Using End-to-End Automatic Speech Recognisers' Internals to Model Disfluencies in Italian Patients with Early-stage Parkinson's Disease.

Loredana Schettino[1,*,†], Vincenzo Norman Vitale[2,†] and Marta Maffia[3,†]

[1]Free University of Bozen-Bolzano, Piazza Università, 1, 39100 Bolzano, Italia
[2]University of Naples Federico II, C.so Umberto I, 40, 80138 Napoli, Italia
[3]University of Naples L'Orientale, Italy, Via Chiatamone 61/62 - 80121 Napoli, Italia

Abstract

Alterations in speakers' articulation and phonation are among the earliest symptoms of Parkinsons' Disease (PD). However, clinical decision-making is currently based on holistic ratings of speech intelligibility, while studies on PD detection mostly involve highly complex and hardly interpretable models. This study builds upon previous works on Italian that showed how the characteristics of disfluency phenomena may be considered as an index of impairment at the very onset of the disease by investigating whether even less complex (supervised) end-to-end speech recognition systems (E2E ASR) can model disfluency phenomena in Italian PD speech and how this could support PD discrimination tasks. Exploiting the ability of E2E ASRs to progressively model useful features for discriminating between PD and non-PD speakers provides valuable insight into the ASRs' internal dynamics as well as for the development of decision support systems.

Keywords

Disfluencies, Spontaneous Speech, Parkinson's Disease, Conformer, Probing

1. Introduction

Parkinson's Disease (PD) is a chronic neurodegenerative disorder steadily on the rise in terms of prevalence and incidence [1, 2]: more than 10 million individuals worldwide are affected by PD, mainly among the population aged 65 and over, and this number is expected to increase in demographically ageing societies. Caused by deterioration or loss of dopaminergic neurons in the *substantia nigra* of the basal ganglia, PD is generally diagnosed based on clinical criteria, such as the medical history and physical/neurological examinations of the patient. Although several experimental studies have shown that speech and voice alterations are among the earliest symptoms of PD [3, 4, 5], this precious information is poorly used in clinical decision-making. In the Unified Parkinson's Disease Rating Scale (UPDRS), the rating tool used to assess the severity and to monitor the progression of the disease [6], only one item (3.1) concerns the patient's speech and suggests an assessment based on the clinician's perception, considering above all intelligibility. The application of advanced and sustainable methods of acoustic data analysis could therefore be beneficial, especially in the diagnostic phase: while a cure for PD has yet to be found, early diagnosis is crucial for access to pharmacological and non-pharmacological interventions.

However, developing machine learning tools for critical areas like early Parkinson's disease detection is significantly hampered by data scarcity. Acquiring the necessary data is both costly, requiring specialized linguistic and medical experts, and complex, given the inherently (and fortunately) small patient sample size. To overcome this limitation, our study explores the use of latent features encoded within pre-trained Automatic Speech Recognition (ASR) models. This approach explores the possibility of efficiently utilizing limited available data by leveraging knowledge distilled from vast quantities of data not originally intended for this purpose. Additionally, to further enhance the procedure, it focuses on specific speech features that were observed to play a significant role in discriminating PD speech, even at the early stages of the disease, namely speech disfluency patterns [7].

We believe that using such a method optimises the use of the available data by integrating domain-specific and computational knowledge and can thus support the development of decision support systems in data-scarce critical contexts, as exemplified by early Parkinson's detection.

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy.
*Corresponding author.
†These authors contributed equally.
lschettino@unibz.it (L. Schettino); vincenzonorman.vitale@unina.it (V. N. Vitale); mmaffia@unior.it (M. Maffia)
0000-0002-3788-3754 (L. Schettino); 0000-0002-0365-8575 (V. N. Vitale); 0000-0002-4913-374X (M. Maffia)

