Parkinson's Disease Detection through Formant and F_0 Analysis at Syllable Level

1st Sevada Hovsepyan
Speech & Audio Processing Group
Idiap Research Institute
Martigny, Switzerland
https://orcid.org/0000-0001-8640-4206

2nd Mathew Magimai.-Doss
Speech & Audio Processing Group
Idiap Research Institute
Martigny, Switzerland
https://orcid.org/0000-0002-8714-1409

***Abstract*—In a recent publication, we put forth a novel approach to syllable-based feature extraction for the detection of Parkinson's disease. The method entails the calculation of standardised spectrotemporal patterns of syllable-like segments (fixed number of frequency and temporal bins), which are then employed as a feature vector for the detection of Parkinson's disease. As the classification performance based on the syllable-level features increased with the inclusion of more frequency bins, we postulated that the standardised spectrotemporal patterns bear resemblance to, or contain, the formant transitions that have been demonstrated to be altered in Parkinson's disease. In this study, we initially demonstrated that the extraction of syllable-level features based on spectrogram energy under the formant and F0 patterns resulted in a significant improvement in classification outcomes. To further test our hypothesis, we statistically compared the eGeMAPS feature set across conditions. This revealed that features related to fundamental frequency and formants are statistically different between Parkinson's disease and healthy conditions. Taken together, our results suggest that syllable-level, formant-informed feature selection can provide reliable PD detection with a relatively small number of features.**

***Index Terms*—Parkinson's disease, speech pathology detection, syllables, formants**

I. Introduction

Parkinson's disease is one of the most prevalent neurodegenerative diseases, affecting a significant number of patients each year [1], [2]. An early diagnosis is of the utmost importance for a more successful treatment [3]. Consequently, in recent years, numerous studies have proposed non-invasive methods for the detection of Parkinson's disease [4], [5]. A significant proportion of this research has been based on speech analysis [6]–[8], where various feature extraction methods were used: starting from hand-crafted feature sets [6], [9], [10] to the use of deep learning methods [11]–[14]. A few notable examples are listed below.

In their 2019 study, Moro-Velazquez and colleagues investigated the potential of distinct phonemic groups for the automated detection of Parkinson's Disease (PD) in speech [15]. The study put forth a novel phonemic grouping technique for the analysis of three distinct speech corpora. The method achieved AUC values between 91% and 98% within individual corpora and between 84% and 95% in cross-corpora trials. The study identified plosives, vowels, and fricatives as key phonemic groups for successful PD detection.

In another study, Lie et al. (2021), the authors initially proposed an automatic, language-independent formant extraction method and demonstrated its efficacy in assessing impairment in vowel articulation in PD patients [16]. The method enabled them to compare datasets from different languages (Finnish, English, and Spanish) and to demonstrate that the vowel articulation of PD patients is impaired in a language-independent manner, thereby substantiating the potential of formants as a means of discerning these differences.

Another recent study [17] suggested a novel speech behavioral test, altering auditory feedback during various speech production tasks (e.g., DDK, whispering). They then used supervised machine learning algorithms to classify between PD and HC groups, achieving an accuracy of 85.4%.

Previously, we proposed a new method [18] for syllable-level feature (SLF) extraction based on standardized spectrograms of syllables, inspired by a number of neurocomputational models of speech perception [19], [20]. Although the proposed method performed well (AUC=78%-89%) [18], it was unclear what underlying information SLFs contained for successful classification. In this paper, in light of previous work showing that formant patterns are altered in the speech of Parkinson's disease patients [6], [7], [21], [22], we hypothesized and attempted to demonstrate that the standardized spectrograms of syllables contain information about formant patterns. We modified the construction of previously reported SLFs [18], and instead of using constant binning for the frequency channels, we calculated the average energy under the formant patterns and F0. In addition, we performed a statistical analysis of the eGeMAPS [23] feature set across conditions to corroborate our findings.

