IDIAP RESEARCH REPORT

SYLLABLE-LEVEL FEATURES FOR SPEECH PATHOLOGY DETECTION: A CASE STUDY OF PARKINSON'S DISEASE

Sevada Hovsepyan[a] Mathew Magimai-Doss

Idiap-RR-02-2026

MAY 2026

[a]Idiap

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Syllable-Level Features for Speech Pathology Detection: A Case Study of Parkinson's Disease

Sevada Hovsepyan[a] and Mathew Magimai.-Doss

Idiap Research Institute, Martigny, CH-1920, Switzerland

In this paper we explore newly proposed syllable level feature extraction for speech pathology detection: focusing on the Parkinson's disease detection from speech. The method is inspired by the spectro-temporal representations used in the neuro-computational models of speech perception, and simply represents a standardized representation of the frequency content of the syllable-like segments. In this study, we demonstrate that this type of representation is powerful enough to successfully detect various speech pathologies from speech samples, independent of speech type, language, and dataset. Furthermore, our analyses show that these representations lead to explainable and interpretable results, and that changes in pathologies are related to both system- and source-related changes in speech production. Overall, syllable-level features suggest themselves as a simple but robust and explainable approach to understand physiological changes in speech production due to different pathologies, as well as to detect those pathologies from speech.

[a]sevada.hovsepyan@idiap.ch

I. INTRODUCTION

Parkinson's disease (PD) stands out as one of the most prevalent neurodegenerative disorders, affecting millions of individuals worldwide. PD impacts both motor (Mazzoni *et al.*, 2012) and non-motor systems, with symptoms ranging from motor planning difficulties, tremors, and bradykinesia to cognitive impairments, neural degeneration, gastrointestinal issues, and neuropsychiatric conditions. Early diagnosis and intervention are crucial for more effective treatment and management of the disease. In recent years, numerous studies have focused on developing non-invasive methods for detecting PD, such as gait and handwriting analysis, which achieve high detection accuracy by assessing motor deficiencies. Blood analysis and cognitive tasks are also employed to evaluate cognitive impairments. However, these methods often require special equipment and settings, and most importantly, rarely capture both cognitive and motor aspects simultaneously, which would be beneficial for a more comprehensive assessment of the disorder.

From this perspective, speech-based PD detection is unique, as it can capture both motor-related impairments (e.g., imprecise articulation) and cognitive impairments (e.g., language planning). Moreover, speech based pathology detection are often easy to deploy, scalable, and not necessarily require special equipment (e.g. phone's microphone can be sufficient). Research in this area has evolved from expert-crafted, hand-engineered features, such as vowel triangles and formant patterns, hypothesized to reflect motor impairments in vocal cord vibration and articulatory control (Karlsson *et al.*, 2020; Michaelis *et al.*, 1998), to large consortia of handcrafted feature sets like ComPARE (Schuller *et al.*, 2010, 2013) and

eGeMAPS (Eyben *et al.*, 2016). The latter, provided more data-driven approach and often operated under "brute-force" paradigm, by capturing a broad spectrum of paralinguistic aspects of speech, though not necessarily relevant to PD. More recently, deep learning approaches, particularly convolutional neural networks (CNNs) and adaptive architectures, have achieved strong empirical performance (Escobar-Grisales *et al.*, 2023; van Gelderen and Tejedor-García, 2024). However, these end-to-end models, while powerful, are typically generic and lack interpretability, a critical factor for clinical adoption (Ananthanarayanan *et al.*, 2025).

Current approaches predominantly focus on acoustic deficits in PD speech, which are largely attributed to motor impairments. Yet, PD, being a neurodegenertaive disease, also affects neural functioning and cognitive processing, that can manifest in produced speech. This oversight is notable given that speech production and perception are tightly coupled processes, shaped not only by motor control and articulation but also by underlying neural activity and linguistic content.

To address this gap, we recently proposed a novel approach for PD speech analysis (Hovsepyan and Magimai.-Doss, 2024) based on the acoustic representations rooted in neurophysiologically plausible generative models of speech (Hovsepyan *et al.*, 2020; Nabé *et al.*, 2021). These models posit that speech is parsed into syllable-like units during perception, and that disruptions in speech production due to PD may leave detectable traces in these perceptual representations. By focusing on syllable-level spectrotemporal patterns, our method aims to capture both motor and neural aspects of PD-related speech changes, offering a more holistic and interpretable approach to pathology detection.

In this study, our objective is to extensively explore and refine our proposed SLF for the detection of Parkinson's disease by speech. We test our method on various speech types and datasets, including the PC-GITA dataset (Orozco *et al.*, 2014) and the Dutch Corpus of Pathological and Normal Speech (COPAS) (Martens *et al.*, 2011). Our objectives are to determine the robustness and generalizability of our feature construction method and to provide insights into the physiological mechanisms underlying PD-related speech changes.

The remainder of the manuscript is organized as follows. In the Background, we provide a quick overview of the aforementioned generative models and delve deeply into the proposed SLF approach. In the Methods section, we detail the datasets used in the study, outline the construction of syllable-level features, and describe the methods used for classification and evaluation of our results. In the Results section, we present our classification outcomes and compare them with baseline feature sets. In the final section, we discuss our findings and outline future directions.

II. BACKGROUND

In this section, we present the primary motivation for examining syllables as the basis for feature extraction in Parkinson's disease speech analysis. We then briefly describe the main findings from our pilot investigation using syllable-level features and outline the specific objectives of this manuscript.

A. Syllables as speech building blocks

Syllables are linguistic units widely recognized as fundamental building blocks of speech, occupying a central role at the intersection of acoustic, linguistic, cognitive, and neural aspects of speech production and perception. From a cognitive perspective, syllables serve as discrete planning units for speech production and as perceptual chunks during speech perception. From a neurophysiological standpoint, it is hypothesized that the brain employs dedicated mechanisms, such as theta oscillations (4–8 Hz), to segment the continuous speech stream into syllable-like, discrete units during perception. These multifaceted roles make syllables ideal candidates for investigating how neurodegenerative diseases such as PD alter speech. Specifically, syllable-level analysis can reveal the sources of deficiencies at multiple levels, extending beyond motor dysfunction to include disruptions in neural, cognitive, and linguistic processing.

B. Speech Production-Perception Cycle and Meurocomputational Models of SPeech Perception

In the neuroscience of speech processing, a seminal framework is the analysis-by-synthesis approach (Bever and Poeppel, 2010; Halle and Stevens, 1962; Nair *et al.*, 2008), a theory rooted in the motor theory of speech perception (Dogonasheva *et al.*, 2025; Hovsepyan *et al.*, 2023; Hyafil *et al.*, 2015; Nabé *et al.*, 2021). This framework posits that speech perception involves the internal generation of motor commands associated with incoming acoustic signals. Contemporary developments, such as predictive coding and the Free Energy Principle

(FEP) (Buckley *et al.*, 2017; Friston, 2010; Friston and Kiebel, 2009), extend this idea, proposing that the brain functions as a statistical, generative system. It anticipates sensory input based on accumulated internal knowledge of the external world and continuously synthesizes expected speech patterns to minimize prediction error and infer incoming sesnory signal, such as speech.

In recent years, neurocomputational models of speech perception have been developed that integrate neural oscillations, syllable-centric processing, and hierarchical predictive coding (Dogonasheva *et al.*, 2025; Hovsepyan *et al.*, 2023; Hyafil *et al.*, 2015; Nabé *et al.*, 2021). These generative models use parameterized spectrotemporal patterns of syllables to simulate expected acoustic signals and refine perceptions by minimizing prediction errors. While primarily designed to explore brain function during speech perception, these models offer a unique perspective for analyzing deficits in speech production. Given that PD disrupts neural processing and motor execution of speech, these disruptions should also manifest in the spectrotemporal patterns of syllables. If these models can infer syllable identity from such patterns, then differences in these patterns should emerge between healthy and PD conditions.

Figure 1 illustrates this conceptual bridge: while traditional approaches analyze the acoustic waveform directly (a bottom-up process), our method leverages the structure of perceptual models to interpret production deficits (a top-down/bidirectional approach).

