Enhancing Detection of Parkinson-Induced Dysarthria with Cross-Lingual Transfer Learning

Andreas Rouvalis[1,2(✉)], Johannes Tröger[2], Julius Steuer[1], Juan Rafael Orozco Arroyave[3], Jan Rusz[4], Jouni Pohjalainen[2], Hali Lindsay[2], Bernd Möbius[1], and Dietrich Klakow[1]

[1] Department of Language Science and Technology, Saarland University, Saarbrücken, Germany
androuvalis@gmail.com
[2] ki elements GmbH, Saarbrücken, Germany
[3] GITA Lab, Faculty of Engineering, University of Antioquia, Medellín, Colombia
[4] Department of Circuit Theory, Czech Technical University in Prague, Prague, Czechia

Abstract. This paper presents a transfer learning approach that leverages cross-linguistic transferability coupled with largely interpretable acoustic features to improve dysarthria detection in Parkinson's disease. It uses features extracted from sustained phonation of /a/ and diadochokinetic exercises in Czech, Spanish, American English, and Italian. The approach addresses data sparsity in clinical settings and accounts for variability due to age and sex. Transfer learning models outperform monolingual classifiers (i.e. classifiers trained and tested on the same language) in most tests, demonstrating the effectiveness of this approach in overcoming data limitations and enhancing Parkinson-induced dysarthria detection.

Keywords: Parkinson's disease · transfer learning · dysarthria detection

1 Introduction

Parkinson's disease (PD) is a chronic, neurodegenerative disorder affecting approximately 1% of the global population over 60 years old (late-onset PD) and a smaller proportion of those aged 30–50 (early-onset PD) [1]. PD is characterized by the progressive loss of dopamine-producing neurons in the substantia nigra [1], leading to a range of motor symptoms—including rigidity, resting tremor, and postural instability—as well as non-motor symptoms such as sleep disturbances and cognitive impairments [2].

Speech impairments are highly prevalent in PD, with around 90% of patients developing hypokinetic dysarthria (HD) [3]. This condition progressively affects

K. Ekštein et al. (Eds.): TSD 2025, LNAI 16029, pp. 218–234, 2026.
https://doi.org/10.1007/978-3-032-02548-7_19

all components of speech production, including respiration, phonation, articulation, and prosody [4,5], significantly reducing speech intelligibility and contributing to social withdrawal and diminished quality of life [6]. HD is characterized by incomplete vocal fold adduction and abduction [7], bradykinesia [8], and reduced articulatory control [9], resulting in increased acoustic noise, reduced voice intensity, breathy and harsh vocal quality, monotone pitch, imprecise consonant articulation, vowel centralization, and irregular speech rate with frequent pauses [10,11].

Current methods for diagnosing dysarthria rely heavily on clinical expertise, with studies showing that trained clinicians and neurologists achieve diagnostic accuracies of 40–60%, depending on the available clinical data [12,13]. Specifically, diagnosing PD is time-consuming, often taking an average of 2.9 years to confirm with a 90% accuracy [14]. These shortcomings highlight the need for objective, efficient, and accessible diagnostic tools that reduces clinician bias and streamline the diagnostic process.

Since speech impairments often precede other motor symptoms in PD, automatic speech analysis offers a promising and non-invasive approach for early PD detection [15]. However, this approach faces two primary challenges: data sparsity and demographic variability. Limited language-specific data pose a significant barrier in training machine learning models, as data collection is time-consuming, costly, and sometimes unfeasible [16]. Additionally, age and sex significantly influence speech patterns, complicating the differentiation between PD-related dysarthria and natural speech variations [17,18].

To address data sparsity, transfer learning (TL) has been applied alongside convolutional neural networks showing promising results in two settings; within PD across languages [19] and across conditions (from PD to Huntington's diseases) [20]. Vasquez et al. [19] trained CNNs time frequency representations on Spanish, German, and Czech PD data. Transfer learning improved generalization by balancing specificity and sensitivity while reducing variability. Despite these advantages, time-frequency features lack interpretability which is crucial for clinical applications.

To further enhance model performance, curriculum learning (CL)—which organizes training data in a meaningful progression—has demonstrated advantages over standard random shuffling, offering performance gains without introducing additional computational overhead [21]. Dhinagar et al. [22] demonstrated a 3.9% performance improvement when training CNNs on T1-weighted MRI scans ordered according to the Hoehn and Yahr (H&Y) scale. They began with more severe, easier-to-classify Parkinson's cases (H&Y = 4) and progressed to milder ones (H&Y = 1), suggesting that data scheduling informed by domain knowledge can improve learning efficiency and robustness.

This study develops neural classifiers using speech data from multiple languages and tasks to improve model robustness. By pretraining on one or more languages and fine-tuning on another, our approach addresses data sparsity and enhances performance in the target language. We pretrain our own classifiers, allowing control over the pretraining data and enabling better understanding

of how its characteristics influence TL effectiveness—unlike the vast, opaque datasets used to train large models. Importantly, our method not only improves dysarthria detection in PD but also leverages a largely interpretable feature set, enhancing its potential clinical utility. To reduce demographic bias, we use linear models to regress out age and sex effects. In addition to random data shuffling, we explore curriculum learning principles through simple, domain-agnostic training schedules—specifically, by ordering the data from healthy controls (HC) to PD individuals and from PD to HC. These simple strategies do not depend on clinical severity scores, making it especially useful when such scores (e.g. HY, UPDRS) are unavailable.