2. Related work

2.1. Parkinson's Disease Speech and Disfluency Patterns

The loss of dopamine in the central nervous system causes motor impairments and has an impact also on laryngeal, respiratory and articulatory functions, with about 90% of individuals with PD suffering from voice and speech disorders [8]. PD-related hypokinetic dysarthria includes a range of alterations: hypophonia (reduced voice volume), dysphonia (changes in voice quality), dysrhytmia, reduced speech rate, monopitch, imprecise articulation [9, 10, 11, 12]. Parkinsonian speech is also commonly referred to as 'disfluent', although a detailed and comprehensive description of the specific characteristics of disrupted PD speech has not yet been provided [13]. Studies have mostly focused on specific types of disfluencies: in [14, 15], for example, stuttering-like disruptions (one-syllable word repetitions, sound and syllable repetitions, sound prolongations, and blocks) were observed in PD patients and healthy speakers, and greater disfluency percentages were found in pathological speech, supporting the relationship between stuttering and the functions of the basal ganglia. In a work on repetitive speech phenomena (both hyperfluent and dysfluent) [16], a positive correlation between the frequency of disfluencies and the duration of PD was found. However, studies have not always considered the functions of disfluency phenomena in PD speech and mostly involved mild-to-severe and strongly disfluent patients in experimental protocols.

A recent study conducted on Italian early-stage PD subjects and on spontaneous monological speech [17], showed that, even at the beginning of the pathology (when patients' speech is completely intelligible), the observation of disfluency phenomena can reveal some alteration in linguistic planning and processing: the speech of PD patients was found to differ from that of sex- and age-matched healthy speakers, in terms of the higher frequency of repairs, the specific functions of hesitations (mostly used by PD patients for lexical retrieval), the location of disfluency phenomena (more within-words in PD speech than in the control group) and the duration of silent pauses (longer in PD than in healthy subjects).

2.2. Parkinson's Disease Automatic Detection

Various studies have been devoted to developing automatic and objective tools to support PD diagnosis and the assessment of its severity [18]. Remarkable PD detection accuracy was achieved by leveraging non-interpretable embeddings obtained with Deep Neural Networks (DNNs)-based self-supervised models, e.g. x-vectors, Wav2Vec 2.0, HuBERT, and TRILLsson representations. Furthermore, a more recent study showed that models based on interpretable features such as prosodic, linguistic, and cognitive descriptors can support the evaluation of speech deterioration in PD patients, whereas models based on non-interpretable features achieve higher detection accuracy [19]. These findings are often based on consideration of highly functional vocal paradigms, such as sustained phonation of isolated segments, and involve rather complex models relying on non-interpretable features or features commonly observed as useful for PD speech discrimination as they become evident in the mid to advanced stages of the disease [18]. Nonetheless, a recent study showed that PD detection trials relying on a restricted number of meaningful features that were extracted from connected speech rather than isolated speech units achieve accurate, as well as economical and interpretable discrimination [7]. Also, studies on the interpretability of DNN-based models, using probing techniques, provided evidence that even smaller and less complex models, such as Conformer-based ones [20], can model speech features and that different features are encoded in DNN layers at different depths [21]. In particular, it was found that higher levels capture phone identity and word identity information, and the last layer before the object function even captures discriminating features of disfluency phenomena, more specifically, filled pauses and prolongations [22].

In substance, this study builds on the following findings from previous work on Italian PD speech:

- relying on natural speech material that results from the usual working dynamic of the vocal apparatus during phonation proves useful for discrimination [7];
- peculiar uses of natural speech characteristics phenomena like disfluency phenomena may be considered as an index of impairment at the very onset of the disease [17];
- less complex supervised end-to-end speech recognition systems (E2E ASR) can model disfluency-related features useful for their discrimination [22].

On this basis, we investigate whether less complex (supervised) E2E ASR systems can model disfluency features in Italian PD speech and how well this could support PD discrimination tasks.

3. Method

3.1. Data and Annotation

The study is based on the data described in [12]. It consists of approximately 40 minutes of monologic speech

produced by 36 Italian native speakers from the Campania region: 18 participants with idiopathic non-demented PD (10 males, 8 females; 51–81 years of age, M= 65) and 18 age-matched Healthy Controls (HC, 10 males, 8 females; 54–77 years of age, M= 64). The patients were recruited from the Movement Disorders Unit of the First Division of Neurology at the University of Campania "Luigi Vanvitelli". PD participants had no prior history of language or speech disorders, had been diagnosed with Parkinson's disease within the past four years (since 2021) and showed no significant cognitive impairment, major or minor depression, or dysthymic disorder. All participants were asked to discuss the positive and negative aspects of the place they were living during data collection. They were encouraged to speak in their usual, conversational tone and at a comfortable volume. Sociolinguistic information for each speaker was gathered via a questionnaire, and all participants provided written consent for the data collection process.