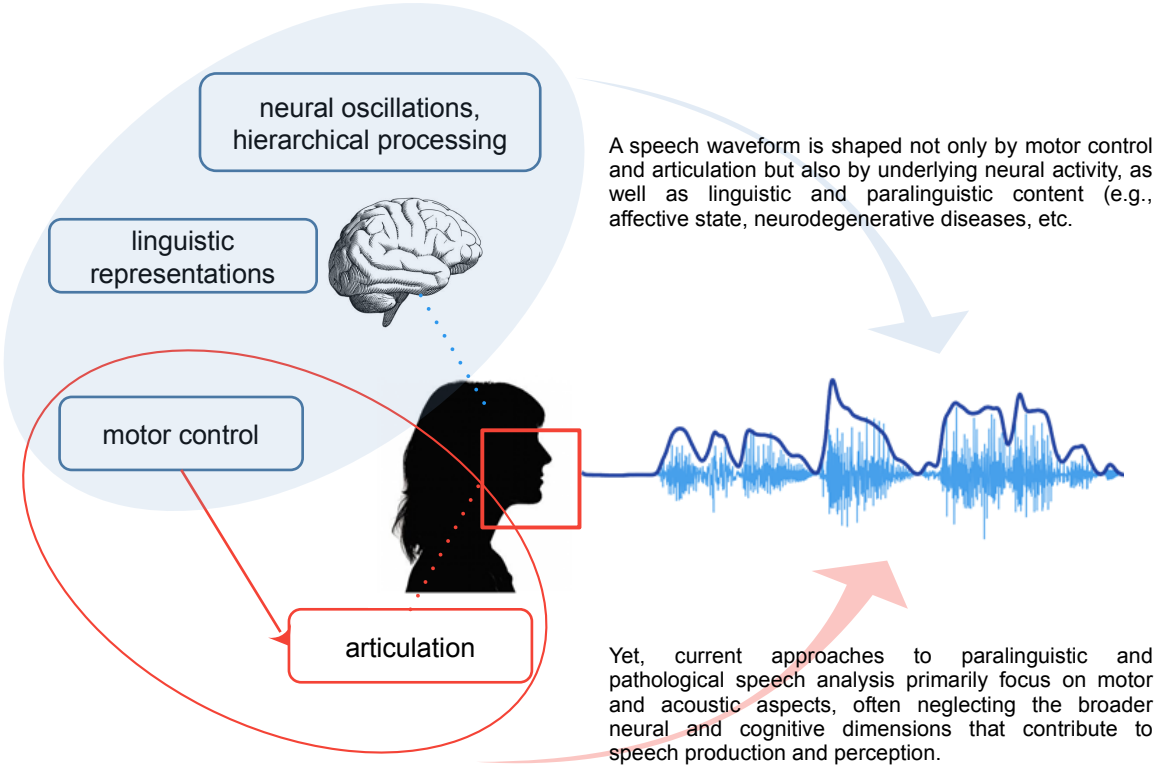

FIG. 1. Speech production and perception are tightly interlinked processes, shaped by motor control, articulatory mechanisms, neural activity (e.g., neural oscillations and hierarchical processing), and linguistic and paralinguistic factors (e.g., emotional tone and cognitive load). This interplay is particularly critical in neurodegenerative disorders, which often disrupt both the motor execution of speech (e.g., articulation) and the neural substrates underlying linguistic and paralinguistic functions (e.g., emotional prosody and cognitive planning). Current approaches to pathological speech analysis, however, predominantly focus on motor and acoustic deficits in the speech waveform. In this manuscript, we argue that valuable insights into pathological speech can also be gained by analyzing components relevant to speech perception.

C. Syllable Level Features for Parkinson's disease detection

Based on the ideas discussed in the previous two sections, that syllables serve as building blocks for speech at multiple levels (acoustic, neural, cognitive, linguistic) and that speech production deficits can be understood from a perception perspective, we recently proposed a feature engineering pipeline for PD speech analysis. The so-called syllable-level features (SLFs) are derived from the spectral representations of syllables used in neurocomputational models of speech perception. The procedure involves calculating a standardized representation of the spectrotemporal patterns of syllables (with a fixed number of frequency and temporal bins) and using these patterns (after flattening) as feature vectors for PD speech classification. In the next section, we detail the construction of SLFs, whereas this section summarizes the results of our pilot study (Hovsepyan and Magimai.-Doss, 2024).

Neurocomputational models aim to mimic brain activity and be neurophysiologically plausible; therefore, to calculate the spectrotemporal pattern of syllables or speech, they often employ a model of the cochlea. Following the same logic, we initially used the same model of the auditory periphery as the aforementioned models, but we also compared other approaches for spectrogram calculation, such as Mel-frequency cepstral coefficients (MFCC) or the Short-Time Fourier Transform (STFT). In addition to comparing different spectrogram calculation methods, we aimed to explore the optimal number of frequency channels used for SLF calculation (specifically, into how many frequency bins the spectrogram should be collapsed).

Our analysis of diadochokinetic (DDK) speech samples from the PC-GITA dataset revealed that the best classification performance was achieved using STFT decomposition, with performance improving as the number of frequency channels increased (we tested 6, 22, and 46 channels). We also examined whether the method of deriving syllable chunks affected classification performance, specifically comparing syllable onset-based segmentation (valleys in the sonority envelope) to syllable nucleus-based segmentation (peaks in the envelope). The analysis showed that, on average, nucleus-based chunking (peaks in the envelope) yielded better performance (83.23% vs. 80.1% AUROC for the 46-channel STFT case). Furthermore, our pilot analysis demonstrated that segmenting long DDK utterances into linguistically informed chunks (such as syllable onsets or nuclei) was more advantageous than using fixed-length chunks of comparable duration (AUROC = 71.85% for 25-50ms segments, AUROC = 71.48% for 25-450ms segments and AUROC = 71.32% for 300-600ms segments).

The results of the pilot study serve as a starting point for the more in-depth analysis performed in this manuscript. Based on our previous results, we focus on the STFT spectrogram of syllables and use syllable nuclei to chunk syllable-like segments of the analyzed speech signal. Beyond these parameters, the goal of this manuscript is to conduct a more intensive and in-depth analysis of syllable-level features. More specifically, we delve into how SLFs can be applied to speech types other than DDK utterances, what kind of acoustic information SLFs capture, and whether the method is applicable to other pathologies and/or languages.

III. METHODS

A. Datasets

In this study, we used two well-established datasets for pathological speech research: the "New Speech Corpus Database for Analysis of People Suffering from Parkinson's Disease" (PC-GITA) (Orozco *et al.*, 2014) and the Dutch Corpus of Pathological and Normal Speech (COPAS) (Martens *et al.*, 2011).

The PC-GITA dataset consists of recordings from 50 individuals with Parkinson's disease (PD) and 50 healthy controls (HC) (Orozco *et al.*, 2014). The PD patients were diagnosed by expert neurologists and were classified as having different levels of severity based on the Unified Parkinson's Disease Rating Scale (UPDRS) (Goetz *et al.*, 2008) and Hoehn and Yahr (H&Y) scales (Hoehn and Yahr, 1967). Both groups were matched by age and gender. The PD group consists of 25 males with a mean age of 62.2 $\pm$ 11.2 years and 25 females with a mean age of 60.1 $\pm$ 7.8 years. The HC group consists of 25 males with a mean age of 61.2 $\pm$ 11.3 years and 25 females with a mean age of 60.7 $\pm$ 7.7 years. The recordings were performed in a soundproof booth with a professional microphone, and the resulting audio files were sampled at 44,100 Hz with 16-bit resolution. The PC-GITA dataset includes various speech types produced by all speakers, such as sustained vowels, rapid syllable repetitions (diadochokinetic (DDK) utterances), monologues, and read text passages and sentences. Since this study focuses on syllable-level features, sustained vowels were excluded from the experiments, and only DDK, read text, monologue, and sentence utterance types were used.

In contrast, the COPAS dataset includes recordings of over 300 speakers performing various speech tasks, such as repeating sentences and reading texts (Martens *et al.*, 2011). The dataset includes recordings of healthy speakers and speakers with speech pathologies, such as hearing impairments, articulation disorders, and dysarthria. Recordings were performed in a quiet clinical setting and sampled at 16 kHz and 16-bit resolution. However, the number of recordings varies greatly depending on the type of pathology. Additionally, not all speakers performed all speech tasks; therefore, the number of recordings per pathology and task varies. For the current experiment, we only needed long-form speech tasks, such as sentences and reading texts. Therefore, we only used two pathologies: dysarthria and hearing impairment, as there were enough recordings for analysis.

B. Control feature sets

As mentioned in the Background section, we chose ComPARE (Schuller *et al.*, 2010) and eGeMAPS (Eyben *et al.*, 2016) feature sets as controls. These feature sets were specifically designed for paralinguistic challenges and have been utilized in various competitions over the years. The ComPARE feature set, for instance, was created for speech analysis challenges, including those at Interspeech (Schuller *et al.*, 2010, 2015, 2016, 2013, 2021), and contains a large number of features related to various speech characteristics such as pitch variations, formants, and harmonics.

eGeMAPS (Eyben *et al.*, 2016), on the other hand, is a derivative of the ComPARE feature set, where a handful of features are specifically selected for their importance in paralinguistic and clinical speech analysis, including PD. After several revisions, the eGeMAPS

feature set now includes 88 low-level descriptors, covering spectral, energy, and frequency aspects of speech. The SMILE toolbox (Eyben *et al.*, 2010), with default parameters, was used to extract eGeMAPS and ComPARE feature sets from both datasets at the utterance level, so that for each utterance, we have a fixed number of features suitable for standard classification procedures.