The key contributions of this research are:

- Evaluating the effectiveness of TL across languages for detecting PD-induced dysarthria.
- Capitalizing on the transferability of largely interpretable acoustic features extracted from diadochokinetic exercises and sustained phonation.
- Evaluating CL strategies agnostic to domain knowledge.

2 Data

2.1 Speech Tasks and Feature Extraction

To assess articulatory and acoustic characteristics, two speech tasks are employed: sustained phonation of /a/ (SPA) and oral diadochokinesis (DDK).

In sustained phonation, participants are tasked to produce the vowel /a/ for as long as possible with consistent pitch and amplitude. Two forms of DDK are used: Alternating Motion Rate (AMR), where participants repeatedly produced a single syllable (/pa/ or /ta/ or /ka/) as quickly and consistently as possible, and Sequential Motion Rate (SMR), where they continuously repeat a syllable sequence (/pa/-/ta/-/ka/).

Twenty task-specific acoustic features are extracted from the SPA task and 62 general acoustic features are extracted from the SPA and DDK tasks[1]. Feature extraction is carried out using our in-house feature extraction engine.

SPA-specific acoustic features capture various aspects of voice stability, variability, and consistency over time. Among these, spectral features describe how energy is distributed across frequencies in the voice signal. For instance, spectral rolloff indicates the frequency below which most of the signal's energy is concentrated, while spectral flatness reflects how noise-like versus tonal the sound is [23]. Harmonic descriptors such as Tonnetz features, which capture relationships between harmonic components, are also included, although they can be more abstract and somewhat difficult to interpret directly [24].

General acoustic features assess voice characteristics across multiple domains for both the DDK and SPA tasks. Articulation measures (e.g., F1 mean frequency, F2 standard deviation) are tied to tongue and vocal tract movement [25].

[1] See Appendix for a full list of these features along with short definitions.

Pitch-related features (e.g., maximum pitch, pitch standard deviation) reflect vocal fold tension and control [26]. Voice variability features, such as jitter and shimmer, reflect the stability and regularity of vocal fold vibrations in terms of frequency and amplitude, respectively [27]. Speech rhythm and dynamics (e.g., pause rate, duration) provide insight into motor control, while loudness features (e.g., mean loudness, loudness peaks) are tied to respiratory strength and vocal fold tension. Finally, spectral/harmonic features offer general speech quality insights reflecting overall vocal tract shape and resonances. Features in this group, such as Mel-Frequency Cepstral Coefficients (MFCCs), are often considered black-box representations. However, recent research has shed light on their interpretability—for instance, MFCC2 reflects a weighted ratio between low- and high-frequency energy, which is linked to voice alterations caused by disease. As we move to higher orders, MFCCs capture faster spectral variations whose biological meaning becomes increasingly difficult to interpret [28].

These features offer a multidimensional profile of speech characteristics, which is valuable because speech alterations in PD stem from impairments across all subsystems of speech production—respiration, phonation, articulation, resonance, and prosody [4,5]. Moreover, the majority of these features are interpretable, making them particularly suitable for clinical applications. General acoustic features are particularly useful for TL, as they can be extracted from both tasks, maximizing data and addressing sparsity. To leverage this, classifiers are trained on both tasks combined and separately for clarity. For sustained phonation, both general and task-specific features are evaluated together and separately. While task-specific features are more limited since they can only be extracted from sustained phonation, they remain valuable due to the task's widespread use in neurodegenerative disease research.

2.2 Datasets

We use speech data from five datasets: PC-GITA (Colombian Spanish) [29] and its extended version (e-PC-GITA) [30], mPower dataset (American English) [31], Parkinson's Voice and Speech (Italian) [32], and the Czech PD dataset [33]. These datasets include recordings from individuals with PD and HC across multiple languages and varying recording conditions, providing a diverse and rich source of speech data. Table 1 presents a summary of the demographic information for the datasets.

We merge the **PC-GITA corpus** [29] with **e-PC-GITA** [30], forming a larger set with 70 PD and 70 HC speakers, with equal sex distribution and balanced age groups. Participants were recorded in either noise-controlled (50 PD, 50 HC) or real-life conditions using smartphones (20 PD, 20 HC). Tasks included sustaining the phonation of the vowel /a/ and repeating syllables and syllable sequences such as /pa-ta-ka/, /pa-ka-ta/, /pa-pa-pa/, /ta-ta-ta/ and /ka-ka-ka/. There are three repetitions available of each task per participant.

A subset of the **mPower dataset** [31] is used, consisting of 797 recordings of sustained phonation of the vowel /a/ from 397 HC and 400 PD individuals.

These recordings, collected using smartphone technology, were selected following a strict exclusion process that removed participants with confounding medical conditions such as Alzheimer's or depression. All participants were English speakers, with a gender-balanced distribution across both groups.

The **Italian Parkinson's Voice and Speech dataset** [32] includes 22 HC (13 females) and 28 individuals with PD (12 females). The tasks involved sustained phonation of /a/ and diadochokinetic exercises (/pa/ and /ta/) in a soundproof room, with two repetitions of each task per participant.

The **Czech PD dataset** [33] consists of 45 HC (17 females) and 35 PD participants (17 females). Tasks included sustained phonation of /a/ and repetition of the syllables /pa-ta-ka/, with two repetitions available for most participants. Recordings were made in a quiet environment with minimal background noise, using a condenser microphone positioned approximately 15 cm from the speaker's mouth.

