A Comparative Study of Parkinsonian Speech Corpora for Deep Learning-Based Detection of Dysarthria

Clara Ponchard, Pierre Serrano
Inria, France
clara.ponchard@inria.fr, pierre.serrano@inria.fr

Abstract

Idiopathic Parkinson's disease is associated with motor speech impairments collectively referred to as hypokinetic dysarthria, which can appear at early disease stages and remain challenging to assess objectively in clinical practice. Most automatic assessment studies rely on individual speech corpora analyzed in isolation, leaving open questions regarding their comparability and their suitability for joint use within unified classification frameworks. This study explicitly investigates the cross-corpus comparability of existing Parkinsonian speech datasets designed for hypokinetic dysarthria assessment. Rather than assuming their compatibility, we evaluate it empirically through the generalization performance of classification systems trained on single or multiple corpora. We examine which datasets can be effectively combined and whether multi-corpus training improves robustness across heterogeneous recording conditions and speech tasks. Four corpora are evaluated under intra-corpus, cross-corpus, and out-of-domain settings. Results demonstrate that multi-corpus training enhances robustness and generalization performance, while also revealing substantial differences in cross-dataset compatibility. These findings provide a clearer understanding of the degree of comparability between existing resources and offer practical guidelines for the design of future corpora and more generalizable tools for the automatic clinical assessment of Parkinsonian speech.

Keywords: Parkinson, dysarthria, deep Learning, self-supervised model

1. Introduction

Idiopathic Parkinson's disease (IPD) is characterized by progressive degeneration of dopaminergic neurons in the substantia nigra, resulting in opaminergic denervation of the striatum (Grabli, 2017). Clinically, it is primarily defined by the motor triad of bradykinesia, rigidity, and resting tremor. Among the additional motor manifestations, speech disorders affect nearly 80% of patients and are perceived as particularly disabling (Hartelius and Svensson, 2009). These impairments, collectively referred to as hypokinetic dysarthria, may involve alterations in respiration, phonation, articulation, resonance, and prosody (Pinto et al., 2010), and can emerge during the early stages of the disease or even in the prodromal (presymptomatic) phase (Logemann et al., 1978; Ho et al., 1998; Sapir, 2014). The most prominent speech abnormalities include monopitch, reduced stress, monoloudness, imprecise consonant production, inappropriate pauses, short rushes of speech, hoarse voice quality, continuous breathiness, altered pitch, and variable speech rate with a tendency toward acceleration (Darley et al., 1969). Perceptual auditory evaluation by clinicians remains the clinical gold standard for the diagnosis and longitudinal monitoring of hypokinetic dysarthria; however,it is widely debated due to its subjective nature (Ghio et al., 2007). Clinicians have therefore emphasized the need for objective and quantifiable tools to complement perceptual assessment and to enable more reliable monitoring of therapeutic interventions (Laaridh, 2017).

Automatic speech processing methods offer a promising alternative by enabling the extraction of discriminative vocal features. However, evaluation on data collected under conditions not represented during training remains essential for real-world applicability. Most existing systems rely on a single corpus or a specific task, which limits their generalization capacity and restricts the broader exploitation of learned representations to analyze data structure, for example through clustering or the identification of similar acoustic profiles. In addition, the majority of studies use embeddings extracted from the last layer of self-supervised audio models such as Wav2Vec 2.0, combined with simple classifiers, which constrains robustness and cross-corpus transferability (Hireš et al., 2023; Klempíř and Krupička, 2024; Ibarra et al., 2023; Postma and Tejedor-Garcia, 2025). Cross-corpus evaluation is therefore crucial to assess the generalization ability of learned representations and to examine the actual comparability of corpora from a classification perspective. However, no standard benchmark currently exists, and commonly used metrics such as accuracy do not always ensure reliable comparisons across studies.