The analysis focused on a series of so-called "disfluency phenomena" defined as speech management phenomena, namely, speech material, e.g. repetitions, segmental prolongations, pauses, and fillers that speakers can use to monitor and effectively manage the online processes of speech planning, coding, articulation, and reception [23]. The phenomena were identified and annotated based on their context of occurrence, following [17] and included the following phenomena specifically involved in the speech planning process (Cohen's k= 0.82, good agreement [24]):

- Prolongations (PRLs), marked prolongation of segmental material, e.g., *la**aa** casa* (the**ee** house);
- Filled Pauses (FPs), non-lexical filler, vocalizations and/or nasalizations, e.g., *eeh, ehm, mhh*;
- Silent Pauses (SPs), marked silences perceived as a hesitant pause in the context of occurrence;
- Lexicalized Filled Pauses (LFPs), lexical fillers, work as discourse markers involved in the coverage of planning times, e.g., *diciamo, insomma, appunto...* (well, let's say, so, ...);
- Repetitions (REPs), repetition of already uttered words or fragments of words, e.g., *di di* (of of) or *d- di* (o- of).

3.2. Probing Approach

Based on previous studies investigating E2E-ASR models' internal behaviour [21, 25, 22, 26], we employ a probing approach to investigate the pre-trained models' ability to capture speaker and speech related markers, i.e., characteristics associated with disfluent speech segments combined with PD biomarkers, and whether these features facilitate PD speech identification.

The employed technique involves:

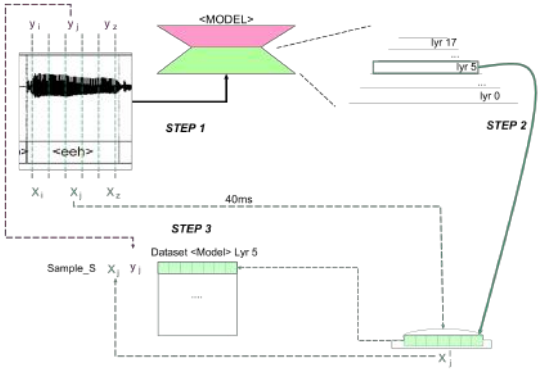


Figure 1: Probing Procedure: Step 1 – The annotated ($Y_{i..j..z}$) input sequence ($X_{i..j..z}$) for sample S is fed to the Probed Model in 40ms chunks. Step 2 – The intermediate encoder's layers' emissions (X_j^I) are captured and associated with the proper label (Y_j). Step 3 – the triplet Sample Index (X_j), label (Y_j)), Intermediate Emission (X_j^I)) builds up in a dataset representing the same sample M in the latent space from the n-th encoder's layer.

- selecting pre-trained models (m). In particular, two publicly available Conformer-based [20] models with different decoding component were selected: one with a Connectionist Temporal Classification (CTC) [27] decoder[1], namely, a non-auto-regressive technique; one with a Recurrent Neural Network Transducer (RNN-T), commonly known as *Transducer*[2], which is an auto-regressive speech transcription technique;
- building Long Short Term Memory (LSTM) and Bidirectional LSTM (BILSTM) classifiers whose inputs are represented by intermediate emissions of the considered model's encoder layers (l), combined with the appropriate sequence of labels based on dataset annotation;
- evaluating the classifications relying on metrics oriented to results safety rather than performance.

3.2.1. Data Preparation

The considered dataset has been prepared based on a set of praat TextGrid annotation files indicating the speaker and the type of disfluency according to the speech signal. More specifically, PRLs, FPs, SPs, LFPs and REPs were considered, resulting in a dataset with a dimension of 850 segments. For each segment, the contextual information preceding and following the disfluency phenomenon has been considered, giving each segment a length of 4

[1] v1.6.0 https://huggingface.co/nvidia/stt_en_conformer_ctc_large
[2] v1.6.0 https://huggingface.co/nvidia/stt_en_conformer_transducer_large