Table 1. Demographics of the Speech Datasets

Dataset	Language	PD Total (F)	PD Age	HC Total (F)	HC Age	Tasks
PC-GITA + e-PC-GITA	Spanish (Colombia)	70 (35F)	61.12 (±10.94)	70 (35F)	61.43 (±9.45)	Sustained /a/, /pa-ta-ka/, /pa-ka-ta/, /pa-pa-pa/, /ka-ka-ka/, /ta-ta-ta/
mPower	English (USA)	400 (129F)	60.48 (±10.61)	397 (128F)	37.55 (±13.72)	Sustained /a/
Parkinson's Voice	Italian	28 (12F)	66.62 (±9.35)	22 (13F)	68.79 (±4.29)	Sustained /a/, /pa-pa-pa/, /ta-ta-ta/
Czech PD	Czech	35 (17F)	63.69 (±9.72)	45 (17F)	62.57 (±10.02)	Sustained /a/, /pa-ta-ka/

3 Methodology

3.1 Data Curation

We impute missing data based on the distribution of each feature, using the Shapiro-Wilk test to assess normality separately for HC and individuals with PD. For normally distributed features, we use the mean; for non-normal ones, the median. Some features may be missing due to extraction limitations—for example, the Amplitude Perturbation Quotient 11 (APQ11), which measures short-term amplitude variation in the voice, requires at least 11 consecutive pitch periods. If the voice sample is too short or irregular, APQ11 cannot be computed and is therefore missing.

We aggregate and average multiple repetitions of each task per participant to reduce noise caused by external factors like stress or distractions. This common practice [34] enhances reliability and provides a more accurate representation of participants' speech profiles.

We perform data normalization (z-score) to standardize feature scales using the training set. This prevents features with large ranges from dominating others

and improves computational efficiency during model training by accelerating convergence.

A multivariate linear regression model is applied to each feature to control for age and sex effects. For each feature, the feature value serves as the dependent variable, with age and sex as predictors. The resulting residuals—which represent feature values with demographic effects removed—are used in subsequent classification tasks. This adjustment helps establish clinically relevant PD-related speech patterns from natural variations due to demographic factors.

All preprocessing steps are embedded within the cross-validation pipeline (see Sect. 3.4) to maintain unbiased model evaluation and prevent data leakage.

3.2 Baseline

To assess the performance of the proposed models, we use monolingual feed-forward neural networks (FFNNs) —trained and test on data from the same language—as baselines. These models serve as benchmarks, allowing us to evaluate the effectiveness of TL approaches.

Four separate FFNNs are trained, one per language. All baseline models follow a consistent architecture designed for binary classification (distinguishing between individuals with PD and HC). Each network includes three fully connected hidden layers, with Leaky ReLU activation functions: a negative slope of 0.2 for the first layer and 0.3 for the subsequent two. A dropout rate of 0.3 is applied after each hidden layer to mitigate overfitting. The binary cross-entropy loss function is used, and optimization is performed using Adam, chosen for its efficiency and fast convergence.

The number of neurons in each hidden layer is determined via hyperparameter search specific to each language's dataset: 90 neurons for Spanish, 32 for Czech, 64 for Italian, and 128 for American English. Additional hyperparameters such as batch size (ranging from 32 to 128), learning rate (from 5e-5 to 5e-4), and the number of training epochs are also optimized per test. Training is performed for a maximum of 120 epochs with early stopping (patience = 20, delta = 0.002). In practice, all models converged within 15-30 epochs.

3.3 Transfer Learning Architecture

The TL models are initialized using the same architecture as their monolingual counterparts, with the same number of layers, neurons, activation functions, dropout rate, optimizer, and loss function. TL experiments are conducted under two setups: (1) pre-training on a single source language and fine-tuning on a target language, and (2) pre-training on a combination of all source languages except the target language, followed by fine-tuning on the target.

To encourage effective weight initialization without full convergence, pre-training is intentionally brief and limited to a maximum of 3 epochs. Fine-tuning is performed with early stopping based on a validation split (20% of the training set), using stratified sampling to preserve class balance. During

fine-tuning, key hyperparameters such as batch size (ranging from 32 to 128), learning rate (ranging from 5e-5 to 5e-4), patience (set to 20), and delta (set to 0.002) are re-optimized for each experiment.

3.4 Model Evaluation

Model evaluation is carried out using repeated stratified K-fold cross-validation to ensure unbiased performance estimation. Specifically, we use 5 folds with 2 repeats, reserving 20% of the training data for testing at each iteration. Stratification maintains class distribution across folds. For tests involving multiple data points per participant (e.g., combining features from DDK and SPA tasks), stratified group K-fold validation with 10 splits is employed to ensure that all samples from a single participant are confined to either the training or test set, thus avoiding data leakage. Model performance is measured using balanced accuracy.

4 Results

Table 2 shows the PD-induced dysarthria detection results for different fine-tuning languages. Transfer learning models achieve the highest balanced accuracy in most tests, surpassing baseline classifiers, highlighting the effectiveness of this approach in PD-induced dysarthria.

In addition to randomly shuffling the data, we evaluate two simple curriculum learning strategies that do not rely on domain knowledge: (1) training with data ordered from HC to PD, and (2) from PD to HC. These ordering schedules are applied consistently across training folds. For clarity, we report results only for the baseline (random shuffling) and the better-performing curriculum strategy: HC-to-PD.

When fine-tuning for Italian, Spanish emerges as the best pre-trained option, surpassing the baseline in three tests. In contrast, pre-training on Czech improves performance for the DDK task and marginally for the combination of tasks, while American English yields results well below the established baseline. In Spanish, pre-training on Italian consistently outperforms not only the baseline but also Czech and American English across all tests. In Czech, there is at least one pre-training language that improves performance over monolingual classifiers in every test. However, Italian emerges as the best pre-training option for most sustained phonation tests, while Spanish outperforms Italian when DDK-extracted features are leveraged. Finally, American English exhibits a trend similar to that of Czech, though less pronounced. All pre-training languages result in a performance boost, but Italian stands out as the best pre-training data, regardless of the features used.