C. Syllable level features

Syllable-level features were extracted using a procedure that closely followed the steps described in a previous study (see (Hovsepyan and Magimai.-Doss, 2024) and Figure 2a). First, syllable nuclei timesteps were detected using the syllable onset detection algorithm proposed by (Räsänen *et al.*, 2018). The spectrogram of each utterance was then extracted using the short-time Fourier transform. Based on the syllable boundaries and STFT information, a standardized representation of each syllable was constructed by averaging the spectral energy across neighboring frequency channels and equally distributed temporal bins (see Figure 2b). Consequently, the resulting spectro-temporal representation maintained consistent temporal and frequency dimensions (64 frequency bins and 8 temporal bins in this study), irrespective of the original syllable's duration. Finally, the standardized representation was flattened into a feature vector for use in classification analyses. The details of each step are outlined below.

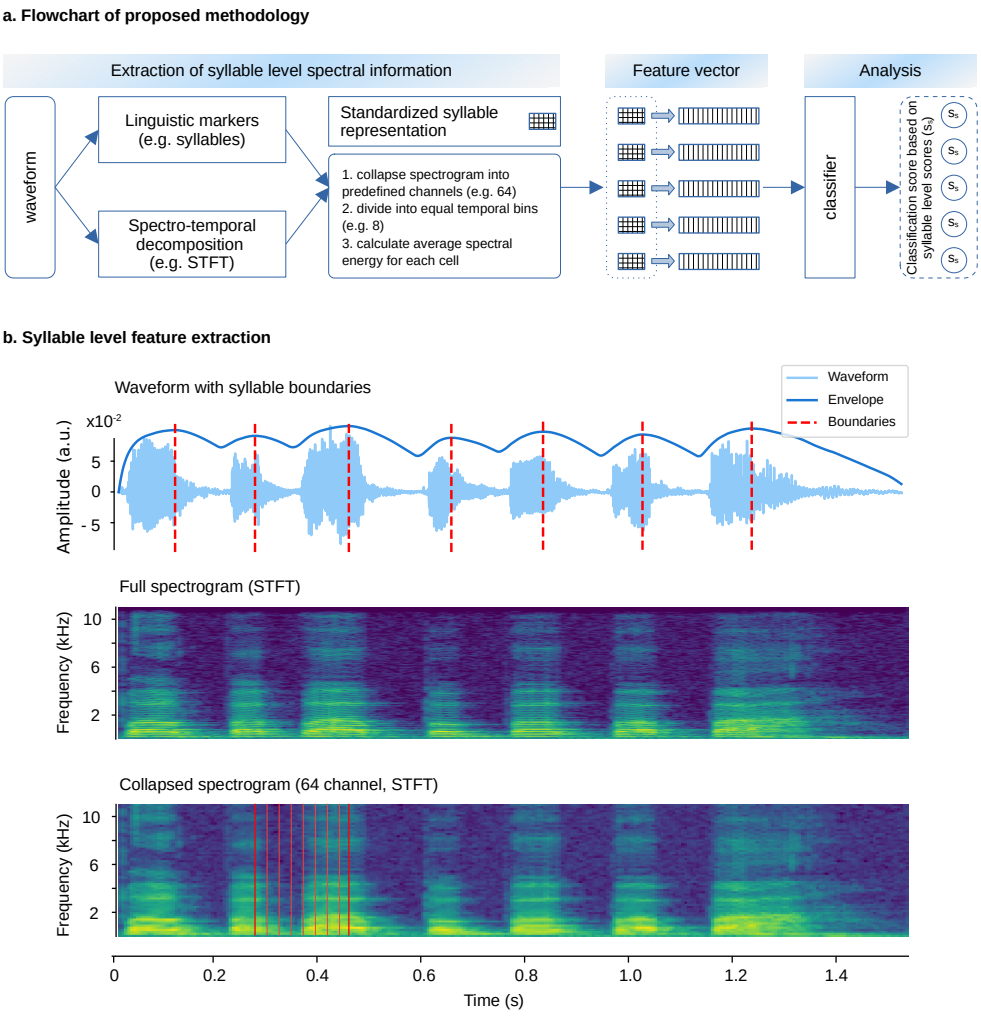

FIG. 2. Section **a** illustrates the pipeline for syllable-level feature extraction. For each speech waveform (top panel, section **b**), linguistic markers, such as syllable boundaries (dashed red lines) and the corresponding spectrogram (middle panel, section B), are first extracted. Next, a standardized syllable representation is computed: the spectrogram (bottom panel, section **b**) is collapsed into predefined frequency bins (e.g., 64 bins), and the syllable's duration is divided into fixed, equal temporal bins (e.g., 8 bins, marked by red vertical lines). The average spectral energy is then calculated for each frequency–temporal bin. As a result, each syllable is represented as a fixed-dimensional matrix (8×64), which is flattened into a feature vector for subsequent classification analysis.

1. Syllable boundaries

We used the algorithm proposed by (Räsänen *et al.*, 2018) for syllable boundary detection. The algorithm is inspired by the oscillatory mechanism that the brain uses to segment continuous speech into syllabic segments by entraining theta (4–8 Hz) oscillations to the speech signal. In their model, (Räsänen *et al.*, 2018) and his colleagues first extracted the sonority envelope of the speech signal by filtering the speech signal into sub-bands using gammatone filter banks and then combining them to obtain the sonority envelope. The resulting signal was then fed to a damped theta oscillator, which shifted the phase of the oscillator to align with the peaks in the sonority envelope, which are typically associated with syllable nuclei. Syllable boundaries were associated with oscillatory phase resets that occurred at the local minima (valleys) of the sonority envelope. Additionally, candidate peaks were filtered based on duration constraints (e.g., peaks closer than 50 ms were merged.). The algorithm's output is the time steps (in milliseconds) of the syllable boundaries (valleys) and nuclei (peaks). In this study, we used the information about syllable nuclei as syllable markers.

Additionally, for each speech type, we calculated the syllable duration distribution, defined as the duration between two consecutive syllable nuclei or peaks of the sonority envelope. For speech types that had more than one utterance per speaker (e.g., DDK utterances), the syllable duration distribution was based on syllable segments from all utterances. Furthermore, syllable duration distributions were approximated using Gaussian kernels, such that for each speech type, there was a Gaussian approximation for the respective syllable

duration distribution. This information was further used for control analysis, highlighting the importance of extracting features based on linguistic markers, such as syllables.

2. Spectrogram: Short Term Fourier Transform

During this study, we used the librosa library (McFee *et al.*, 2015) to calculate the power spectrogram with the short-term Fourier transform of each utterance. The following parameters were used throughout the simulations: 'nfft points' = 511 and 'hop length' = 128. Since the PC-GITA dataset (Orozco *et al.*, 2014) includes audio files sampled at 44.1 kHz, we loaded those files with the librosa default settings, resulting in a sampling rate of 22.05 kHz. For the COPAS dataset (Martens *et al.*, 2011), which is sampled at 16 kHz, we did not resample the audio signal during the STFT calculation. This yielded to 'hop length' of 23 ms and 30 ms for the PC-GITA and COPAS datasets, respectively.

3. Syllable based segmentation

We calculated the syllable-level features for each utterance based on the extracted power spectrogram and the corresponding syllable boundaries (nuclei timesteps), following the procedure below.

First, we averaged the power in adjacent frequency channels of the power spectrogram to collapse it into 64 frequency channels. We then divided the resulting collapsed spectrogram into syllable segments, with each segment consisting of the spectral information between two consecutive syllable nuclei.

Furthermore, each syllable segment was divided into eight equal temporal bins (red vertical lines in the bottom panel of Figure 2b), and the average activity of each frequency channel within each bin was calculated. Thus, each syllable segment is represented in a standardized way with sixsty-four frequency and eight temporal bins. Finally, we transformed the standardized spectrogram for each syllable into the decibel (db) scale, and then flattened it to obtain a syllable-level feature vector with dimensions of 1x(8x64=512).

4. Control segmentation

To assess whether segmentation based on linguistic markers, such as syllables, is critical for feature extraction, we extracted SLF-like representations using randomly segmented utterances. The procedure mirrored that used for SLF extraction, with one key difference: utterances were not segmented according to syllable markers. Instead, segment durations were randomly sampled from a Gaussian distribution, derived before, approximating the natural syllable duration distribution.

Next, each random segment, similar to the syllable segment, was divided into 8 equal parts, and the average power of each bin was calculated. The resulting random segment spectrogram, with a size of 64 frequency bins and 8 time bins, was transformed into a dB scale, and the resulting Random Segment Features (RSF) were calculated.