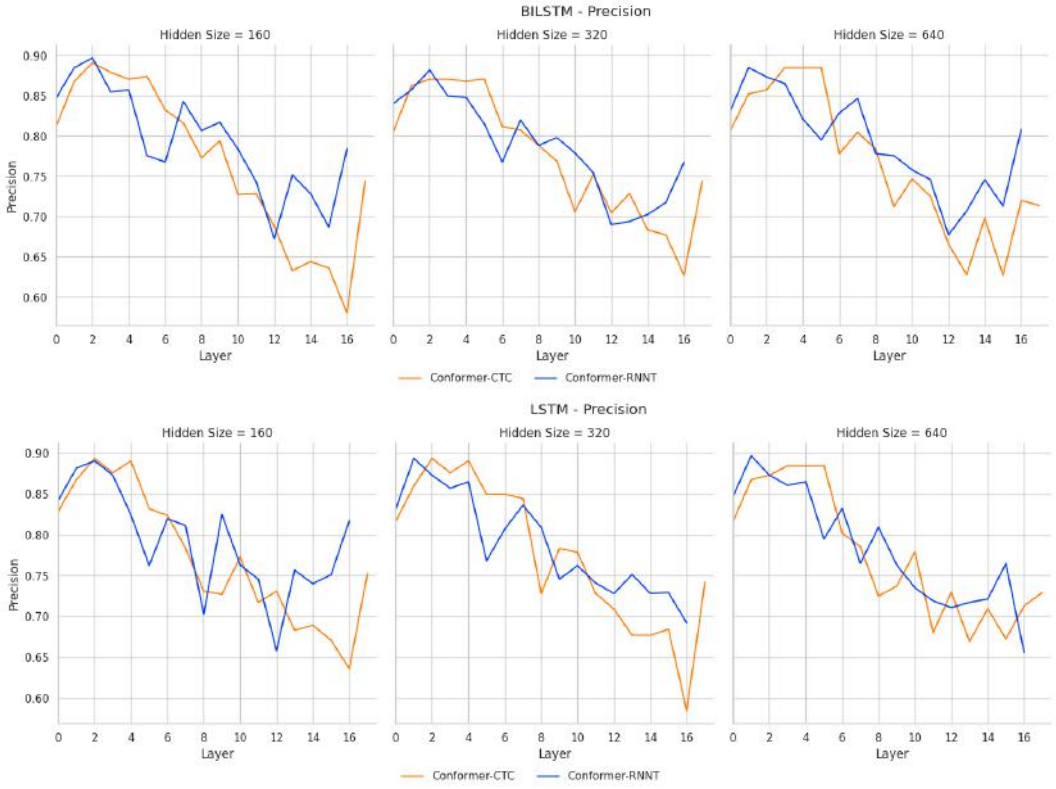


Figure 2: The precision achieved by differently sized (160,320,640) BILSTM (Top) and LSTM (Bottom) classifiers trained on the considered dataset in different latent spaces. The pre-trained models (orange and blue) along with the encoding layer on the x-axis indicate the latent space in which the dataset has been considered for training and evaluation.

seconds. Then, for each encoding layer from a considered pre-trained model, we extract a representation of segments in the corresponding latent space following the procedure described in Figure 1. In particular, for each segment, we obtain:

- A *sequence of intermediate emissions*, namely fragment representations in the corresponding layer's latent space. Each fragment corresponds to a portion of t milliseconds of the input signal, where t depends on the considered model's characteristics.
- A *sequence of labels* associated with each fragment, indicating whether that fragment belongs to a disfluency or not and, if so, whether the speaker is PD or HC.

The resulting dataset consists of pairs of sequences of emissions (i.e., distilled features) and corresponding labels identified by the model and the layer from which they were extracted.

3.2.2. Pre-trained Models

We selected two publicly available Conformer-based [20] pre-trained models built with the NVIDIA Nemo toolkit[3], both with a fragment dimension $t = 40 milliseconds$ and only differing in the decoding component.