The remainder of the manuscript is structured as follows: In Section II, we provide an overview of our previous work on syllable-level feature extraction for the detection of Parkinson's disease from speech. Section III outlines the proposed modification to the original approach. It also includes information about the used dataset, baseline feature set, as well as classification analysis and statistical tests. Section IV presents

This work was funded by the Swiss National Science Foundation through the Bridge Discovery project EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (grant no. 40B2 0 194794)

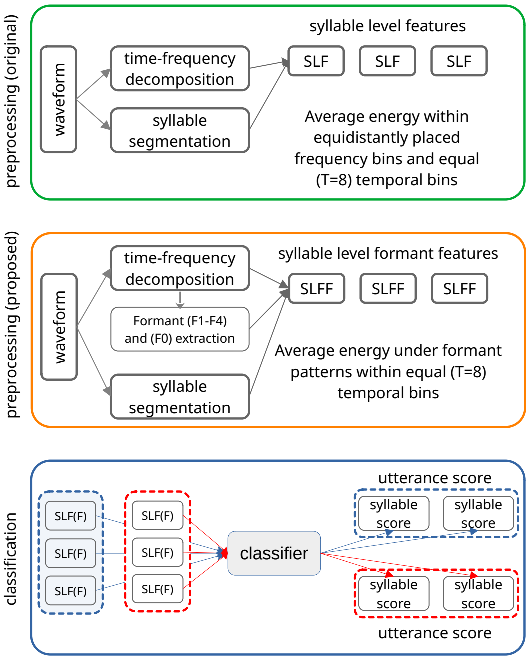

Fig. 1. Overview of the proposed methods: The top panel (green) corresponds to the originally proposed syllable-level feature extraction, which takes the average activity within equally spaced frequency channels within T=8 temporal bins. The second panel (orange) represents the currently proposed methodology for syllable level feature extraction, which, unlike the original method, averages the energy under the formant patterns. Finally, SLFs or SLFFs are used for the classification process, providing both syllable-level and utterance-level scores.

the new classification results, comparisons with the results from the original investigation, as well as the results of the statistical analysis. Section V concludes our findings.

II. Background

In this section, we provide a detailed explanation of the syllable-level feature (SLF) extraction proposed in our prior investigation [18]. The proposed approach rests on the notion that the spectrogramm of each syllable is divided into equal temporal bins (e.g., eight bins) and that the average energy of each frequency channel/bin within each temporal bin is calculated (Fig 1, green box). This method draws inspiration from a series of neurocomputational models of speech perception [19], [20], thereby enabling a standardised representation of spectrotempral patterns of syllables, irrespective of their original duration. Consequently, it offers a straightforward and intuitive approach to utilising linguistic markers, such as syllables, as a ”frame” window for feature extraction, ensuring a consistent number of features for each window.

In the original study we have tested different numbers of frequency channels, different spectrorgam types and linguistic markers for syllable segmentation and found that the best classification results are obtained with STFT and peak-to-peak segmentation (roughly corresponding to syllable nuclei to syllable nuclei) [18]. The performance increased with the inclusion of more frequency channels and the optimal performance (mean AUC = 83.2%) was achieved with the 46-channel short-term Fourier transform (STFT), whereby each SLF was based on 46 homogeneously separated frequency bins spanning from 0 up to around 10kHz. We attribute this finding to the idea that the more channels there are, the more spectral information is available for classification.

III. Methods

In this section we outline the main steps for replicating the study. We start by explaining the proposed methodology and highlighting the differences from the original study. We then describe the dataset used and the baseline feature set. Next, the classification protocol used is described and the section concludes with the description of the statistical analysis performed.

A. Proposed Methodology

In this section, we have described the modification of the previously proposed approach [18] to extract formant (and F0) features at the syllable level. The modification is based on the idea that instead of using constant binning for the frequency channels, we calculated the average energy under the formant patterns (Fig. 1, orange box, Fig. 2).

The procedure consists of the following steps: 1. for each utterance, the power STFT and the syllable boundaries are extracted. 2. additionally, the patterns of the fundamental frequency (F0) and the first 4 formants (F1-F4) are extracted. 3. for each syllable-like segment (between 2 consecutive syllable boundaries), the energy (spectrogram amplitude) under the pitch (F0) and formant (F1-F4) patterns was derived (Fig. 2). Then the syllable segment was divided into T=8 equal temporal bins and the average energy under F0 and F1-F4 patterns was derived. This results in a 5x8 representation of a syllable.

Finally, the flattened representation (1x40) is used for the classification procedure (Fig. 1 blue box). Syllable level formant (&F0) features (SLFFs) from each utterance are used for the classification proposal, then the classification score corresponding to each utterance is combined to derive the utterance level score.