5. Formant based syllable level features

Original SLF construction, the spectrum of a syllable was binned into fixed freqeuncy bins to get the collapsed spectrogram with reduced number of frequency channels. We further

explored the benefits of dynamic frequency binning, e.g. aggregating spectral information across formant and foundamental frequency patterns. For this analysis, we made the following modifications to the previously described SLF extraction procedure. First, for each utterance, we extracted the power STFT and the syllable boundaries, along with the patterns of the fundamental frequency (F0) and the first four formants (F1-F4). Additionally, formant bandwidths were extracted, whereas for the fundamental frequency, a fixed bandwidth equal to a quarter of the F0 was used. The formant patterns (with their respective bandwidths) and the fundamental frequency patterns were extracted using the PRAAT (Boersma and Weenink, 2003) implementation in Python via the Parselmouth library (Jadoul *et al.*, 2018). Next, for each syllable-like segment, the average energy (spectrogram amplitude) around the pitch (F0 ± fixed bandwidth) and formant (F1-F4) patterns (± respective bandwidth) was derived. Each syllable segment was then divided into T=8 equal temporal bins, and the average energy associated with the F0 and formant patterns and their respective bandwidths was derived. This procedure results in a 5x8 spectrotemporal representation of a syllabic segment. Finally, the flattened representation (1x40) is used for the classification procedure, as in the earlier proposed SLFs.

D. Classification

Throughout the manuscript, we used a Support Vector Machine (SVM) with a cubic polynomial kernel for all classification. We used the standard parameters with a fixed random state from the *scikit-learn* Python package without hyperparameter tuning. Different classification protocols were used for the two datasets. For the PC-GITA dataset (Orozco *et al.*,

2014), which has balanced samples of healthy and pathological speech, we used the leave-one-sample-out (LOSO) protocol. For the COPAS dataset (Martens *et al.*, 2011), which has imbalanced samples of healthy and pathological speech, we used a stratified K-fold protocol (K = 4 for hearing impairment and K = 5 for dysarthria).

In each case, the classification results are reported based on the accumulated classification scores from each fold. The scores are obtained by either combining the classification scores for each syllable in the utterance (called syllable-level scores) or calculating the average of the syllable scores for each utterance, which leads to an utterance-level score.

E. Statistical analysis

We ran a linear mixed random effects model to test whether eGeMAPS features were statistically different between conditions. To do this, we constructed the linear model with fixed effects of eGeMAPS features and PD vs. HC condition, with speakers coded as a random effect. We then looked specifically at the interaction term between features and conditions and extracted which features were statistically significant ($p < 0.05$) across conditions (corrected for multiple comparisons using the false discovery rate).

IV. RESULTS

In this section we extensively test the proposed SLF method for speech pathology detection. We first extend our earlier report on PD detection from DDK speech utterances from PC-GITA dataset. Consequently we run classification analysis based on SLF on different speech types, such as text reading, monologues and sentences. Furthermore, we explore

what kind of information SLF capture and lastly we generalize the proposed method by testing it on COPAS dataset on different pathologies/languages.

A. Parkinson's disease detection from speech with syllable level features

We first expanded our earlier report on SLF's capacity to detect PD from speech by conducting classification analyses on additional speech samples and types from the PC-GITA dataset (Orozco *et al.*, 2014). The results are visualized in Figure 3, with detailed performance metrics provided in Supplementary Tables 1a and 1b for syllable- and sentence-level scores, respectively. For comparison, we also include AUC values derived from classification using established feature sets, such as eGeMAPS and ComPARE (Figure 4), as controls for each speech type. In Figure 4, each bar plot shows the average AUC for the respective speech type, for each feature set. Detailed performance metrics for control features sets can be found in Supplementary Tables 1c and 1d for eGeMAPS and ComPARE features respectively.

We extend our earlier report on SLF's capacity to detect PD from speech by conducting classification analyses on additional speech samples and types from the PC-GITA dataset (Orozco *et al.*, 2014). Results are visualized in Figure 3, with detailed performance metrics provided in Supplementary Tables 1a and 1b for syllable- and utterance-level scores, respectively.

For comparison, we also include AUC values obtained using established feature sets, eGeMAPS and ComPARE, as controls for each speech type (Figure 4). In Figure 4, each bar represents the mean AUC across participants for the respective speech type (grey dots

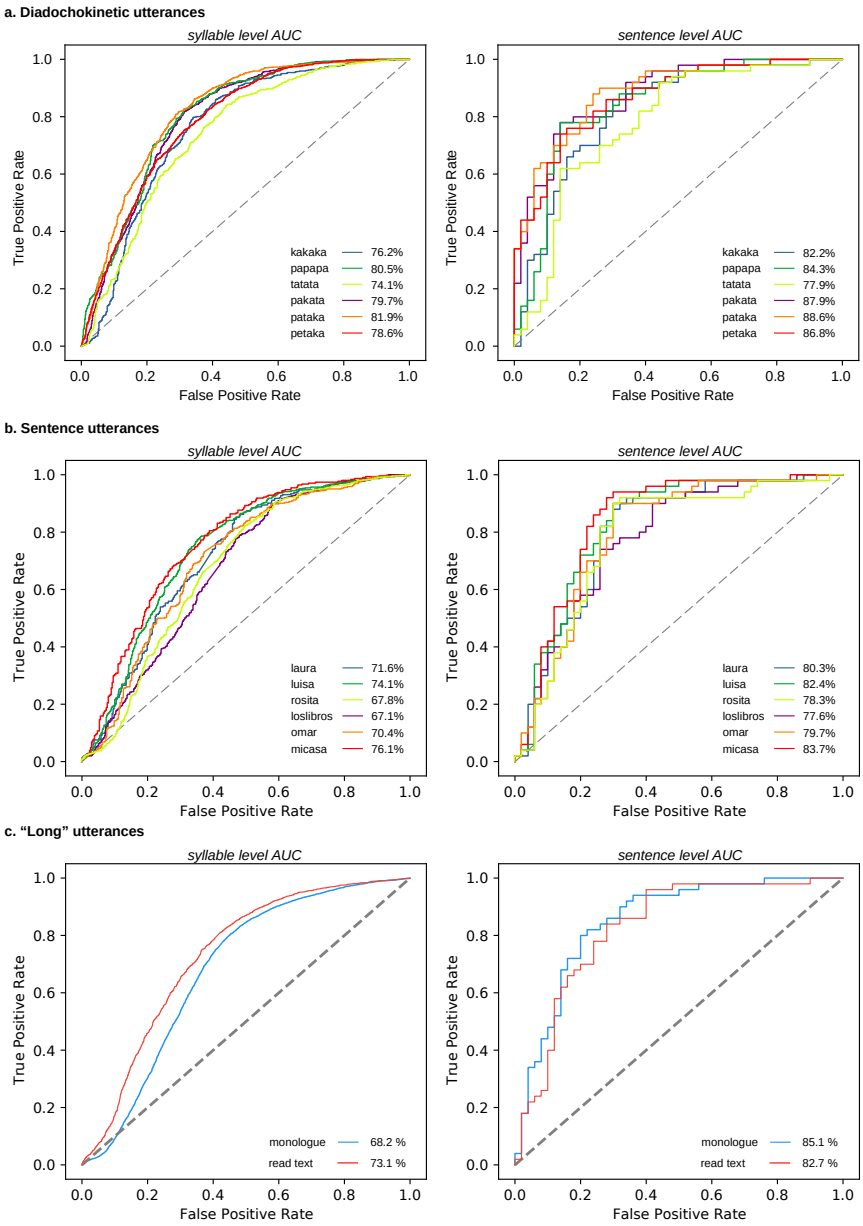

FIG. 3. Classification results on the PC-GITA dataset using SLF (utterance-level scores) and control feature sets are shown in the figure. As illustrated, the proposed SLF approach matches or outperforms the classification performance of all control feature sets across all speech types.

represent different utterances within speech type) and feature set. Detailed performance

metrics for the control feature sets are reported in Supplementary Tables 1c and 1d for eGeMAPS and ComPARE, respectively.

The results demonstrate consistently strong classification performance across all speech types, reinforcing the suitability of SLF as a robust candidate for PD detection from speech. Notably, utterance-level scores systematically outperform syllable-level scores in classification accuracy. However, syllable-level scores remain well above chance, indicating that even at this finer granularity, sufficient discriminative information is captured to detect PD-related speech alterations. This suggests that while individual syllables provide meaningful evidence, aggregating information across multiple syllable segments within an utterance further enhances classification performance by leveraging cumulative evidence.

Moreover, when compared to the control feature sets, the SLF approach, focusing on utterance-level scores (to align with the utterance-based nature of eGeMAPS and ComPARE), either outperforms or yields results comparable to those of the established feature sets (Figure 4). This suggests that the proposed SLF-based classification performance is on par with that of more traditional feature sets.