Table 2. Balanced Accuracy (%) of dysarthria detection across different feature sets. Results from random shuffling (shown in parentheses) serve as the baseline, while scores outside parentheses reflect the performance achieved using the HC-to-PD ordering strategy. **DDK** = general acoustic features from DDK tasks, **SPA** = general and task-specific acoustic features from sustained phonation of /a/, **SPA (general)** = general acoustic features from sustained phonation of /a/, **SPA (specific)** = task-specific acoustic features from sustained phonation of /a/, **DDK + SPA** = general acoustic features from both tasks. The first column either specifies the pre-training language or the baseline. The cases in which transfer learning improves upon the baseline are indicated in bold.

Pre-training language	Fine-tuning language	DDK	SPA	SPA (general)	SPA (specific)	DDK + SPA
Italian (baseline)	—	83.33 (80.42)	71.67 (68)	79.58 (77.59)	69.17 (66.1)	84.16 (81.09)
Spanish	Italian	**87.91** (84.58)	**74.17** (73.14)	76.67 (74.87)	57.91 (56.31)	**85.38** (83.74)
Czech		85.41 (83.33)	72.92 (71.27)	77.07 (76.58)	58.75 (56.45)	84.72 (84.11)
Am. English		—	64.58 (62.90)	70.42 (68.33)	60.42 (58.56)	—
All		84.16 (80.58)	67.45 (66.07)	72.17 (70.18)	60.75 (57.91)	85 (81.68)
Spanish (baseline)	—	73.46 (72.15)	68.93 (65.88)	68.21 (65)	56.07 (55.7)	69.51 (67.28)
Italian	Spanish	**77.37** (75.7)	**70.71** (67.85)	**70** (67.14)	**57.86** (53.57)	72.01 (71.85)
Czech		76.06 (74.59)	68.21 (67.14)	69.64 (66.42)	56.07 (52.14)	69.96 (68.46)
Am. English		—	67.14 (64.28)	67.50 (65.71)	57.14 (55.71)	—
All		74.22 (74.95)	71.42 (68.57)	67.85 (75.42)	57.85 (54.28)	**73.63** (71.12)
Czech (baseline)	—	67.99 (66.38)	65.11 (63.47)	63.79 (61.77)	60.61 59.2)	68.16 (65.6)
Italian	Czech	**71.58** (69.74)	69.13 (63.25)	**69.08** (63.91)	**67.30** (56)	66.67 (66.94)
Spanish		70.35 (69.5)	68.25 (63.83)	67.78 (63.58)	63.17 (58)	**69.23** (64.89)
Am. English		—	**71.59** (67.83)	64.76 (69.33)	65.00 (57.25)	—
All		69 (70.45)	65.16 (64.08)	68.49 (64.66)	66.58 (55.83)	64.51 (68.56)
Am. English (baseline)	—	—	61.31 (58.83)	59.17 (57)	60.72 (56.88)	—
Italian	Am. English	—	**63.85** (62.72)	**62.34** (59.08)	**64.09** (60)	—
Spanish		—	62.16 (60.46)	61.33 (57.71)	61.85 (61.96)	—
Czech		—	63.10 (59.58)	60.96 (56.82)	62.88 (58.5)	—
All		—	63.35 (61.02)	60.21 (58.46)	61.33 (59.45)	—

5 Discussion

The detection of PD-induced dysarthria using machine learning has advanced significantly in recent years. However, training reliable models requires extensive data, which is often challenging, costly, and time-consuming to collect in clinical settings. Current experiments confirm previous studies [1,15] showing that diadochokinetic exercises and sustained phonation tasks can be used to effectively detect dysarthric speech in PD. In addition, current experiments indicate that TL, when coupled with the acoustic features utilized, presents a promising solution to address data sparsity while benefiting from a feature space that remains largely interpretable to clinicians.

As observed in various tasks [35], we find here that larger datasets do not always guarantee better outcomes; instead, the quality of the dataset is the decisive factor. If size were the sole determinant, the mPower dataset—significantly

larger than the others—would consistently outperform smaller datasets as pre-training data. However, it often produced suboptimal results, with smaller datasets like Italian outperforming it. This underscores the nuanced role of dataset size in TL, where factors such as linguistic proximity, separability of pathological groups, and recording conditions often outweigh sheer volume.

Linguistic proximity plays a crucial role in the success of TL, as observed across many tasks [36], with closely related languages exhibiting enhanced transferability. This is particularly evident in the performance of Spanish and Italian, both Romance languages, where TL demonstrates notable success in both directions.

Additionally, in line with [19], it is observed that datasets more easily separable, namely Spanish and Italian, tend to perform better as pre-training data in TL scenarios. In this context, separability refers to how easily data points can be distinguished and correctly assigned to their respective categories—typically evident through high baseline model performance in monolingual classification. This suggests that the intrinsic structure of datasets plays a critical role in facilitating effective knowledge transfer. The clearer distinctions within datasets enable models to extract and generalize patterns more efficiently, resulting in better performance when transferring knowledge to new languages. Other than linguistic proximity, the bidirectional success of TL between Spanish and Italian could be explained by dataset separability. Italian is the best-performing pre-training language for Spanish because it offers data more easily separable compared to Czech and American English. In fact, the inherent separability of the dataset may explain why knowledge transfer, while bidirectional, is more successful from Italian to Spanish than in the reverse direction, as Italian is more easily separable than Spanish.