B. Dataset

In this study, we used a Spanish speech corpus database containing recordings from 50 patients with Parkinson’s disease [24]. As a control, the database also includes recordings from 50 healthy patients. Both groups are gender and age matched. The PD dataset consists of 25 males (mean age 62.2±11.2 years) and 25 females (mean age 60.1±7.8 years). Similarly, the HC group consists of 25 men and 25 women with a mean age of 61.2±11.3 years and 60.7±7.7 years respectively. All subjects perform the same set of speech tasks, e.g. vowel sustaining, free speech, diadochokinetic (DDK) utterances, etc. All recordings in the dataset are sampled at 44.1 kHz with 16-bit resolution.

In this study, we have only used DDK utterances as they are designed to stress the articulatory system - and therefore

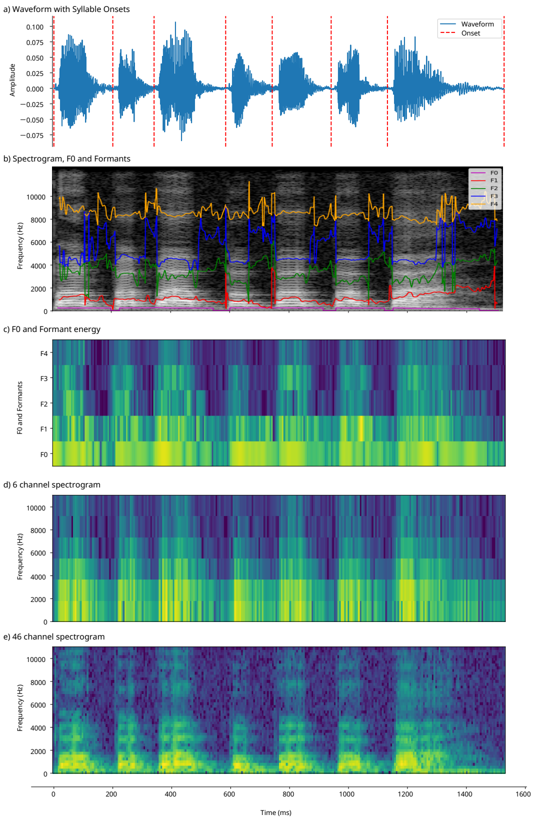

Fig. 2. Example of DDK utterance (PD condition), where we show detected onsets (a). full spectrogram with F0 and F1-F4 and reduced spectorgrams

may better reveal differences in articulation due to Parkinson's disease. Each participant utters 6 different DDK utterances (`papapa`, `tatata`, `kakaka`, `pataka`, `pakata`, `petaka`). Therefore, for each type of DDK utterance there are 100 recordings, 50 corresponding to the PD condition and 50 to the HC.

C. *Baseline feature set*

As baseline feature sets, we considered the ComParE [25] and eGeMAPS [23] feature sets, which, although primarily designed for paralinguistic challenges, have also been used successfully for PD detection [26], [27]. Here, we use eGeMAPS as baseline because, first, in our initial investigation [18], we saw that on average (mean AUC = 76.2%) it provides better classification results on the used dataset compared to ComParE (mean AUC = 65.33%). Second, it consists of 88 carefully selected features, so its features are more interpretable and can be conditionally divided into two subgroups: those related to either the source or the system of speech production [28].

D. *Classification protocol*

We applied the leave-one-subject-out (LOSO) protocol for each DDK utterance type. Therefore, for each utterance, we first extracted (SLFFs) and used them as test set. The SLFFs for the remaining utterances were used for training the classifier. In this case, we used a cubic kernel support vector machine without hyperparameter optimisation. For each fold, the classifier outputs the classification scores for each syllable of omitted utterance, which we then combined by averaging to obtain the utterance level score.