B. Importance of syllable segments

Having established that SLF provide sufficient information to discriminate between PD speech and HC speech, we then further focused on the importance of the segments being based on linguistic markers, such as syllable nuclei. To do this, we conducted an additional set of analysis comparing the SLF approach with a similar approach where segments are

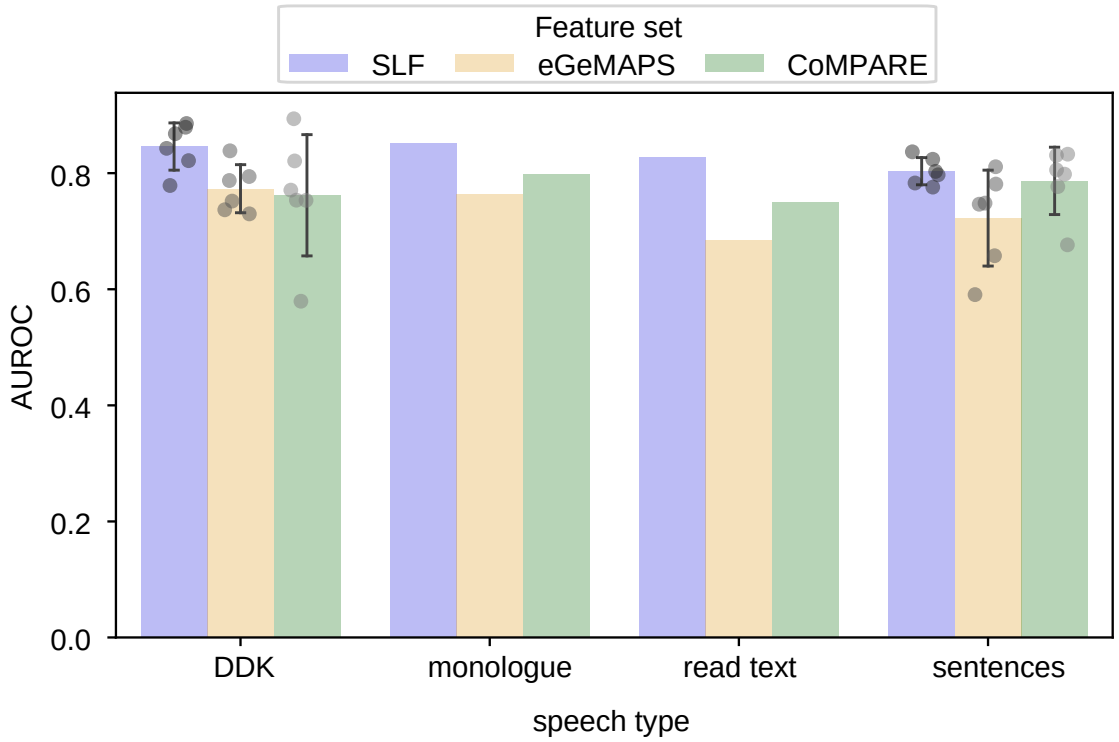

FIG. 4. The bars in the figure represent the classification performance (AUROC values, y-axis) using SLF (blue), eGeMAPS (red), and ComPARE (green) as feature sets, evaluated across all speech types from the PC-GITA dataset (x-axis).

not aligned on linguistic markers (syllable nuclei), but rather were randomly drawn from respective syllable duration distribution (see Methods for more details).

The classification procedure was identical to that used for SLF (e.g. in Figure 4), and the results for each speech type are presented in Figure 5. For consistency, we again aggregated scores from all segments into utterance-level scores; classification results are shown in Figure 5. Our (null) hypothesis is that syllable-based features are more advantageous and lead to better discrimination of Parkinsonian speech than random segment-based features. To test this hypothesis, we performed a Wilcoxon signed-rank test (restricted to speech types with

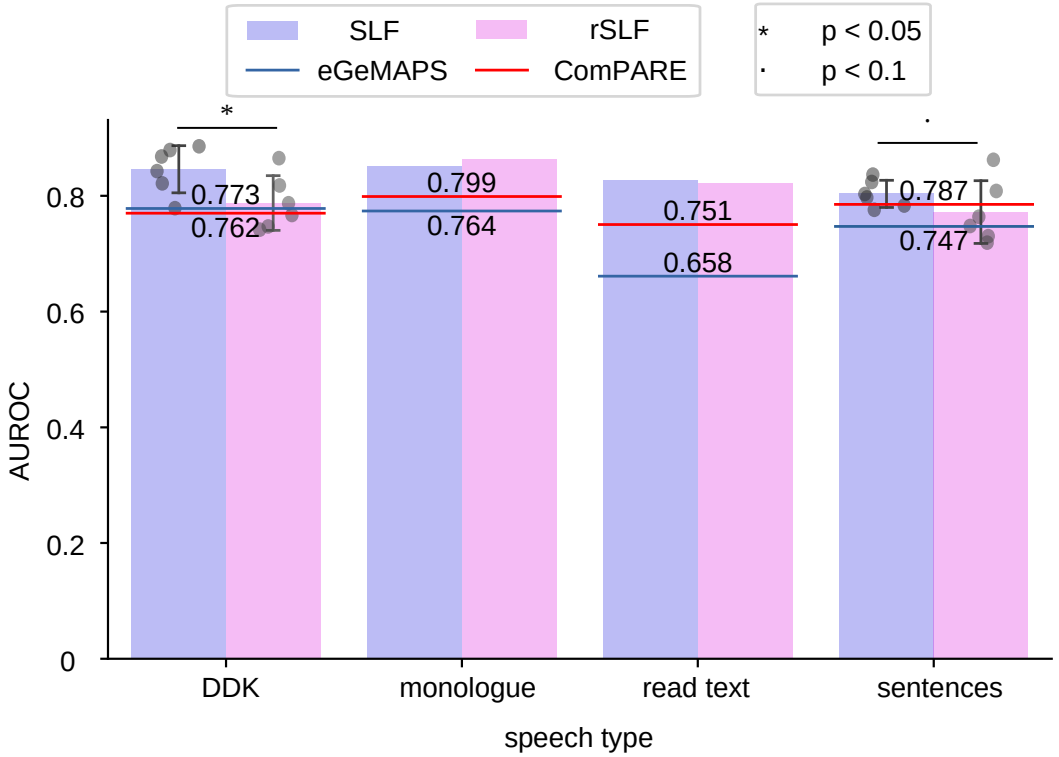

FIG. 5. The figure represents the classification results (measured as AUROC) with syllable based features (blue) versus random syllable-like segment (magenta).

multiple utterances; DDK and sentences). The difference was especially prominent for articulatorily constrained speech types, specifically, DDK speech (Wilcoxon W = 21, p = 0.016, rank biserial correlation = 1.0), and for shorter "free form" speech, sentences (Wilcoxon W = 18, p = 0.078, rank biserial correlation = 0.714). Overall, these results indicate that aligning segments with linguistic markers, such as syllable nuclei, is an important factor in the proposed feature extraction procedure.

C. Interpretability

To elucidate the information captured by SLF that enables effective classification of PD speech relative to HC speech, we investigated whether standardized syllable spectrograms encode formant patterns. Previous studies have reported alterations in formant patterns in the speech of PD patients (Convey *et al.*, 2023; Liu *et al.*, 2021, 2023). We therefore hypothesized that these patterns contribute to classification performance and sought to demonstrate their presence in SLF representations. To test this, we modified the construction of previously reported SLFs by replacing constant frequency binning with a targeted analysis of average energy around formant frequencies and the fundamental frequency. To further validate our findings, we conducted a statistical analysis of the eGeMAPS feature set across conditions.

Figure 6 demonstrates that the formant-based SLF, despite its lower dimensionality (5×8 vs. 64×8), achieves performance well above chance and its results are comparable to the classification results coming from established feature sets. More specifically, for DDK utterances, the average AUROC was 84.6% for SLF, 76.2% for ComPARE, 77.2% for eGeMAPS, and 75.75% for the formant-based SLF. This trend holds across other speech types: formant-based SLF performance is 5–10% lower than SLF but remains comparable to eGeMAPS and ComPARE.

These findings indicate that the standard SLF encodes formant and F0 information while also capturing supplementary features that enhance classification performance.