In Czech, TL outperformed monolingual models in all tests. Italian and Spanish proved to be the most effective pre-training dataset for Czech. Similarly, for American English, Italian emerged as the best pre-training language across all conducted tests. Separability is the most likely explanation. This suggests that the inherent characteristics of these datasets, rather than linguistic similarity, drives their success in boosting performance for Czech and American English.

The separability of the dataset was significantly improved when training data was arranged sequentially—beginning with HC and then progressing to patients—compared to either random shuffling or ordering with patients before HC. Notably, this HC-to-patient scheduling consistently outperformed random shuffling across all tested configurations. We attribute this improvement to the model's ability to first establish a stable baseline from the presumed lower variability in healthy speech, which facilitates the subsequent learning of more complex pathological patterns. Importantly, this ordering strategy enhances performance without requiring domain knowledge or access to clinical severity scores. As such, it offers a simple yet effective approach that is especially advantageous in clinical settings where detailed annotations (e.g., H&Y or UPDRS scores) may not be readily available.

Despite technological advancements, smartphone microphones still fall short of professional-grade quality used in controlled data collection. TL models pre-trained or fine-tuned on American English data showed only modest improve-

ments, largely due to data quality limitations rather than flaws in the TL approach itself. Unlike the Spanish dataset, the lack of complementary data collected in ideal acoustic conditions limits the models' ability to generalize. Notably, higher accuracies achieved with Spanish data suggest that while noisy, real-world data can improve robustness, when used as part of training, relying solely on it poses significant challenges.

Pre-training on multiple languages and fine-tuning on just one often results in suboptimal performance due to the conflicting features learned from linguistically diverse datasets. When multiple datasets are combined, the key factors that previously contributed to successful transfer learning—such as linguistic proximity, high data separability, and consistent recording quality—are weakened, reducing the effectiveness of the learned representations.

Lastly, models trained on DDK-derived general acoustic features consistently demonstrate the highest predictive performance, both in monolingual and transfer learning scenarios. Following this, general acoustic features extracted from SPA also show good predictive performance and cross-linguistic transferability, although they are slightly less effective than their DDK-derived counterparts. In contrast, models trained solely on SPA-specific acoustic features tend to perform worse, often exhibiting lower balanced accuracies. Kruskal-Wallis tests showed that general acoustic features—whether from DDK or SPA—contained a greater number of statistically significant features capable of distinguishing between HC and individuals with PD. Meanwhile, only a small subset of SPA-specific features reached statistical significance. Given their higher transferability, broader applicability across tasks, and higher interpretability, general acoustic features—particularly those derived from DDK—are better suited for clinical applications.

6 Conclusions

The primary objectives of this research were threefold. First, it evaluated the transferabillity of various tasks and the features derived from them in dysarthria detection. Second, it sought to verify the effectiveness of TL in this task, leveraging data from four languages. Third, evaluate CL strategies agnostic to domain knowledge.

DDK-derived features showed superior predictive capability and transferabillity compared to those derived from sustained phonation. TL offers a promising solution to data scarcity in clinical machine learning by leveraging knowledge from various languages to develop more robust models. Its success depends on factors like linguistic proximity, the initial separability of pre-training data, and recording conditions. Notably, TL performance improved significantly when data was scheduled sequentially from HC to patients rather than randomly shuffled or reversed. This approach is particularly advantageous as it does not require domain knowledge.

Looking forward, further studies should continue exploring additional interpretable speech-derived features beyond those used in the current research (e.g. vowel space related features). Expanding the set of clinically meaningful features could further enhance model robustness and improve dysarthria detection in Parkinson's disease.

A key limitation of this research is the lack of a broader range of languages, particularly non-Indo-European ones. Expanding language diversity in future work is crucial for improving our understanding of dysarthria in Parkinson's disease and for further exploring the potential of transfer learning to enhance performance.

In conclusion, acoustic features extracted from diadochokinetic exercises and sustained phonation tasks can be effectively used to discriminate healthy from dysarthric speech. More importantly, the implementation of transfer learning in this research, through language-based knowledge transfer, demonstrates its potential to address critical gaps in data availability in Parkinson's disease and related conditions.

Acknowledgments. This work is part of A.R.'s thesis research and was made possible through the support of the Onassis Foundation scholarship program, which funded A.R.'s studies at Saarland University.

JR received grant funding from The Czech Ministry of Health (NW24-04-00211) and the National Institute for Neurological Research (Programme EXCELES, ID Project No. LX22NPO5107), funded by the European Union—Next Generation EU.

A Features

Table 3. Summary of all acoustic features used in this study. The table lists each feature's name, a brief definition, and its category. Features are divided into SPA-specific acoustic features, which are extracted only from the sustained phonation of /a/, and general acoustic features, which are extracted from both the SPA and diadochokinetic tasks.

