metric for evaluating age prediction from voice acoustics, e.g. [21]. In [21], Novotny et al., focused on predicting children's age and achieved precise estimates even using sustained vowels, albeit with specialized algorithms, not foundation models.

As observed in Datasets 1 and 3, sustained vowel phonation (/a/) consistently led to substantial underestimation of age. Importantly, this bias was not only specific to HC and PD (Table 1; HC MAD = 27.0; PD MAD = 25.0; Bland–Altman Supplementary Figs. S1 and S2) but was also present in atypical parkinsonian syndromes (Table 3), indicating a generally limited age-discriminative capacity of the model when restricted to isolated vowel production. Nevertheless, the accurate preservation of gender distributions across all diagnostic categories suggests that certain speaker characteristics remain robustly encoded even under these constrained conditions. The systematic failure of age estimation from sustained vowels may have implications for downstream representation learning. Specifically, the weak age sensitivity observed for vowel-only recordings may be advantageous when using Wav2Vec-based embeddings for classification tasks, as it suggests that age-related variation plays a limited role and is less likely to dominate the learned representations. In contrast, given the consistent and robust performance observed for gender estimation across all tasks and datasets, the use of gender-balanced datasets appears advisable to prevent unintended bias propagation in speech-based modeling pipelines.

We acknowledge that gender recognition from speech might raise important ethical considerations. In this study, our objective is not to infer or enforce personal identity attributes, but rather to distinguish subject subgroups in neurodisordered datasets where systematic differences in speech production are known to exist. Prior research has consistently demonstrated gender-related differences in speech characteristics, including in neurological populations [6, 22, 23]. Capturing such variability is essential for correctly characterizing datasets and avoiding confounding effects in downstream analyses. At the same time, we recognize that some individuals may not wish to disclose or be categorized by gender, which further motivates the need for transparent and responsible use of such annotations. In this context, age and gender stratification strategies, such as training on one subgroup and testing on another, are particularly relevant for studies employing speech machine learning models for condition detection. Such stratified analyses can help disentangle disease-related speech markers from demographic effects, improve interpretability, and provide a more rigorous assessment of model generalization across populations.

Beyond controlled diagnostic-oriented datasets, automated estimation of age and gender may also be valuable in scenarios where demographic information is completely missing or poorly documented, such as speech samples collected from publicly available sources (e.g., online videos or interviews). Recent studies, including ParkCeleb [24] and ADCeleb [25], illustrated a promising trend toward leveraging publicly available recordings of individuals who have disclosed neurological conditions, enabling large-scale and, in some cases, longitudinal analyses of speech. As pretraining datasets continue to grow in size and diversity, the use of such data is expected to increase, and existing foundation models may play an important role in the automated annotation and initial characterization of these samples. An additional practical benefit arises in studies where fully balanced datasets are not available. In such cases, reporting other “baseline” performance metrics, beyond straightforward sample-size baselines, can provide valuable context.

A pretrained model can be useful not only in cases where age and gender information is missing. Having access to a general-purpose, large pretrained model for age and gender estimation that performs reasonably well in PD and other voice or speech disorders offers advantages even when demographic metadata are available. In such settings, the model can serve as an independent reference for identifying potential data quality issues. For example, during the development of speech data pipelines, recordings may be incorrectly paired, or metadata may be inconsistently assigned to individual samples. Discrepancies between predicted and recorded demographic attributes can thus function as a form of automated self-quality check, highlighting samples that warrant closer inspection.

A known limitation of the current approach is that, when the goal is to achieve the highest possible predictive accuracy, particularly for age estimation, performance is likely to be substantially lower than what could be achieved with task-specific fine-tuning. In such cases, fine-tuning the pretrained model on disorder-specific data, for example PD or other dysarthric speech, would be necessary to improve age prediction accuracy. While our study demonstrated the intrinsic capability of a dataset-agnostic foundation model, we believe that fine-tuning remains the most effective strategy for applications that demand precise demographic predictions.

Future work should investigate multimodal phenotyping strategies combining speech-derived features with records of clinical care or administrative health data. For conditions such as PD or AD, reliably distinguishing demographic structure is only a first step; more complex voice-based signatures of disease severity, motor impairment, or cognitive decline could help address known limitations of structured data sources, where staging and functional trajectories are difficult to recover at scale.