To address these limitations, we propose a deep learning-based system for the automatic detection of hypokinetic dysarthria, inspired by techniques from speaker recognition. The system's robustness and generalization ability are evaluated in a cross-corpus setting using four datasets (Neurovoz, IPVSD, MDVR-KCL, and AHN), covering multiple languages, speech tasks, and recording conditions, and totaling 380 speakers, including 211 patients with Parkinson's disease. This cross-evaluation

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 2–8
11 May 2026.

aims not only to assess system robustness but also to determine to what extent these corpora are comparable and can be jointly used within a multi-corpus framework. Among them, the French AHN corpus, previously unused for classification tasks, is the only dataset including recordings in the OFF-DOPA condition, that is, when patients are recorded without the effect of dopaminergic medication. It also includes a range of UPDRS speech scores, offering a unique opportunity to analyze system performance in clinically realistic and more challenging conditions.

The main objectives of this study are: (i) to analyze the impact of single- and multi-corpus training on performance under intra-corpus, cross-corpus, and out-of-domain conditions; (ii) to assess the comparability of existing corpora through system generalization and identify effective corpus combinations for joint training; and (iii) to propose a reproducible and clinically relevant evaluation protocol based on appropriate metrics, such as fixed-point AUC.

This approach enables the quantification of model robustness to Parkinsonian speech variability, provides insight into the compatibility of existing resources, and highlights the limitations of current systems, particularly in realistic screening scenarios. It thus contributes to the development of evaluation strategies that improve the clinical relevance and transferability of automatic hypokinetic dysarthria detection systems.

2. Materials and Methods

2.1. Corpus

In this study, four corpora were used, the first three of which are freely accessible: NeuroVoz (Mendes-Laureano et al., 2024), IPVSD (Dimauro et al., 2017), MDVR-KCL (Jaeger et al., 2019), and AHN (Ghio et al., 2021). A brief description of each corpus is presented below. Table 1 summarizes the number of participants in the Parkinson's disease (PD) and healthy control (HC) groups, along with their distribution by sex and mean age.

2.1.1. NeuroVoz

The NeuroVoz dataset, which is publicly available, is a collection of speech recordings designed for the development and validation of machine learning models for the diagnosis and monitoring of Parkinson's disease (PD) (Mendes-Laureano et al., 2024). It was recorded jointly by the Bioengineering and Optoelectronics Group (ByO) at the Universidad Politécnica de Madrid (UPM) and the Departments of Otorhinolaryngology and Neurology at Hospital General Universitario Gregorio Marañón (HGUGM) and Hospital Universitario de Fuenlabrada (HUF). We analyze a subset comprising voice recordings from 107 native speakers of Castilian Spanish, including 54 healthy control subjects and 53 patients with PD. Patients were recorded in the ON state, after taking their prescribed medication between two and five hours prior to the recording session. The protocol includes four types of speech tasks: (1) sustained phonation of vowels; (2) repetition of 15 predefined sentences; (3) a diadochokinetic (DDK) task based on the rapid repetition of /pa-ta-ka/; and (4) a free monologue based on the description of an illustration. Recordings were conducted under standardized conditions using an AKG C420 head-mounted microphone, with a sampling rate of 44.1 kHz and a 16-bit resolution.

2.1.2. IPVSD

The Italian Parkinson's Voice and Speech Database (IPVSD) corpus was created to assess speech intelligibility in patients with Parkinson's disease using automatic speech recognition systems (Dimauro et al., 2017). It includes recordings from 65 Italian speakers, mainly from the Bari region, divided into three groups: 15 young healthy controls aged 19 to 29 years (13 men, 2 women), 22 older healthy controls aged 60 to 77 years (10 men, 12 women), and 28 patients with Parkinson's disease aged 40 to 80 years (19 men, 9 women). The Parkinsonian patients were recorded in the ON state, after taking their prescribed medication between two and five hours prior to the recording session. The protocol includes several speech tasks: (1) reading of phonetically balanced texts, words, and sentences; (2) sustained vowel phonation; and (3) diadochokinetic (DDK) tasks (/pa/ and /ta/). The acoustic signals were recorded in uncompressed WAV format, with a sampling rate of 44.1 kHz.