Feature	Definition	Category
first loudness drop time	The first point in time (in seconds) where the loudness drops by 16.66% of the loudness range and at least 4dB. If there is no such drop, the duration of the speech signal is returned.	SPA-specific
spectral contrast mean	Average energy difference between spectral peaks and valleys across sub-bands.	SPA-specific
spectral contrast sd	Standard deviation of energy differences between spectral peaks and valleys.	SPA-specific
spectral flatness mean	Average ratio of geometric to arithmetic mean of the power spectrum; indicates noisiness.	SPA-specific
spectral flatness sd	Standard deviation of spectral flatness; reflects variation in tonality vs. noisiness.	SPA-specific

(*continued*)

Table 3. (*continued*)

Feature	Definition	Category
spectral rolloff mean	Average frequency below which 85% of the spectral energy is concentrated.	SPA-specific
spectral rolloff sd	Standard deviation of roll-off frequency across time; indicates variability in spectral high-end content.	SPA-specific
tonnetz fifth x mean	Mean x-coordinate of tonal centroid for the perfect fifth interval.	SPA-specific
tonnetz fifth x sd	Standard deviation of the x-coordinate for the perfect fifth interval.	SPA-specific
tonnetz fifth y mean	Mean y-coordinate of tonal centroid for the perfect fifth interval.	SPA-specific
tonnetz fifth y sd	Standard deviation of the y-coordinate for the perfect fifth interval.	SPA-specific
tonnetz major third x mean	Mean x-coordinate of tonal centroid for the major third interval.	SPA-specific
tonnetz major third x sd	Standard deviation of the x-coordinate for the major third interval.	SPA-specific
tonnetz major third y mean	Mean y-coordinate of tonal centroid for the major third interval.	SPA-specific
tonnetz major third y sd	Standard deviation of the y-coordinate for the major third interval.	SPA-specific
tonnetz minor third x mean	Mean x-coordinate of tonal centroid for the minor third interval.	SPA-specific
tonnetz minor third x sd	Standard deviation of the x-coordinate for the minor third interval.	SPA-specific
tonnetz minor third y mean	Mean y-coordinate of tonal centroid for the minor third interval.	SPA-specific
tonnetz minor third y sd	Standard deviation of the y-coordinate for the minor third interval.	SPA-specific
vowel duration	Duration of the spoken vowel in seconds.	SPA-specific
average mfccs 1	Average of the 1st MFCC, reflecting overall spectral slope or broadband energy.	General
average mfccs 2	Average of the 2nd MFCC, interpreted as a low-to-high frequency energy ratio.	General
average mfccs 3	Average of the 3rd MFCC, interpretable as a mid-frequency to (low+high) frequency energy ratio.	General
average mfccs 4	Average of the 4th MFCC, representing the energy ratio between two frequency bands and two slightly higher bands.	General
alpha ratio mean	Mean ratio of summed energy in the 50âĂŞ1000 Hz and 1âĂŞ5 kHz.	General
alpha ratio sd	Standard deviation of the ratio of summed energy in the 50âĂŞ1000 Hz and 1âĂŞ5 kHz.	General
apq3 shimmer	Three-point Amplitude Perturbation Quotient—average absolute difference between the amplitude of a period and the mean amplitude of its two neighbors, divided by the average amplitude.	General

(*continued*)

Table 3. (*continued*)

Feature	Definition	Category
apq5 shimmer	Five-point Amplitude Perturbation Quotient—average absolute difference between the amplitude of a period and the mean amplitude of it and its four closest neighbors, divided by the average amplitude.	General
apq11 shimmer	Eleven-point Amplitude Perturbation Quotient—average absolute difference between the amplitude of a period and the mean amplitude of it and its ten closest neighbors, divided by the average amplitude. A value above 3.070% is considered pathological (MDVP standard).	General
dda shimmer	Average absolute difference between consecutive differences in the amplitudes of successive periods.	General
ddp jitter	Average absolute difference between consecutive differences in durations of successive periods, divided by the average period. Reflects rapid pitch variability (jitter).	General
duration	Duration in seconds.	General
f1 bandwidth mean	Mean bandwidth of the first formant.	General
f1 bandwidth sd	Standard deviation of the first formant bandwidth.	General
f1 frequency mean	Mean center frequency of the first formant (F1).	General
f1 frequency sd	Standard deviation of F1 center frequency.	General
f1 relative energy mean	Mean amplitude of the spectral envelope at F1 relative to the spectral F0 peak.	General
f1 relative energy sd	Standard deviation of amplitude at F1 relative to the spectral F0 peak.	General
f2 bandwidth mean	Mean bandwidth of the second formant (F2).	General
f2 bandwidth sd	Standard deviation of the second formant bandwidth.	General
f2 frequency mean	Mean center frequency of the second formant (F2).	General
f2 frequency sd	Standard deviation of F2 center frequency.	General
f2 relative energy mean	Mean amplitude of the spectral envelope at F2 relative to the spectral F0 peak.	General
f2 relative energy sd	Standard deviation of amplitude at F2 relative to the spectral F0 peak.	General
f3 bandwidth mean	Mean bandwidth of the third formant (F3).	General
f3 bandwidth sd	Standard deviation of the third formant bandwidth.	General
f3 frequency mean	Mean center frequency of the third formant (F3).	General
f3 frequency sd	Standard deviation of F3 center frequency.	General
f3 relative energy mean	Mean amplitude of the spectral envelope at F3 relative to the spectral F0 peak.	General
f3 relative energy sd	Standard deviation of amplitude at F3 relative to the spectral F0 peak.	General
h1 a3 harmonic difference mean	Mean ratio of energy of the first F0 harmonic (H1) to the highest harmonic in the third formant region (A3).	General
h1 a3 harmonic difference sd	Standard deviation of H1 to A3 energy ratio.	General
h1 h2 harmonic difference mean	Mean ratio of energy of the first F0 harmonic (H1) to the second F0 harmonic (H2).	General

(*continued*)

Table 3. (*continued*)