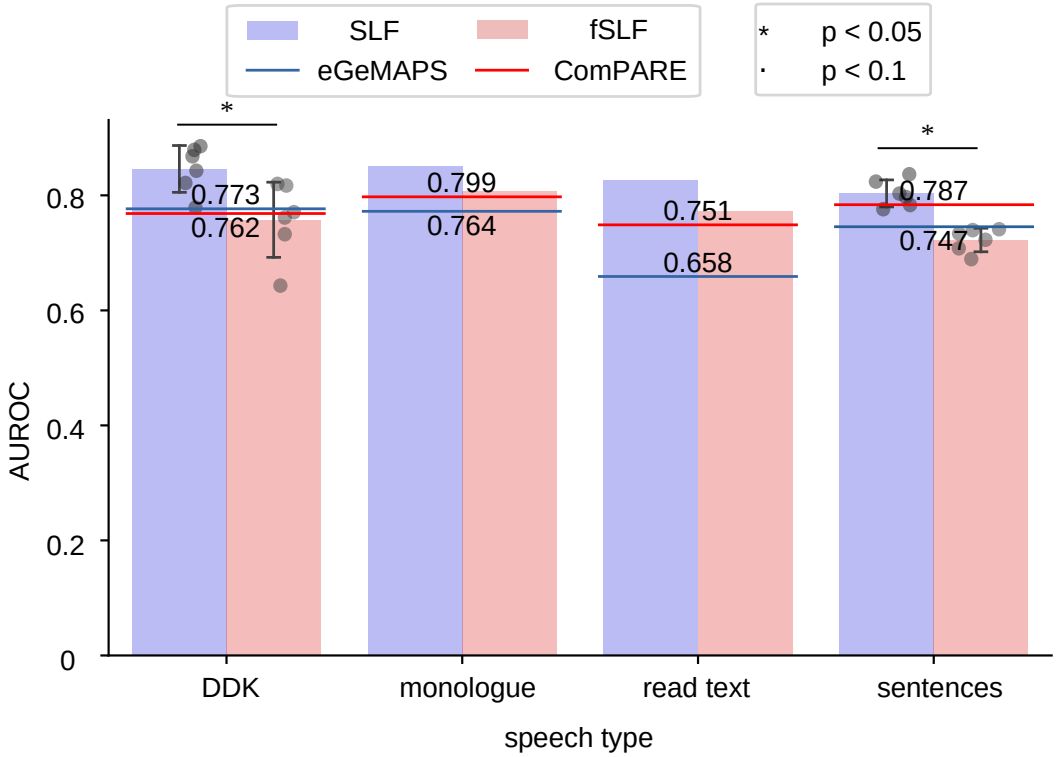

FIG. 6. Classification results for SLF (blue) and fSLF (red). The Wilcoxon signed-rank test indicates that SLF yields significantly better discriminatory power than fSLF for the DDK (Wilcoxon W = 21, p=0.031, rank biserial corrrelation=1) and sentence (Wilcoxon W = 21, p=0.031, rank biserial corrrelation=1) speech types.

To further validate our hypothesis, we conducted a linear mixed-model analysis of the eGeMAPS feature set. We selected eGeMAPS over ComPARE due to its curated set of 88 features, which directly relate to both the source and system related components of speech production. This feature set offers greater interpretability and a clearer link to the physiological mechanisms underlying speech (Dubagunta *et al.*, 2022).

In our analysis, we modeled fixed effects for eGeMAPS features and PD vs. HC condition, with speaker included as a random effect. We then examined the interaction between features

feature group	DDK	monologue	read text	sentences	totoal
F0	20	3		6	29
F1	6	1		6	13
F2	6	2		6	14
F3	10	1		2	13
mffc			1	3	4
other	1				1

TABLE I. The number of significant interactions between eGeMAPS features and PD vs HC conditions for each speech type in PC-GITA dataset.

and conditions, identifying statistically significant differences ($p < 0.05$) after correcting for multiple comparisons using the false discovery rate. The results for each speech and utterance type are presented in Table I.

Our findings reveal statistically significant differences across conditions for nearly all speech types, particularly in features related to formants and/or fundamental frequency. In some cases, MFCC-related features also showed significance, for more details, such as exact feature name for each speech type and utterance, as well as respective effect size and exact p-value, please see the Supplementary table 2a-d. These results align with our formant-based SLF analyses, reinforcing the conclusion that SLF effectively captures critical information about source and system of speech production, such as formant patterns and fundamental frequency at the syllable level.

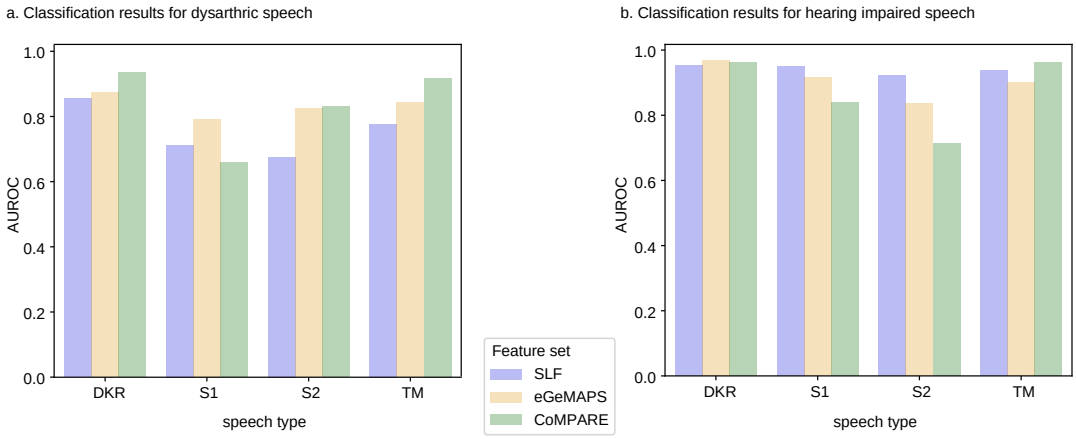

FIG. 7. Classification results on COPAS dataset

D. Speech Pathology detection / Generalizability beyond PD

To evaluate the generalizability of the SLF approach, we applied it to the COPAS dataset (Martens *et al.*, 2011), a well-established resource for pathological speech research. As described in the Methods section, this dataset comprises Dutch recordings from participants with various speech pathologies and includes multiple speech types. For this analysis, we focused exclusively on recordings from individuals with hearing impairment and dysarthria, as these were the pathologies with sufficient amount of recordings for robust classification analysis.

Classification results, obtained via K-fold cross-validation, are presented in Figure 7, with more detailed performance metrics presented in Supplementary Table 3a and 3b for dysarthic and hearing impairment cases, respectively. Due to differences in sample size, we used K=4 folds for hearing loss and K=5 folds for dysarthria. The results demonstrate that SLF achieves consistently high performance across all speech types for hearing loss, either

outperforming or matching the control feature sets. For dysarthria, however, performance was more modest, with control feature sets outperforming SLF across all speech types.

V. DISCUSSION

In this manuscript, we build upon our earlier proposal of a novel feature extraction method for speech pathology detection (Hovsepyan and Magimai.-Doss, 2024), drawing inspiration from neurocomputational models of speech perception (Hovsepyan *et al.*, 2020; Su *et al.*, 2023; Yildiz *et al.*, 2013). We conduct extensive testing of the new feature set across diverse speech types, expanding our initial analysis, previously limited to DDK tasks, to include longer speech samples such as sentences, monologues, and read text passages. Furthermore, we evaluate the proposed SLF method using Dutch dataset for pathological speech (Martens *et al.*, 2011), which includes recordings from individuals with pathologies such as hearing impairment and dysarthria. Our findings demonstrate that the SLF method is not constrained by speech type, language, or dataset, indicating its potential for broader generalizability.

Furthermore, to assess the role of syllable boundary segmentation, we compared linguistically informed segments with randomly segmented intervals of equivalent duration. The results reveal that linguistically derived segments yield superior performance, underscoring the critical role of syllable-like units in this approach.

Finally, to explore the interpretability of our results, we investigated whether SLF captures PD-related disruptions in vowel and formant articulation. Accordingly, we adapted the SLF creation process to capture dynamic frequency patterns, specifically those related

to formants and pitch, rather than relying solely on average spectral energy across uniformly distributed channels. To validate these findings, we employed classifier-independent statistical analyses to identify features with significant differences between PD and HC groups. These analyses corroborate that formant- and pitch-related information is most prominently affected in PD compared to HC conditions.

A. Importance of syllable segments

A critical aspect of our methodology is the segmentation of continuous speech into syllable-like units, a choice motivated by analysis-by-synthesis theories of speech perception and production (Bever and Poeppel, 2010; Halle and Stevens, 1962; Liberman and Mattingly, 1985). These theories posit that the brain decodes speech by generating and comparing internal articulatory gestures to match the incoming speech signal. Impairments in speech production may therefore manifest as detectable traces in perceptual analysis. We designed the SLF extraction process to mirror generative models of speech perception, which recognize syllables by iteratively comparing predicted and actual spectrograms. This approach has already demonstrated its ability to distinguish between PD patients and HC, suggesting that syllable-based representations can capture condition-specific differences.

A key question is whether syllable-based segmentation, or any segmentation with similar durations, is inherently advantageous. To test this, we first derived the syllable duration distribution for each speech type in the PC-GITA dataset (Orozco *et al.*, 2014). We then recalculated SLF using randomly sampled segment durations from these distributions.

Results revealed a significant performance drop for nearly all speech and utterance types. On average, performance declined by 7% for DDK, 15% for monologues, and 3% for sentences, with minimal change for read text. The latter two cases were somewhat unexpected: most probably the classification results for sentences were influenced by an outlier (micasa). By contrast, the results for the read text task can likely be explained by the fact that read text is more distinctly articulated (compared to monologues) and far longer (compared to sentences and DDK tasks). As a result, even randomly segmented chunks may closely resemble syllabic segments and, due to their longer durations, accumulate sufficient segmental information to yield performance comparable to syllable-based segmentation at the utterance level. Importantly, all utterance types except the outlier exhibited some performance drop, confirming the importance of syllable-based segmentation in our method.