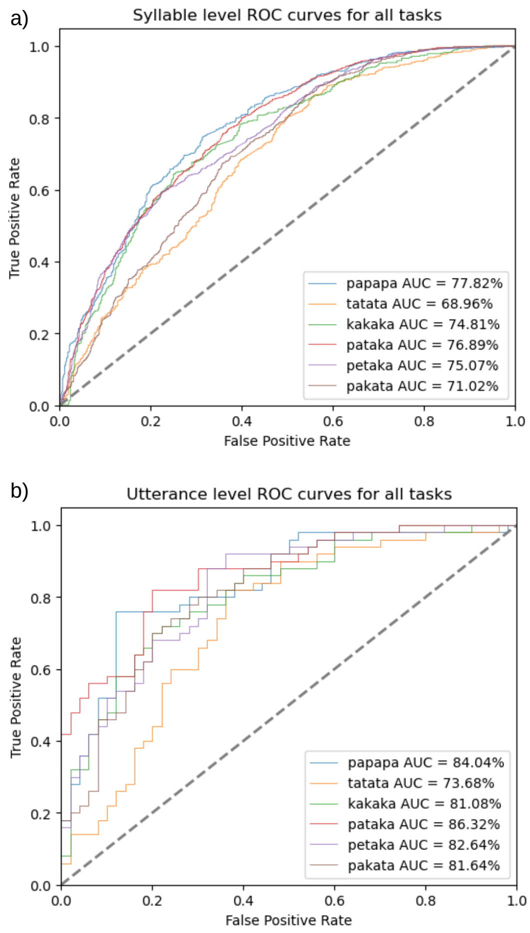

Fig. 3. ROC curves and AUC values for syllable- and utterance-level scores

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 modelled 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).

F. *Toolboxes*

The following toolboxes were used for during this study: The syllable segmentation was performed with the Python implementation of syllable segmentation algorithm based on sonority envelope and neural oscillations [29]. STFT spectrograms were calculated with the `librosa` library [30], whereas formats and fundamental frequency were extracted

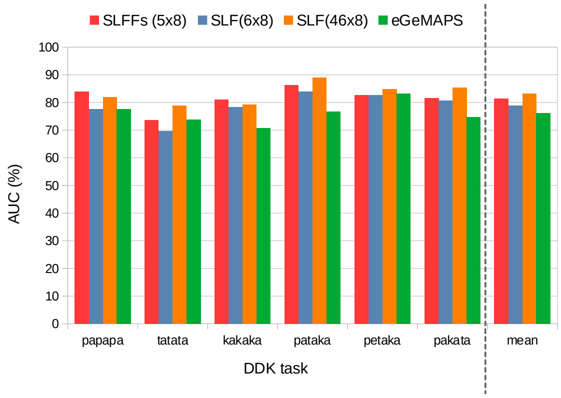

Fig. 4. Comparison with SLF and baseline eGeMAPS

with Python implementation of Praaat software [31] with the `parselmount` library [32]. Finally, eGeMAPS [23] feature set was extracted with the OpenSMILE toolbox [33].

IV. Results

In this section we present the classification results for each DDK utterance type with the LOSO protocol as well as the statistical test analysis.

A. *Classification analysis*

Subpanel a) in Fig. 3 is based on the classification results based on the syllable level scores, while subpanel b) corresponds to the classification results based on the utterance level scores. As we have also seen from the original study, combining the scores of all syllables within an utterance and using the utterance level score for classification leads to better classification results. Overall, we see good classification results and in all cases well above chance level.

Moreover, in Fig. 4 we compare the results of the original study with 46-channel SLF (mean AUC = 83.23%) and the baseline with eGeMAPS (mean AUC = 76.2 %), the new formant-based features perform at the same level (mean AUC = 81.6 %), despite having a significantly lower number of features per syllable. The SLFF also outperform, the original SLF approach with the similar number of features (6 channel STFT, mean AUC = 78%).

In the original paper [18], we also showed that there is a performance gain when syllable-level features are extracted based on peak-to-peak segmentation (46-channel STFT, mean AUC=83.23%) compared to valley-to-valley segmentation (46-channel STFT, mean AUC=80.1%). This observation was consistent across all tested spectrogram types and number of frequency channels. We attributed this observation to the notion that peak-based segmentation might capture consonant transitions that might be more informative. However, as can be seen in TABLE I, for SLFFs, both peak- and valley-based segmentation lead to similar results, which may be due to the notion that features are based on F0 and formants (which are primarily vowel attributes), information about consonants is missing. This can also explain why the 46 channel STFT still performs better (Fig. 2), since it contains more details (possibly also about consonants).