5. Conclusion

SFMs, such as Wav2Vec 2.0, have already demonstrated strong utility in representation learning, classification, and automatic speech recognition for PD. Here, we extend their application to the annotation of PD speech recordings for demographic estimation, specifically

gender and age. Our results showed that gender can be predicted with excellent accuracy across healthy and pathological speakers. Age estimation was less precise, particularly for older individuals, but the model still captures meaningful structure, allowing characterization of datasets in both HC and, importantly, PD patients.

These findings suggest broader potential applications of general pretrained speech models. They can support quality control of recordings in data pipelines for voice disorder research, detect mislabeled samples, and highlight inconsistencies within datasets, serving as exploratory tools without requiring task-specific training. While fine-tuning the model on the target data could improve age estimation, the current approach demonstrates that large pretrained models already provide valuable insights and can characterize datasets reliably with zero/minimal exposure to training data.

Contributions

Study concept and design: OK, RK, AT. Acquisition, analysis, or interpretation of data: OK, RK. Drafting of the manuscript: OK. Critical revision of the manuscript for important intellectual content: all authors. Study supervision: RK. Funding: RK, AT. All authors accept responsibility to submit the manuscript for publication.

Acknowledgement

This research was supported by the Czech Health Research Council grant no. NW24-04-00259. This research was partially supported by the project CZ.02.01.01/00/23_025/0008743, funded by the European Union under the Operational Programme Johannes Amos Comenius (OP JAK).

Competing interests

The authors declare no competing interests.

References

[1] J. Cai, Y. Song, J. Wu, X. Chen, Voice Disorder Classification Using Wav2vec 2.0 Feature Extraction, Journal Of Voice. (2024). https://doi.org/10.1016/j.jvoice.2024.09.002.

[2] A. Favaro, Y.-T. Tsai, A. Butala, T. Thebaud, J. Villalba, N. Dehak, L. Moro-Velázquez, Interpretable speech features vs. DNN embeddings: What to use in the automatic assessment of Parkinson's disease in multi-lingual scenarios, Computers In Biology And Medicine. 166 (2023) 107559. https://doi.org/10.1016/j.compbiomed.2023.107559.

[3] M.E. Morris, Movement Disorders in People With Parkinson Disease: A Model for Physical Therapy, Physical Therapy. 80 (2000) 578-597. https://doi.org/10.1093/ptj/80.6.578.

[4] R. Bhidayasiri, L.V. Kalia, B.R. Bloem, Tackling Parkinson's Disease as a Global Challenge, Journal Of Parkinson's Disease. 13 (2023) 1277-1280. https://doi.org/10.3233/jpd-239005.

[5] T.Y. Zhong, E. Janse, C. Tejedor-Garcia, L. ten Bosch, M. Larson, Evaluating the Usefulness of Non-Diagnostic Speech Data for Developing Parkinson's Disease Classifiers, in: Interspeech 2025, ISCA, ISCA, 2025: pp. 3738-3742. https://doi.org/10.21437/interspeech.2025-1805.