Interestingly, random segmentation results closely matched those of control feature sets such as ComPARE and eGeMAPS, sometimes even surpassing them. One possible explanation for this similarity could arise because both control features and random SLF disregard linguistic boundaries, relying instead on fixed/small sliding windows that blur these boundaries. Random segment-based features thus function similarly to sliding windows, with classifiers identifying patterns across segments. The proposed approach, even with random segments, occasionally outperforms control features by providing multi-point evidence per utterance, in contrast to the single-point evidence offered by control sets. This suggests that pathological speech differences may not be uniformly distributed across utterance duration but are instead concentrated in specific segments. By dividing speech into multiple parts and analyzing each segment, we capture the most prominent differences, thereby enhancing clas-

sification performance. In contrast, the sliding window technique employed by eGeMAPS and ComPARE averages frame-level features across the entire utterance, tending to dilute these differences. This aligns with another key finding: syllable-level classification yielded lower AUC values than utterance-level classification, which aggregates syllable-level scores.

Lastly, it is well established that PD effects on speech is also manifested with disrupted speech rhythmicity, which is reflected in the syllable duration and rate of pathological speech. While the SLF method does not directly encode syllable duration, durations are normalized into fixed temporal bins, it may still indirectly capture temporal irregularities in PD speech. For example, nucleus-to-nucleus segmentation, as used in SLF extraction in this study, inherently incorporates the increased intersyllabic or inter-word intervals characteristic of PD. When spectral energy is averaged across fixed bins, the inclusion of "silent" segments reduces overall energy levels, particularly in free-form speech such as monologues, potentially improving classification performance.

Though effective, this approach highlights opportunities for refinement. Introducing dynamic sampling rates, using unequally spaced temporal bins to target specific syllable segments, could enhance temporal resolution. Such an approach, which has proven successful in syllable recognition, may further improve the SLF method's sensitivity and utility in detecting speech pathologies.

B. Interpretability

To better understand the information captured by the SLF method, we pursued two complementary approaches: first, we tested the hypothesis that SLF encodes formant and pitch-

related information; second, we performed a classifier-independent analysis of the eGeMAPS feature set to identify systematically differing features across conditions.

This hypothesis is grounded in research demonstrating that PD disrupts both vowel articulation and pitch control. These variations are particularly evident in sustained vowel tasks, where maintaining a stable vocal tract configuration over time is challenging for individuals with motor control impairments, such as those with PD.

To test this, we transformed the spectrogram representation of syllables by averaging spectral energy across formant- and pitch-related channels, rather than uniformly distributed frequency bins. This reduced the SLF feature vector from 64 to just 5 channels—over a tenfold reduction—while maintaining classification performance comparable to control feature sets. This result suggests that formant and pitch information is indeed central to SLF's effectiveness.

We further validated this by statistically comparing eGeMAPS features across conditions. Formant-related features, fundamental frequency, and select MFCCs consistently emerged as significant discriminators across all speech types.

While these findings confirm that SLF retains formant and pitch information through average spectral energy, they also reveal that SLF captures additional, richer details. Performance declined when using only formant- and pitch-related channels, indicating that SLF's strength lies in its ability to encode consonant articulation—information not directly measured by eGeMAPS or ComPARE. Since formants are tied to vowels and our segmentation focuses on syllable nuclei, SLF uniquely captures transitions between vowels, including

the execution of intervening consonants. This broader scope explains why the full SLF method outperforms feature sets limited to vowel- and pitch-related metrics.

C. Generalizability of the SLF Method

In this study, we expanded our neurocomputationally inspired method for speech pathology detection, which uses syllable spectrogram representations. This approach is both simple and effective, achieving classification accuracy comparable to established feature sets. We tested its generalizability by applying it to PD detection across diverse speech types and datasets, including the COPAS dataset in Dutch, which features hearing impairment and dysarthria.

The COPAS dataset allowed us to evaluate SLF on two distinct pathologies. For dysarthria, performance varied: SLF matched control feature sets in structured tasks like diadochokinetic utterances (DKR) but lagged in longer speech types. This discrepancy may stem from the diverse etiologies of dysarthria in COPAS, which can disrupt syllable boundary detection and reduce SLF's effectiveness compared to PD. In contrast, hearing impairment, a pathology rooted in perceptual deficits, aligned well with SLF's theoretical foundation. Analysis-by-synthesis theories suggest that production artifacts in hearing impairment arise from impaired efference copy, a feedback mechanism critical for motor control. SLF achieved over 90% AUROC across all speech types for hearing impairment, often surpassing control features and demonstrating remarkable consistency.

These results highlight SLF's robustness in capturing production deficits linked to perceptual mechanisms. The method's ability to generalize across languages, speech types, and pathologies underscores its potential as a versatile tool for speech pathology detection.

VI. CONCLUSION AND OUTLOOK

In this manuscript, we evaluated the use of syllable-level features (SLF) for detecting speech pathologies. Inspired by neurophysiologically plausible generative models of speech perception, SLF is grounded in the analysis-by-synthesis theory, which posits that the brain internally generates speech during perception. Our hypothesis was that production deficits due to pathology would manifest in perceptual representations, enabling detection from the perception side: a departure from traditional motor-focused approaches.

Syllables were chosen as the feature extraction unit because they are fundamental to speech production and perception, with distinct acoustic patterns and cognitive relevance. By extracting features from linguistically motivated segments, SLF captures linguistic, acoustic, and cognitive aspects of speech. Our results confirm that this approach reliably detects pathologies across datasets, speech types, and languages, with particularly strong performance for PD and hearing impairment. While SLF performed respectably for dysarthria, further refinement may be needed for pathologies with complex or variable effects on speech.

The SLF method is inherently flexible. It is, in principle, agnostic to specific segmentation or spectrogram calculation methods, allowing for optimization tailored to different pathologies. For example, dynamic frequency binning, focusing on formant and F0 patterns

affected by PD, could enhance targeted feature construction. Similarly, dynamic temporal sampling could improve resolution in regions critical for consonant articulation, further reducing feature dimensionality.

Future directions include exploring dynamic temporal and frequency sampling to focus on pathology-specific regions, as well as adapting SLF for long-term monitoring in noisy environments or on-device applications.

Overall, SLF demonstrates that linguistically driven feature extraction can advance speech pathology detection, offering a biologically plausible, low-profile alternative to traditional methods.

ACKNOWLEDGMENTS

VII. REFERENCES

Ananthanarayanan, A., Senivarapu, S., and Murari, A. (**2025**). "Towards Causal Interpretability in Deep Learning for Parkinson's Detection from Voice Data" doi: 10.1101/2025.04.25.25326311.

Bever, T. G., and Poeppel, D. (**2010**). "Analysis by Synthesis: A (Re-)Emerging Program of Research for Language and Vision," Biolinguistics 27, doi: 10.5964/bioling.8783.

Boersma, P., and Weenink, D. (**2003**). "Praat: doing phonetics by computer," https://api.semanticscholar.org/CorpusID:60594797.

COPAS				
Condition	**Speech Type**	**Subjects**	**M** (n, age ± std)	**F** (n, age ± std)
control	DKR	83	34 (40.50 ± 16.37)	49 (33.27 ± 17.05)
	S1	82	33 (41.09 ± 16.25)	49 (33.27 ± 17.05)
	S2	82	33 (41.12 ± 16.21)	49 (33.27 ± 17.05)
	TM	82	34 (40.50 ± 16.37)	48 (32.96 ± 17.09)
dysarthia	DKR	55	32 (44.56 ± 21.78)	23 (44.52 ± 25.05)
	S1*	50	28 (43.46 ± 22.62)	21 (44.76 ± 24.94)
	S2*	50	28 (43.57 ± 23.11)	21 (44.76 ± 24.94)
	TM*	48	28 (43.46 ± 22.43)	19 (41.79 ± 26.09)
hearing loss	DKR	24	9 (31.33 ± 7.97)	15 (35.93 ± 15.36)
	S1	24	9 (31.33 ± 7.97)	15 (35.93 ± 15.36)
	S2	24	9 (31.33 ± 7.97)	15 (35.93 ± 15.36)
	TM	26	9 (31.33 ± 7.97)	17 (38.29 ± 15.98)
PC-GITA				
Healthy Control	DDK, sentences, read text, monologue	50	25 (61.2 ± 11.3)	25(60.7 ± 7.7)
Parkinson's Disease	DDK, sentences, Read text, monologue	50	25 (62.2 ± 11.2)	25(60.1 ± 7.8)

TABLE II. The table shows the number of speakers, their mean age (± standard deviation) by gender for each pathology and speech type in the COPAS and PC-GITA datasets. An asterisk indicates that, for a particular speech type, there was a speaker with no gender information.

586 Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. (**2017**). "The free energy principle
587 for action and perception: A mathematical review," Journal of Mathematical Psychology
588 **81**, 55–79, doi: 10/gcnkg7.

Convey, R. B., Ihalainen, T., Liu, Y., Räsänen, O., Ylinen, S., and Penttilä, N. (**2023**). "A comparative study of automatic vowel articulation index and auditory-perceptual assessments of speech intelligibility in Parkinson's disease," International Journal of Speech-Language Pathology 1–11, doi: 10.1080/17549507.2023.2251725.