TABLE I
Peak and Valley Based Segmentation

task	peaks	valleys
papapa	84.04	83.20
tatata	73.68	77.92
kakaka	81.00	80.08
pataka	86.32	82.68
pakata	82.64	80.92
petaka	81.64	85.40
mean	81.6	81.7

B. *Significant features across conditions*

The results of the statistical analysis (see Methods: Statistical analysis for details) are presented in TABLE II, where we have listed all statistically significant (p<0.05) features and whether they were related to F0 or formats. The numbers in the table represent how many signficicant interection were for each DDK utterance type and feature category (e.g. related to F0) pair. Interestingly, all the significant features were related to either the fundamental frequency or formats. We can also see that the DDK utterances of `tatata` and `kakaka` have the fewest number of significant features, which seems to agree with the classification results, as this number was also the lowest for these utterance types. Its is also worth mentioning, that the DDK utterances with higher performance are those that show differences in both source (F0) and system-related (formants) features, suggesting that PD affects both aspects of speech production.

TABLE II
Statistically Significant Features Across Conditions

task	F0	F1	F2	F3	total
papapa	4	1	1	2	8
tatata	2	-	1	1	4
kakaka	2	-	-	-	2
pataka	4	2	1	2	9
pakata	4	1	1	2	8
petaka	4	1	1	2	8
total	20	5	5	9	

V. Conclusions

The revised syllable-level feature extraction method was found to be an effective approach overall. The classification results were consistent with those of the original approach, but with a reduction in the number of features per syllable, thereby enhancing efficiency. The results support the hypothesis that reduced spectrograms in SLF capture formant dynamics, which aid in PD vs. HC classification. Statistical analysis demonstrated that the features most affected are related to fundamental frequency and formants. Therefore, it can be concluded that syllabic-based feature extraction effectively captures articulatory differences due to PD, especially when frequency binning is based on formant (F0)-informed variable frequency patterns.

References

[1] Jinee Goyal, Padmavati Khandnor, and Trilok Chand Aseri, "Classification, Prediction, and Monitoring of Parkinson's disease using Computer Assisted Technologies: A Comparative Analysis," *Engineering Applications of Artificial Intelligence*, vol. 96, pp. 103955, Nov. 2020.

[2] M. C. de Rijk, L. J. Launer, K. Berger, M. M. Breteler, J. F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, and A. Hofman, "Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group," *Neurology*, vol. 54, no. 11 Suppl 5, pp. S21–23, 2000.

[3] Melissa J. Armstrong and Michael S. Okun, "Diagnosis and Treatment of Parkinson Disease: A Review," vol. 323, no. 6, pp. 548–560.

[4] Quoc Cuong Ngo, Mohammod Abdul Motin, Nemuel Daniel Pah, Peter Drotár, Peter Kempster, and Dinesh Kumar, "Computerized analysis of speech and voice for Parkinson's disease: A systematic review," vol. 226, pp. 107133.

[5] Liaqat Ali, Ashir Javeed, Adeeb Noor, Hafiz Tayyab Rauf, Seifedine Kadry, and Amir H. Gandomi, "Parkinson's disease detection based on features refinement through L1 regularized SVM and deep neural network," vol. 14, no. 1, pp. 1333.

[6] Sabine Skodda, Wenke Visser, and Uwe Schlegel, "Vowel Articulation in Parkinson's Disease," vol. 25, no. 4, pp. 467–472.

[7] Yuanyuan Liu, Mittapalle Kiran Reddy, Nelly Penttilä, Tiina Ihalainen, Paavo Alku, and Okko Räsänen, "Automatic Assessment of Parkinson's Disease Using Speech Representations of Phonation and Articulation," vol. 31, pp. 242–255.

[8] Clayton R. Pereira, Danilo R. Pereira, Silke A. T. Weber, Christian Hook, prefix=de useprefix=true family=Albuquerque, given=Victor Hugo C., and João P. Papa, "A survey on computer-assisted Parkinson's Disease diagnosis," vol. 95, pp. 48–63.

[9] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*. Aug. 2013, pp. 148–152, ISCA.

[10] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," vol. 56, no. 4, pp. 1015.

[11] Marek Wodzinski, Andrzej Skalski, Daria Hemmerling, Juan Rafael Orozco-Arroyave, and Elmar Nöth, "Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 717–720.

[12] Anu Iyer, Aaron Kemp, Yasir Rahmatallah, Lakshmi Pillai, Aliyah Glover, Fred Prior, Linda Larson-Prior, and Tuhin Virmani, "A machine learning method to process voice samples for identification of Parkinson's disease," vol. 13, no. 1, pp. 20615.