2.1.3. MDVR-KCL

The MDVR-KCL corpus, which is publicly available, was recorded at King's College London (KCL) Hospital under conditions designed to replicate a realistic telephone call scenario, with participants holding the phone to their preferred ear and the microphone positioned close to the mouth (Jaeger et al., 2019). It includes voice recordings from 37 native English speakers, comprising 21 healthy control subjects and 16 patients with Parkinson's disease (PD). No information is provided regarding the participants' sex distribution or age, nor about whether patients had taken medication prior to the recording sessions. The protocol includes reading aloud the text "The North Wind and the Sun", optionally followed by the reading of a technical passage, and a spontaneous verbal exchange with the examiner. Recordings were made using a Motorola Moto G4

Corpus	Language	Participants					Age (mean ± SD)			
		Total	Female		Male		Female		Male	
			PD	HC	PD	HC	PD	HC	PD	HC
Neurovoz	Spanish	107	20	26	33	28	67.2 ± 9.1	66.6 ± 12.3	69.8 ± 11.4	61.6 ± 7.5
IPVSD	Italian	65	9	14	19	23	64.3 ± 12.2	58.7 ± 17.0	68.6 ± 6.4	42.0 ± 24.8
MDVR-KCL	English	37	-	-	-	-	-	-	-	-
AHN	French	171	43	40	77	11	64.1 ± 9.9	62.2 ± 8.6	65.8 ± 10.0	66.6 ± 14.1

Table 1: Distribution of Parkinson's disease (PD) and healthy control (HC) participants across corpora, including language, sex, and mean age (± standard deviation, SD).

smartphone via a dedicated application (Toggle Recording App), stored in uncompressed WAV format, with a sampling rate of 44.1 kHz and a 16-bit resolution.

2.1.4. AHN

The AHN (Aix Hôpital Neurologie) corpus consists of acoustic and aerodynamic recordings collected over more than twenty years by the Neurology Department of the Aix-en-Provence Hospital Center (Ghio et al., 2021). It is distinguished by the presence of aerodynamic signals synchronized with the acoustic recordings, as well as by the diversity of recording conditions (with or without L-DOPA, with or without deep brain stimulation). We analyze a subset of the corpus comprising 171 French-speaking participants, including 120 patients with Parkinson's disease (PD) and 51 healthy control subjects. The Parkinsonian patients were treated exclusively with L-DOPA. Most patients were recorded in two pharmacological states: OFF-DOPA, after at least 12 hours of medication withdrawal, and ON-DOPA, after a minimum delay of 1.5 hours following L-DOPA intake. Before each recording session, dysarthria severity was assessed by a neurologist using Item 18 of the UPDRS scale (Fahn and Elton, 1987), with scores ranging from 0 to 3. The analyzed task consisted of repeating the sentence "papa ne m'a pas parlé de beau papa", designed for the evaluation of plosive consonants. Data were collected using the Assisted Voice Evaluation system EVA2 (Teston et al., 1999), which enables synchronized acquisition of acoustic and aerodynamic signals via a handheld device mounted on an adjustable stand and coupled with a sealed silicone mask fitted around the speaker's mouth. The acoustic signal was recorded in WAV format using an AKG C419 directional microphone positioned 4 cm from the mask, with a sampling rate of 25 kHz.

2.2. Method

The task consists of a binary classification aimed at predicting whether a speaker is a healthy control (HC) or a Parkinson's disease (PD) patient from an audio recording. The architecture combines a self-supervised model, Wav2Vec 2.0 XLS-R (Babu et al., 2022), pre-trained on a large multilingual corpus, with a multi-head feature aggregation (MHFA) classification head that aggregates representations from multiple SSL layers with a hidden dimension of 128 (Peng et al., 2022). The parameters of the SSL model are frozen, and only those of the MHFA head are optimized, allowing the model to implicitly select the most informative layers. This approach is inspired by modern solutions for speaker recognition and audio deepfake detection (Wang et al., 2024).