Dogonasheva, O., Doelling, K. B., Zakharov, D., Giraud, A.-L., and Gutkin, B. (**2025**). "Rhythm-based hierarchical predictive computations support acoustic-semantic transformation in speech processing," Nature Computational Science **5**(10), 915–926, doi: 10.1038/s43588-025-00876-9.

Dubagunta, S. P., Magimai.-Doss, M., Eleni Theocharopoulos, and Mathew Magimai Doss (**2022**). "Towards Automatic Prediction of Non-Expert Perceived Speech Fluency Ratings," ICMI Companion doi: 10.1145/3536220.3563689.

Escobar-Grisales, D., Ríos-Urrego, C. D., and Orozco-Arroyave, J. R. (**2023**). "Deep Learning and Artificial Intelligence Applied to Model Speech and Language in Parkinson's Disease," Diagnostics **13**(13), 2163, doi: 10.3390/diagnostics13132163.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (**2016**). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," IEEE Transactions on Affective Computing **7**(2), 190–202, doi: 10.1109/TAFFC.2015.2457417.

Eyben, F., Wöllmer, M., and Schuller, B. (**2010**). "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, Association for Computing Machinery, New York,

NY, USA, pp. 1459–1462, doi: 10.1145/1873951.1874246.

Friston, K. (**2010**). "The free-energy principle: A unified brain theory?," Nature Reviews Neuroscience **11**(2), 127–138, doi: 10/bj3r7f.

Friston, K., and Kiebel, S. (**2009**). "Predictive coding under the free-energy principle," Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1521), 1211–1221, doi: 10/fcswfc.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., Olanow, C. W., Rascol, O., Schrag, A., Teresi, J. A., Van Hilten, J. J., and LaPelle, N. (**2008**). "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," Movement Disorders **23**(15), 2129–2170, doi: 10.1002/mds.22340.

Halle, M., and Stevens, K. (**1962**). "Speech recognition: A model and a program for research," IRE Transactions on Information Theory **8**(2), 155–159, doi: 10.1109/TIT.1962.1057686.

Hoehn, M. M., and Yahr, M. D. (**1967**). "Parkinsonism: Onset, progression and mortality," Neurology **17**(5), 427–442, doi: 10.1212/wnl.17.5.427.

Hovsepyan, S., and Magimai.-Doss, M. (**2024**). "Syllable Level Features for Parkinson's Disease Detection from Speech," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Seoul, Korea, Republic of, pp. 11416–11420, doi: 10.1109/ICASSP48485.2024.10446484.

633 Hovsepyan, S., Olasagasti, I., and Giraud, A.-L. (**2020**). "Combining predictive coding and
634 neural oscillations enables online syllable recognition in natural speech," Nature Commu-
635 nications **11**(1), 3117, doi: 10/gg2n6w.
636 Hovsepyan, S., Olasagasti, I., and Giraud, A.-L. (**2023**). "Rhythmic modulation of pre-
637 diction errors: A top-down gating role for the beta-range in speech processing," PLOS
638 Computational Biology **19**(11), e1011595, doi: 10.1371/journal.pcbi.1011595.
639 Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., and Giraud, A. L. (**2015**). "Speech
640 encoding by coupled cortical theta and gamma oscillations," eLife **4**(MAY), 1–45, doi:
641 10/gdh97p.
642 Jadoul, Y., Thompson, B., and De Boer, B. (**2018**). "Introducing Parselmouth: A Python
643 interface to Praat," Journal of Phonetics **71**, 1–15, doi: 10.1016/j.wocn.2018.07.001.
644 Karlsson, F., Schalling, E., Laakso, K., Johansson, K., and Hartelius, L. (**2020**). "Assess-
645 ment of speech impairment in patients with Parkinson's disease from acoustic quantifica-
646 tions of oral diadochokinetic sequences," The Journal of the Acoustical Society of America
647 **147**(2), 839–851, doi: 10.1121/10.0000581.
648 Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception
649 revised," Cognition **21**(1), 1–36, doi: 10/d79rrw.
650 Liu, Y., Penttilä, N., Ihalainen, T., Lintula, J., Convey, R., and Räsänen, O. (**2021**).
651 "Language-Independent Approach for Automatic Computation of Vowel Articulation Fea-
652 tures in Dysarthric Speech Assessment," IEEE/ACM Transactions on Audio, Speech, and
653 Language Processing **29**, 2228–2243, doi: 10.1109/TASLP.2021.3090973.

Liu, Y., Reddy, M. K., Penttilä, N., Ihalainen, T., Alku, P., and Räsänen, O. (**2023**). "Automatic Assessment of Parkinson's Disease Using Speech Representations of Phonation and Articulation," IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 242–255, doi: 10.1109/TASLP.2022.3212829.

Martens, J., Bodt, M., Nuffelen, G. V., and Middag, C. (**2011**). "Corpus of Pathological and Normal Speech (COPAS)," .

Mazzoni, P., Shabbott, B., and Cortés, J. C. (**2012**). "Motor Control Abnormalities in Parkinson's Disease," Cold Spring Harbor Perspectives in Medicine **2**(6), a009282, doi: 10.1101/cshperspect.a009282.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., and Nieto, O. (**2015**). "Librosa: Audio and Music Signal Analysis in Python," in *Python in Science Conference*, Austin, Texas, pp. 18–24, doi: 10.25080/Majora-7b98e3ed-003.

Michaelis, D., Fröhlich, M., and Strube, H. W. (**1998**). "Selection and combination of acoustic features for the description of pathologic voices," The Journal of the Acoustical Society of America **103**(3), 1628–1639, doi: 10.1121/1.421305.

Nabé, M., Schwartz, J.-L., and Diard, J. (**2021**). "COSMO-Onset: A Neurally-Inspired Computational Model of Spoken Word Recognition, Combining Top-Down Prediction and Bottom-Up Detection of Syllabic Onsets," Frontiers in Systems Neuroscience **15**, 75, doi: 10.3389/fnsys.2021.653975.

Nair, V., Susskind, J., and Hinton, G. E. (**2008**). "Analysis-by-Synthesis by Learning to Invert Generative Black Boxes," in *Artificial Neural Networks - ICANN 2008*, edited by V. Kůrková, R. Neruda, and J. Koutník, **5163** (Springer Berlin Heidelberg, Berlin, Hei-

delberg), pp. 971–981, doi: 10.1007/978-3-540-87536-9_99.

Orozco, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J., González-Rátiva, M., and Noeth, E. (**2014**). "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *International Conference on Language Resources and Evaluation (LREC)*.

Räsänen, O., Doyle, G., and Frank, M. C. (**2018**). "Pre-linguistic segmentation of speech into syllable-like units," Cognition **171**, 130–150, doi: 10.1016/j.cognition.2017.11.003.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. S. (**2010**). "The INTERSPEECH 2010 paralinguistic challenge," in *Interspeech 2010*, ISCA, pp. 2794–2797, doi: 10.21437/Interspeech.2010-739.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., Nöth, E., Zhang, Y., and Weninger, F. (**2015**). "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition," in *Interspeech 2015*, ISCA, pp. 478–482, doi: 10.21437/Interspeech.2015-179.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., and Evanini, K. (**2016**). "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Interspeech 2016*, ISCA, pp. 2001–2005, doi: 10.21437/Interspeech.2016-129.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (**2013**). "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of the Annual*

Conference of the International Speech Communication Association, INTERSPEECH, pp. 148–152, doi: 10.21437/Interspeech.2013-56.

Schuller, B. W., Batliner, A., Bergler, C., Mascolo, C., Han, J., Lefter, I., Kaya, H., Amiriparian, S., Baird, A., Stappen, L., Ottl, S., Gerczuk, M., Tzirakis, P., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., Rothkrantz, L. J. M., Zwerts, J. A., Treep, J., and Kaandorp, C. S. (**2021**). "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates," in *Proc. Interspeech 2021*, pp. 431–435, doi: 10.21437/Interspeech.2021-19.

Su, Y., MacGregor, L. J., Olasagasti, I., and Giraud, A.-L. (**2023**). "A deep hierarchy of predictions enables online meaning extraction in a computational model of human speech comprehension," PLOS Biology **21**(3), e3002046, doi: 10.1371/journal.pbio.3002046.

van Gelderen, L., and Tejedor-García, C. (**2024**). "Innovative Speech-Based Deep Learning Approaches for Parkinson's Disease Classification: A Systematic Review," Applied Sciences **14**(17), 7873, doi: 10.3390/app14177873.

Yildiz, I. B., von Kriegstein, K., and Kiebel, S. J. (**2013**). "From Birdsong to Human Speech Recognition: Bayesian Inference on a Hierarchy of Nonlinear Dynamical Systems," PLoS Computational Biology **9**(9), e1003219–e1003219, doi: 10/gf26tp.