[13] Máté Hireš, Matej Gazda, Peter Drotár, Nemuel Daniel Pah, Mohammod Abdul Motin, and Dinesh Kant Kumar, "Convolutional neural network ensemble for Parkinson's disease detection from voice recordings," vol. 141, pp. 105021.

[14] Awais Mahmood, Muhammad Mehroz Khan, Muhammad Imran, Omar Alhajlah, Habib Dhahri, and Tehmina Karamat, "End-to-End Deep Learning Method for Detection of Invasive Parkinson's Disease," vol. 13, no. 6, pp. 1088.

[15] Laureano Moro-Velazquez, Jorge A. Gomez-Garcia, Juan I. Godino-Llorente, Francisco Grandas-Perez, Stefanie Shattuck-Hufnagel, Virginia Yagüe-Jimenez, and Najim Dehak, "Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's Disease," vol. 9, no. 1, pp. 19066.

[16] Yuanyuan Liu, Nelly Penttilä, Tiina Ihalainen, Juulia Lintula, Rachel Convey, and Okko Räsänen, "Language-Independent Approach for Automatic Computation of Vowel Articulation Features in Dysarthric Speech Assessment," vol. 29, pp. 2228–2243.

[17] Ángeles Piña Méndez, Alan Taitz, Oscar Palacios Rodríguez, Ildefonso Rodríguez Leyva, and M. Florencia Assaneo, "Speech's syllabic rhythm and articulatory features produced under different auditory feedback conditions identify Parkinsonism," vol. 14, no. 1, pp. 15787.

[18] Sevada Hovsepyan and Mathew Magimai.-Doss, "Syllable Level Features for Parkinson's Disease Detection from Speech," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 11416–11420, IEEE.

[19] Sevada Hovsepyan, Itsaso Olasagasti, and Anne-Lise Giraud, "Combining predictive coding and neural oscillations enables online syllable recognition in natural speech," *Nature Communications*, vol. 11, no. 1, pp. 3117, June 2020.

[20] Izzet B. Yildiz, Katharina von Kriegstein, and Stefan J. Kiebel, "From Birdsong to Human Speech Recognition: Bayesian Inference on a Hierarchy of Nonlinear Dynamical Systems," *PLoS Computational Biology*, vol. 9, no. 9, pp. e1003219–e1003219, Sept. 2013.

[21] Rachel B. Convey, Tiina Ihalainen, Yuanyuan Liu, Okko Räsänen, Sari Ylinen, and Nelly Penttilä, "A comparative study of automatic vowel articulation index and auditory-perceptual assessments of speech intelligibility in Parkinson's disease," pp. 1–11.

[22] Daniel Escobar-Grisales, Tomás Arias-Vergara, Cristian David Ríos-Urrego, Elmar Nöth, Adolfo M. García, and Juan Rafael Orozco-Arroyave, "An Automatic Multimodal Approach to Analyze Linguistic and Acoustic Cues on Parkinson's Disease Patients," in *INTERSPEECH 2023*. pp. 1703–1707, ISCA.

[23] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[24] Juan Rafael Orozco, Julian D. Arias-Londoño, J. Vargas-Bonilla, María González-Rátiva, and Elmar Noeth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," May 2014.

[25] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Interspeech 2016*. Sept. 2016, pp. 2001–2005, ISCA.

[26] Alireza Bayestehtashk, Meysam Asgari, Izhak Shafran, and James McNames, "Fully Automated Assessment of the Severity of Parkinson's Disease from Speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 172–185, Jan. 2015.

[27] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Interspeech 2010*. Sept. 2010, pp. 2794–2797, ISCA.

[28] S. Pavankumar Dubagunta, Mathew Magimai.-Doss, Eleni Theocharopoulos, and Mathew Magimai Doss, "Towards Automatic Prediction of Non-Expert Perceived Speech Fluency Ratings," .

[29] Okko Räsänen, Gabriel Doyle, and Michael C. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, Feb. 2018.

[30] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8.

[31] Paul Boersma and David Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009.

[32] Yannick Jadoul, Bill Thompson, and Bart de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[33] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, Firenze Italy, Oct. 2010, pp. 1459–1462, ACM.