The corpora described above are used in the experiments. For each corpus, the data are split into 80% for training and 20% for testing, with 20% of the training set dedicated to development. Five independent random splits are generated to assess robustness (Postma and Tejedor-Garcia, 2025; Gimeno-Gómez et al., 2025), ensuring strict speaker separation. Performance is reported as mean and standard deviation. Audio recordings are resampled to 16 kHz and segmented into randomly selected one-second chunks for training, with balanced batches and optimization using Adam with a cross-entropy loss function. Training runs for at least 100 epochs, and the last checkpoint is retained. During inference, the full recording is provided to the model, and the MHFA head aggregates temporal representations into a single prediction.

Evaluation is based on accuracy and AUC. Intra-corpus evaluation uses the test set from the same corpus as training, whereas inter-corpus evaluation tests on a different corpus, representing a more challenging out-of-domain scenario. Models can be trained in either mono-corpus or multi-corpus settings. All predictions are made at the recording level, without speaker-level aggregation. The MDVR-KCL corpus, whose size does not allow a split into training and test sets, is used exclusively as an out-of-domain evaluation corpus.

Test →	Neurovoz		IPVSD		AHN		MDVR-KCL	
Train ↓	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)
Neurovoz (N)	76.8 ± 0.7	84.9 ± 2.3	69.9 ± 7.4	43.4 ± 7.3	79.6 ± 1.4	51.2 ± 3.3	83.6 ± 4.3	90.2 ± 3.2
IPVSD (I)	65.9 ± 3.0	68.2 ± 5.0	94.4 ± 2.7	95.8 ± 4.1	80.8 ± 4.5	66.7 ± 8.0	70.4 ± 2.5	72.7 ± 4.2
AHN (A)	59.5 ± 3.6	54.7 ± 3.8	69.4 ± 6.6	63.9 ± 2.1	83.4 ± 4.6	79.2 ± 7.4	69.9 ± 6.0	67.1 ± 9.7

Table 2: Cross-corpus performance of systems trained on a single corpus. Accuracy (Acc) and area under the ROC curve (AUC) are reported with their standard deviations (±) computed over five independent splits. Grey cells indicate that evaluation is performed on the test set of the corpus used for training.

Test →	Neurovoz		IPVSD		AHN		MDVR-KCL	
Train ↓	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)
I + A	61.2 ± 2.9	61.4 ± 6.6	93.8 ± 4.3	92.9 ± 9.2	83.0 ± 4.6	78.4 ± 8.7	74.5 ± 3.9	74.1 ± 4.6
N + A	75.9 ± 1.8	84.1 ± 3.8	66.8 ± 9.6	35.5 ± 10.1	83.8 ± 2.3	81.3 ± 3.6	79.2 ± 1.8	84.0 ± 3.6
N + I	78.6 ± 2.2	85.9 ± 3.7	91.5 ± 3.6	95.0 ± 4.9	77.7 ± 2.6	59.2 ± 6.7	82.2 ± 1.9	87.2 ± 2.6
N + I + A	77.7 ± 3.4	85.4 ± 4.5	91.5 ± 3.1	93.6 ± 7.1	83.7 ± 3.4	84.3 ± 4.2	80.8 ± 1.4	85.3 ± 1.6

Table 3: Cross-corpus performance of systems trained on corpus combinations. Metrics and conventions follow those of Table 2. Grey cells indicate intra-corpus evaluation.

3. Results

The performance of systems trained on a single corpus is presented in Table 2. In the intra-corpus setting, mono-corpus models generally achieve strong performance. The model trained on IPVSD (I) obtains the highest AUC (95.8%), followed by Neurovoz (N) (84.9%) and AHN (A) (79.2%).

Cross-corpus evaluation reveals more heterogeneous generalization capabilities. The N model transfers effectively to MDVR-KCL (AUC of 90.2%), whereas the I and A models exhibit more limited transfer performance. Notably, applying the N model to the A corpus results in a high accuracy (79.6%) but a near-random AUC (51.2%), indicating unstable or poorly calibrated discrimination behavior.