[6] F. Cao, A.P. Vogel, P. Gharahkhani, M.E. Renteria, Speech and language biomarkers for Parkinson's disease prediction, early diagnosis and progression, Npj Parkinson's Disease. 11 (2025). https://doi.org/10.1038/s41531-025-00913-4.
[7] S. Scimeca, F. Amato, G. Olmo, F. Asci, A. Suppa, G. Costantini, G. Saggio, Robust and language-independent acoustic features in Parkinson's disease, Frontiers In Neurology. 14 (2023). https://doi.org/10.3389/fneur.2023.1198058.
[8] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, Björn W. Schuller, Speech-based Age and Gender Prediction with Transformers, Arxiv (Cornell University). (2023). https://doi.org/10.48550/arxiv.2306.16962.
[9] Common-Voice-Gender-Detection, Hugging Face. https://huggingface.co/prithivMLmods/Common-Voice-Gender-Detection (accessed December 19, 2025).
[10] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, in: Interspeech 2020, ISCA, ISCA, 2020. https://doi.org/10.21437/interspeech.2020-2650.
[11] O. Klempíř, R. Krupička, Analyzing Wav2Vec 1.0 Embeddings for Cross-Database Parkinson's Disease Detection and Speech Features Extraction, Sensors. 24 (2024) 5520. https://doi.org/10.3390/s24175520.
[12] S.A. Jones, J. Cosgrove, H. Wang, R.K. Mathew, Automated Detection of Speech Disorders in Parkinson's Disease using Deep Convolutional Neural Networks: A Pilot Study, (2025). https://doi.org/10.1101/2025.07.18.25331737.
[13] K.H. Kim, B.-J. Lee, H.-W. Koo, Feasibility Study of Parkinson's Speech Disorder Evaluation With Pre-Trained Deep Learning Model for Speech-to-Text Analysis, Korean Journal Of Neurotrauma. 20 (2024) 168. https://doi.org/10.13004/kjnt.2024.20.e30.
[14] J.-A. Gómez-García, L. Moro-Velázquez, J.I. Godino-Llorente, G. Castellanos-Domínguez, Automatic age detection in normal and pathological voice, in: Interspeech 2015, ISCA, ISCA, 2015: pp. 3739-3743. https://doi.org/10.21437/interspeech.2015-742.
[15] T. Tuncer, S. Dogan, A novel octopus based Parkinson's disease and gender recognition method using vowels, Applied Acoustics. 155 (2019) 75-83. https://doi.org/10.1016/j.apacoust.2019.05.019.
[16] K. Zhao, E. Fleisch, F. Barata, GeHirNet: A Gender-Aware Hierarchical Model for Voice Pathology Classification, Arxiv (Cornell University). (2025). https://doi.org/10.48550/arxiv.2508.01172.
[17] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. González-Rátiva, E. Nöth, New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 342–347, Reykjavik, Iceland. European Language Resources Association (ELRA). https://aclanthology.org/L14-1549/
[18] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, F. Girardi, Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System, Ieee Access. 5 (2017) 22199-22208. https://doi.org/10.1109/access.2017.2762475.
[19] J. Hlavnicka, R. Cmejla, J. Klempir, E. Ruzicka, J. Rusz, Acoustic Tracking of Pitch, Modal, and Subharmonic Vibrations of Vocal Folds in Parkinson's Disease and Parkinsonism, Ieee Access. 7 (2019) 150339-150354. https://doi.org/10.1109/access.2019.2945874.

[20] Model for Age and Gender Recognition based on Wav2vec 2.0 (24 layers), Hugging Face. https://huggingface.co/audeering/wav2vec2-large-robust-24-ft-age-gender (accessed December 19, 2025).
[21] M. Novotny, R. Cmejla, T. Tykalova, Automated prediction of children's age from voice acoustics, Biomedical Signal Processing And Control. 81 (2023) 104490. https://doi.org/10.1016/j.bspc.2022.104490.
[22] S. Skodda, W. Visser, U. Schlegel, Gender-Related Patterns of Dysprosody in Parkinson Disease and Correlation Between Speech Variables and Motor Symptoms, Journal Of Voice. 25 (2011) 76-82. https://doi.org/10.1016/j.jvoice.2009.07.005.
[23] T. Tykalova, J. Rusz, J. Klempir, M.T. Pellecchia, Speech impairment in a large sample of patients with multiple system atrophy: effect of gender and language, Journal Of Neural Transmission. (2025). https://doi.org/10.1007/s00702-025-03016-9.
[24] A. Favaro, A. Butala, T. Thebaud, J. Villalba, N. Dehak, L. Moro-Velázquez, Unveiling early signs of Parkinson’s disease via a longitudinal analysis of celebrity speech recordings, Npj Parkinson's Disease. 10 (2024). https://doi.org/10.1038/s41531-024-00817-9.
[25] K. Gao, A. Favaro, N. Dehak, L.M. Velazquez, ADCeleb: A Longitudinal Speech Dataset from Public Figures for Early Detection of Alzheimer’s Disease, in: Interspeech 2025, ISCA, ISCA, 2025: pp. 5688-5692. https://doi.org/10.21437/interspeech.2025-523.