Feature	Definition	Category
h1 h2 harmonic difference sd	Standard deviation of H1 to H2 energy ratio.	General
hammarberg index mean	Mean ratio of the strongest spectral peak in 0âĂŞ2 kHz to the strongest peak in 2âĂŞ5 kHz.	General
hammarberg index sd	Standard deviation of the ratio of strongest spectral peaks in 0âĂŞ2 kHz vs. 2âĂŞ5 kHz.	General
hnr mean	Mean ratio of the strongest spectral peak in 0âĂŞ2 kHz to the strongest peak in 2âĂŞ5 kHz.	General
hnr sd	Standard deviation of the harmonic-to-noise energy ratio.	General
local jitter	Average timing variability between pitch periods within a segment of a signal, expressed in seconds.	General
local absolute jitter	Mean absolute deviation in timing between pitch periods within a segment, in seconds.	General
local dB shimmer	Average difference in amplitude (in dB) between consecutive F0 periods.	General
local shimmer	Average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude; reflects cycle-to-cycle amplitude variability.	General
loudness mean	Mean loudness in decibels during voiced intervals.	General
loudness sd	Standard deviation of loudness in decibels during voiced intervals.	General
ppq5 jitter	Five-point Period Perturbation Quotient—average absolute difference between a period and the average of it and its four nearest neighbors, divided by the average period.	General
rap jitter	Relative Average Perturbation—average absolute difference between a period and the average of it and its two neighbors, divided by the average period.	General
rate loudness peaks	Number of loudness peaks per second in voiced intervals.	General
spectral slope 0 500 mean	Mean linear slope of the log power spectrum in the 0âĂŞ500 Hz band.	General
spectral slope 0 500 sd	Standard deviation of spectral slope in the 0âĂŞ500 Hz band.	General
spectral slope 500 1500 mean	Mean linear slope of the log power spectrum in the 500âĂŞ1500 Hz band.	General
spectral slope 500 1500 sd	Standard deviation of spectral slope in the 500âĂŞ1500 Hz band.	General
speech ratio	Percentage of the audio signal identified as speech.	General
utterance durations sum	Total duration of all detected utterances.	General
utterance durations mean	Mean duration of detected utterances.	General
vocal tremor	Frequency of the most intense low-frequency modulation in the fundamental frequency.	General
pitch max	Maximum perceived frequency of a sound, corresponding to the rate of vibrations of the sound wave.	General

(*continued*)

Table 3. (*continued*)

Feature	Definition	Category
pitch min	Minimum perceived frequency of a sound, corresponding to the rate of vibrations of the sound wave.	General
pitch mean	Average perceived frequency of a sound, corresponding to the rate of vibrations of the sound wave.	General
pitch std	Standard deviation of the perceived frequency of a sound, corresponding to the rate of vibrations of the sound wave.	General
pitch range	Difference between the minimum and maximum perceived frequencies of a sound, corresponding to the rate of vibrations of the sound wave.	General
pause rate	Total length of pauses divided by the total length of speech (including pauses).	General
number of pauses	Number of pauses between speech segments based on voiced intervals.	General

References

1. Gündüz, H.: deep learning-based Parkinson's disease classification using vocal feature sets. IEEE Access **7**, 115540–115551 (2019). https://doi.org/10.1109/ACCESS.2019.2936564
2. Tsanas, A., Little, M.A., McSharry, P.E., Spielman, J., Ramig, L.O.: Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans. Biomed. Eng. **59**, 1264–1271 (2012)
3. Ho, A.K., Iansek, R., Marigliani, C., Bradshaw, J.L., Gates, S.: Speech impairment in a large sample of patients with Parkinson's disease. Behav. Neurol. **11**, 131–137 (1998). https://doi.org/10.1155/1999/327643
4. Pinto, S., Chan, A., Guimarães, I., Rothe-Neves, R., Sadat, J.: A cross-linguistic perspective to the study of dysarthria in Parkinson's disease. J. Phonetics **64**, 156–167 (2017). https://doi.org/10.1016/j.wocn.2017.01.009
5. Magee, M., Copland, D., Vogel, A.P.: Motor speech and non-motor language endophenotypes of Parkinson's disease. Expert Rev. Neurother. **19**, 1191–1200 (2019)
6. Chu, S.Y., Tan, C.L.: Subjective self-rated speech intelligibility and quality of life in patients with Parkinson's disease in a Malaysian sample. Open Public Health J. **12**, 485–493 (2018). https://doi.org/10.2174/1874944501811010485
7. Vojtech, J.M., Van Stan, J.H., Mehta, D.D., Hillman, R.E.: Effects of age and Parkinson's Disease on the relationship between vocal fold Abductory kinematics and relative fundamental frequency. J. Voice **38**(5), 1008–1022 (2024). https://doi.org/10.1016/j.jvoice.2023.01.012
8. Williams, D.: Why So slow? models of parkinsonian bradykinesia. Nat. Rev. Neurosci. **25**, 573–586 (2024). https://doi.org/10.1038/s41583-024-00830-0
9. Orozco-Arroyave, J.R., Hönig, F., Arias-Londoño, J.D., Vargas-Bonilla, J.F., Skodda, S., Rusz, J., Nöth, E.: Voiced/unvoiced transitions in speech as a poten-