Bi-corpus combinations (Table 3) show contrasting synergies. The N+I pair stands out in its generalization to MDVR-KCL (87.2% AUC), but fails on AHN (59.2%). All pairs systematically fail on at least one unseen corpus. In contrast, multi-corpus training on all three corpora (N+I+A) yields robust performance across all datasets and avoids the severe degradation observed with mono- or bi-corpus models. Results on MDVR, which was never seen during training, further confirm the robustness of this approach (AUC 85.3%).

ROC curves (Figure 1) indicate that global AUC does not always reflect clinically meaningful performance. At operating points with low false positive rates (FPR), which are more relevant in clinical settings, performance gains differ across corpora. The evaluation also reveals strong sensitivity to random splits, particularly for IPVSD and AHN, highlighting limitations related to the small size of these datasets.

A post-hoc analysis of the N+I+A detector was conducted on a representative split (close to the average performance across the five training runs). [1]

4. Discussion

The results show that multi-corpus training on three corpora (N+I+A) provides the best trade-off between specialization and generalization. While mono-corpus models achieve strong intra-corpus performance, their cross-corpus transfer capabilities remain limited. All bi-corpus combinations fail on at least one unseen corpus, confirming that exposure to only two protocols is insufficient to capture the acoustic diversity associated with Parkinson's disease, as observed in (Ibarra et al., 2023).

The analysis of ROC curves highlights the limitations of accuracy (and, by extension, the F1-score) as primary evaluation metrics for hypokinetic dysarthria corpora, given their strong heterogeneity and often limited size. Systems displaying high accuracy may nevertheless exhibit near-random discrimination ability, as illustrated by the inference of the N system on the AHN corpus (Table 3). Evaluation based on AUC or on fixed operating points, particularly at low false positive rates, appears essential for comparing systems and more faithfully assessing their real clinical applicability, beyond the considerations typically reported in the literature (Ozbolt et al., 2022). Furthermore, the absence of a fixed, shared evaluation corpus strictly independent from training constitutes a major obstacle to reliable

[1] It is not possible to merge systems obtained from different training runs, as such an approach would require a common calibration and the use of identical training data.

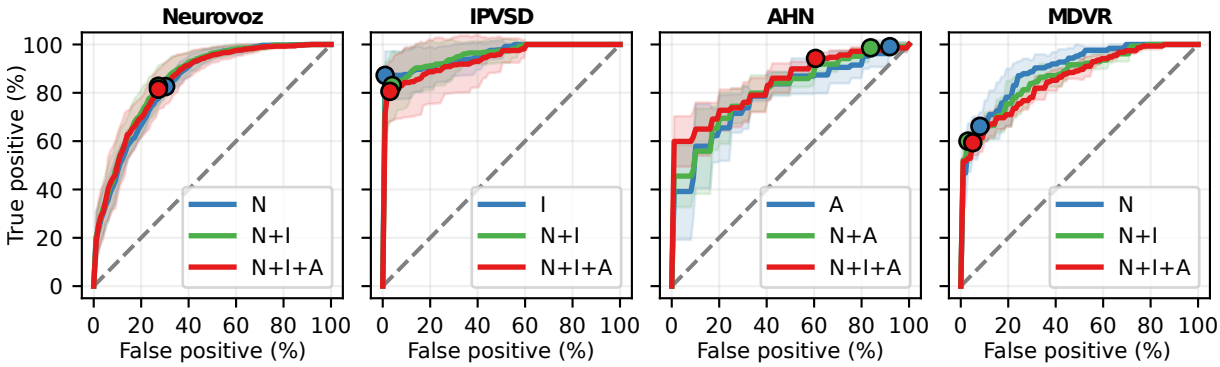

Figure 1: Mean ROC curves (± standard deviation) for models trained on one or several of the Neurovoz (N), IPVSD (I), and AHN (A) corpora. Dots indicate the operating point maximizing classification accuracy for each evaluated system.

comparisons across systems, architectures, and learning parameter configurations.