tial bio-marker to detect Parkinson's disease. In: Proc. Interspeech 2015, pp. 95–99 (2015). https://doi.org/10.21437/Interspeech.2015-34
10. Darley, F.L., Aronson, A.E., Brown, J.R.: Differential diagnostic patterns of dysarthria. J. Speech Hear. Res. **12**(2), 246–269 (1969). https://doi.org/10.1044/jshr.1202.246
11. Yang, S., et al.: The physical significance of acoustic parameters and its clinical significance of dysarthria in Parkinson's disease. Sci. Rep. **10** (2020). https://doi.org/10.1038/s41598-020-68754-0
12. van der Graaff, M.M., et al.: Clinical identification of dysarthria types among neurologists, residents in neurology and speech therapists. Eur. Neurol. **61**, 295–300 (2009)
13. Zyski, B.J., Weisiger, B.E.: Identification of dysarthria types based on perceptual analysis. J. Commun. Disord. **20**(5), 367–378 (1987). https://doi.org/10.1016/0021-9924(87)90025-6
14. Rizzo, G., Copetti, M., Arcuti, S., Martino, D., Fontana, A., Logroscino, G.: Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis. Neurology **86**(6), 566–576 (2016). https://doi.org/10.1212/WNL.0000000000002350
15. Hosseini-Kivanani, N., Vásquez-Correa, J.C., Schommer, C., Nöth, E.: Exploring the use of phonological features for Parkinson's disease detection. In: Proceedings 20th International Congress of the Phonetic Sciences (ICPhS 2023), pp. 3897–3901. Prague Congress Center, Czech Republic (2023)
16. Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., Azim, M.: Transfer learning: a friendly introduction. J. Big Data **9** (2022). https://doi.org/10.1186/s40537-022-00652-w
17. Simpson, A.P., Ericsdotter Nordgren, C.: Sex-specific differences in f0 and vowel space. In: Proceedings of the XVIth International Congress of Phonetic Sciences (ICPhS 2007), pp. 933–936. Saarbrücken, Germany (2007)
18. Fougeron, C., Guitard-Ivent, F., Delvaux, V.: Multi-dimensional variation in adult speech as a function of age. Languages **6**(4), 176 (2021). https://doi.org/10.3390/languages6040176
19. Vásquez-Correa, J.C., et al.: Convolutional neural networks and a transfer learning strategy to classify Parkinson's disease from speech in three different languages. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: CIARP 2019, pp. 697–706. Springer (2019)
20. Vásquez-Correa, J.C., et al.: Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages. Pattern Recogn. Lett. **150**, 272–279 (2021). https://doi.org/10.1016/j.patrec.2021.04.011
21. Soviany, P., Ionescu, R.T., Rota, P., Sebe, N.: Curriculum learning: a survey. Int. J. Comput. Vis. **130**, 1526–1565 (2022). https://doi.org/10.1007/s11263-022-01611-x
22. Dhinagar, N.J., et al.: Curriculum based multi-task learning for Parkinson's disease detection. In: Proceedings IEEE 20th International Symposium on Biomedical Imaging (ISBI 2023) (2023)
23. Tekindor, A., Aydın, E.: Feature selection improves speech based Parkinson's disease detection performance. In: Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 1: BIOSIGNALS, pp. 726–732. SciTePress (2024). https://doi.org/10.5220/0012347300003657

24. Harte, C., Sandler, M., Gasser, M.: Detecting harmonic change in musical audio. In: Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia, pp. 21–26. ACM Press, Santa Barbara (2006). https://doi.org/10.1145/1178723.1178727
25. Ladefoged, P., Johnson, K.: A Course in Phonetics. Cengage Learning (2011). https://books.google.de/books?id=FjLc1XtqJUUC
26. Dichter, B.K., Breshears, J.D., Leonard, M.K., Chang, E.F.: The control of vocal pitch in human laryngeal motor cortex. Cell **174**(1), 21-31.e9 (2018). https://doi.org/10.1016/j.cell.2018.05.016
27. Kováč, D., et al.: Exploring language-independent digital speech biomarkers of hypokinetic dysarthria (2022). https://doi.org/10.1101/2022.10.24.22281459
28. Tracey, B., Volfson, D., Glass, J., et al.: Towards interpretable speech biomarkers: exploring MFCCs. Sci. Rep. **13**, 22787 (2023). https://doi.org/10.1038/s41598-023-49352-2
29. Orozco-Arroyave, J.R., Arias-Londoño, J.D., Vargas-Bonilla, J.F., González-Rátiva, M.C., Nöth, E.: New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In: Proceedings 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 342–347. ELRA, Reykjavik (2014)
30. La Quatra, M., Turco, M.F., Svendsen, T., Salvi, G., Orozco-Arroyave, J.R., Siniscalchi, S.M.: Exploiting foundation models and speech enhancement for parkinson's disease detection from speech in real-world operative conditions. arXiv preprint arXiv:2406.16128 (2024)
31. Bot, B.M., et al.: The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci. Data **3**, 160011 (2016). https://doi.org/10.1038/sdata.2016.11
32. Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D., Girardi, F.: Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system. IEEE Access **5**, 22199–22208 (2017). https://doi.org/10.1109/ACCESS.2017.2762475
33. Tröger, J., et al.: An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders. Front. Digit. Health **6** (2024). https://doi.org/10.3389/fdgth.2024.1440986
34. Chen, Y., Xu, C., Liang, C., Tao, Y., Shi, C.: Speech-based clinical depression screening: an empirical study. arXiv preprint arXiv:2406.03510 (2024)
35. Agnew, C., Scanlan, A., Denny, P., Grua, E., van de Ven, P., Eising, C.: Annotation quality versus quantity for object detection and instance segmentation. IEEE Access **12**, 140958–140977 (2024). https://doi.org/10.1109/ACCESS.2024.3467008
36. Bardet, A., Bougares, F., Barrault, L.: A study on multilingual transfer learning in neural machine translation: finding the balance between languages. In: Martín-Vide, C., Purver, M., Pollak, S. (eds.) SLSP 2019. LNCS (LNAI), vol. 11816, pp. 59–70. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31372-2_5
37. Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G.: Training medical image analysis systems like radiologists. arXiv preprint arXiv:1805.10884 (2019)