The use of a frozen self-supervised learning (SSL) model combined with a multi-layer classification head helps mitigate overfitting in a low-data setting. This approach achieves performance comparable to, and in some cases exceeding, the state of the art across all corpora. Although differences in evaluation protocols make direct comparisons challenging, our results are competitive: for the DDK task of the Neurovoz corpus, we achieve 90% accuracy, compared to 85% reported by (Postma and Tejedor-Garcia, 2025) and 88.6% by (Ibarra et al., 2023).

The robustness of our approach is further confirmed by the results on MDVR, which was never seen during training, and by consistently high AUC values across all test sets. A preliminary analysis suggests that intermediate layers, particularly around the 20th layer, play a key role, likely encoding prosodic information relevant for hypokinetic dysarthria detection. Beyond metric selection, evaluation protocols also raise important issues in comparisons of SSL models. While several studies compare such models, these comparisons most often rely on simplified protocols, typically involving a linear classifier or an SVM applied to the final layer representation. However, as recently discussed in (Zaiem et al., 2025), such configurations are not appropriate for establishing reliable benchmarks, particularly in low-resource settings. Our results show that this type of approach may underexploit SSL representations and lead to unstable conclusions, reinforcing the previously highlighted need for stricter and more clinically meaningful evaluation protocols.

5. Conclusion

The use of pre-trained self-supervised audio models, combined with multi-corpus training, enables the development of a hypokinetic dysarthria detection system capable of generalizing across languages, datasets, and recording conditions, including out-of-domain scenarios. Multi-corpus training therefore emerges as a central lever to overcome the limitations of approaches evaluated on single, homogeneous datasets.

Beyond predictive performance, the embeddings extracted from the classification head of the proposed system—explicitly constrained by the hypokinetic dysarthria detection task and validated in a multi-corpus setting—constitute a relevant foundation for future post hoc analyses of clinical databases. They open the way to investigations such as clustering of recordings or patients, with the objective of identifying homogeneous acoustic profiles and examining their associations with clinical or protocol-related factors.

The results highlight the importance of rigorous evaluation protocols and appropriate metrics, such as fixed-operating-point AUC, as well as the need for shared evaluation corpora that are independent from training data and cover diverse clinical conditions.

Finally, post hoc analyses conducted on the AHN corpus reveal the current limitations of the models in cases of mild dysarthria (low UPDRS scores) and under realistic screening conditions (OFF medication state). These findings suggest that targeted data augmentation strategies, fine-grained balancing of clinical tasks, and multimodal approaches represent promising directions to enhance both robustness and clinical relevance.

6. Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2026-AD011014982R2 made by GENCI.

7. Bibliographical References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Frederic L. Darley, Arnold E. Aronson, and J. R. Brown. 1969. Clusters of Deviant Speech Dimensions in the Dysarthrias. *Journal of Speech and Hearing Research*, 12(3):462–496.

Stanley Fahn and Richard L. Elton. 1987. The unified parkinson's disease rating scale. In Stanley Fahn, C. David Marsden, Donald Calne, and Melvin Goldstein, editors, *Recent Developments in Parkinson's Disease*, volume 2, pages 153–163. Macmillan Health Care Information, Florham Park.

Alain Ghio, Gilles Pouchoulin, Antoine Giovanni, Corinne Fredouille, Bernard Teston, Joana Révis, Jean-François Bonastre, Danièle Robert-Rochet, Ping Yu, Maurice Ouaknine, Marie-Dominique Guarella, Christine Spezza, Thierry Legou, and Alain Marchal. 2007. Approches complémentaires pour l'évaluation des dysphonies : bilan méthodologique et perspectives. *Travaux interdisciplinaires du Laboratoire Parole et Langage*, 26:33–74. Autorisation No.3240 : TIPA est la revue du Laboratoire Parole et Langage.

David Gimeno-Gómez, Catarina Botelho, Anna Pompili, Alberto Abad, and Carlos-D Martínez-Hinarejos. 2025. Unveiling interpretability in self-supervised speech representations for parkinson's diagnosis. *IEEE Journal of Selected Topics in Signal Processing*.

David Grabli. 2017. Maladie de parkinson et syndromes parkinsoniens : les signes moteurs. *La Presse Médicale*, 46(2, Part 1):187–194.

L. Hartelius and P. Svensson. 2009. Speech and swallowing symptoms associated with parkinson's disease and multiple sclerosis: A survey. *Folia Phoniatrica et Logopaedica*, 46(1):9–17.

M. Hireš, P. Drotár, N. D. Pah, Q. C. Ngo, and D. K. Kumar. 2023. On the inter-dataset generalization of machine learning approaches to parkinson's disease detection from voice. *International Journal of Medical Informatics*, 179:105237.

A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates. 1998. Speech impairment in a large sample of patients with parkinson's disease. *Journal of Behavioural Neurology*, 11:131–137.

Emiro J Ibarra, Julián D Arias-Londoño, Matías Zañartu, and Juan I Godino-Llorente. 2023. Towards a corpus (and language)-independent screening of parkinson's disease from voice and speech through domain adaptation. *Bioengineering*, 10(11):1316.

Ondrej Klempíř and Roman Krupička. 2024. Analyzing wav2vec 1.0 embeddings for cross-database parkinson's disease detection and speech features extraction. *Sensors*, 24:5520.

I. Laaridh. 2017. *Évaluation de la parole dysarthrique*. Ph.D. thesis, Université d'Avignon.

J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky. 1978. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients. *Journal of Speech and Hearing Disorders*, 43(1):47–57.

Ana S. Ozbolt, Laura Moro-Velazquez, Irene Lina, Adam A. Butala, and Najim Dehak. 2022. Things to consider when automatically detecting parkinson's disease using the phonation of sustained vowels: Analysis of methodological issues. *Applied Sciences*, 12(3):991.

J. Peng, O. Plchot, T. Stafylakis, L. Mošner, L. Burget, and J. Černocký. 2022. An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 555–562.

S. Pinto, A. Ghio, B. Teston, and F. Viallet. 2010. La dysarthrie au cours de la maladie de parkinson. *Revue Neurologique*, 166(10):800–810.

Emmy Postma and Cristian Tejedor-Garcia. 2025. Evaluating the Effectiveness of Pre-Trained Audio Embeddings for Classification of Parkinson's Disease Speech Data. In *Interspeech 2025*, pages 4603–4607.

S. Sapir. 2014. Multiple factors are involved in the dysarthria associated with parkinson's disease. *Journal of Speech, Language, and Hearing Research*, 57(4):1330–1343.

Bernard Teston, Alain Ghio, and Bernard Galindo. 1999. A multisensor data acquisition and processing system for speech production investigation. pages 2251–2254.

X. Wang, H. Delgado, H. Tak, et al. 2024. Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *Proceedings of the ASVspoof 2024 Workshop*, pages 1–8.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. 2025. Speech self-supervised representations benchmarking: a case for larger probing heads. *Computer Speech & Language*, 89:101695.

8. Language Resource References

Dimauro, Giovanni and Di Nicola, Vincenzo and Bevilacqua, Vitoantonio and Caivano, Danilo and Girardi, Francesco. 2017. *Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System*.

Ghio, Alain and Pouchoulin, Gilles and Viallet, François and Giovanni, Antoine and Woisard, Virginie and Crevier-Buchman, Lise and Hirsch, Fabrice and Fauth, Camille and Fredouille, Corinne. 2021. *Du recueil à l'exploitation des corpus de parole « pathologique » : comment accéder à la variation physiopathologique ?* Corpus. PID https://doi.org/10.4000/corpus.5677.

Jaeger, Hagen and Trivedi, Dhaval and Stadtschnitzer, Michael. 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. PID https://doi.org/10.5281/zenodo.2867216.

Mendes-Laureano, J. and Gómez-García, J. A. and Guerrero-López, A. and others. 2024. *NeuroVoz: a Castillian Spanish corpus of parkinsonian speech*. PID https://doi.org/10.1038/s41597-024-04186-z.