On the one hand, we considered a CTC decoder, one of the most popular decoding techniques. It consists of a non-auto-regressive speech transcription technique that collapses consecutive, all-equal, transcription labels (character, word piece, etc.) to one label unless a special label separates them. The result is a sequence of labels shorter than or equal to the input vector sequence length. Being non-auto-regressive, it is also considered computationally effective, requiring less time and resources for training and inference phases. On the other hand, we considered a Transducer, which is an auto-regressive speech transcription technique that overcomes CTC's limitations, being non-auto-regressive and subject to limited label sequence length. The Transducer decoding technique can produce label-transcription sequences longer than

[3]Nemo version 1.21.0.

the input vector sequence and models inter-dependency in long-term transcription elements. A Transducer typically comprises two sub-decoding modules: one that forecasts the next transcription label based on the previous transcriptions (prediction network) and the other that combines the encoder and prediction-network outputs to produce a new transcription label (joiner network). These features improve transcription speed and performance (compared to CTC), while requiring more training and computational resources [28]. Also, the two techniques should provide different representations and contributions during the training phase (backdrop) due to their different dynamics in forward propagation.

Note that both considered pre-trained models rely on the same encoder architecture, but the Conformer-CTC model has 18 encoding layers, while the Conformer-Transducer encoder has 17 layers. This resulted in 35 different latent space representations for the considered dataset.

3.2.3. Classifiers

The classifiers internally consist of a LSTM or a BILSTM module, followed by a Feed Forward Neural Network (FFNN). The choice of LSTM and BILSTM modules is driven by their capacity to capture the temporal dependencies in the input, which fits well with our objective of modeling temporal dependencies in the latent space representation of the speech signal.

Since the LSTM/BILSTM hidden-layer size is a crucial parameter, we investigate the impact of three different layer sizes (hidden-layer size, h), namely 160, 320 and 640. So, an LSTM-based classifier processes a sequence of $\{e_{l,m}\}$ emission vectors (each of length n) and produces a new sequence of vectors with size h. The two sequences are aligned over time. At each time step t, based on the LSTM/BILSTM hidden-layer output, the FFNN produces a label indicating whether the considered input represents a disfluency segment, pronounced by a PD or HC speaker, or not. In summary, we train and evaluate many different RNN classifiers/detectors ($L_{h,m,l}$) for all possible h, m, and l combinations to search for the evidence of disfluencies-related pathological biomarker properties in the models' decisions.footnoteFrameworks used to implement and train the classifiers: torch==2.2.1 and pytorch-lightning==2.0.7 and BILSTM based classifier were trained, resulting in $\sim$ 200 models. Note that the temporal sensitivity of our classifier/detector, namely the minimum difference between consecutive time steps, is 40 ms because the considered ASR models produce emissions at that rate.

During the training phase, the considered corpus was split into train, validation, and test sets using 60%, 20%, and 20% percentages while ensuring that these sets did not share the same speakers. Each classifier has been trained for a maximum of 100 epochs using an Adam optimizer with an initial $lr = 0.00001$. To reduce the risk of overfitting, we introduce a *dropout* neuron-selection strategy for the LSTM/BILSTM gates, which statistically excludes (with a 0.1 probability) one neuron and its weights during each training iteration [29]. Finally, an early stopping mechanism was used to avoid wasting computational resources. In particular, the training phase ends if the validation-loss does not decrease by a minimum of 0.001 during the last 20 epochs, which is the patience threshold.

3.2.4. Evaluation

Since our aim is to investigate whether pre-trained E2E ASR models encode features useful for the identification of disfluency phenomena in PD speech, and whether and how they enable the discrimination between PD and HC, we decided to rely on metrics oriented to results safety rather than performance. Note that a sample is classified as PD or HC if the portion related to the disfluency is (1) detected and (2) at least 60% of frames are correctly labeled as either PD or HC. The reliability of the approach is assessed by inspecting the confusion matrices for the best LSTM or BiLSTM, CTC-based and RNNT-based classifiers, which provide a breakdown of correct and incorrect predictions for each class. The quantitative analysis was further supported by a qualitative exploration of the acoustic features emerging as relevant for discrimination with reference to previous literature [18]. To this aim, the eGeMAPSv02 [30] feature set from the OpenSmile toolkit [31] was selected as the basic feature set and inspected using the Orange software [32].

4. Results and Discussion

In this study, we considered two distinct ASR architectures, namely Conformer-CTC and Conformer-Transducer, selected for the differing capabilities of their decoding components. These decoding mechanisms, i.e., CTC and Transducer, being respectively non-autoregressive and autoregressive by nature, are inherently capable of capturing diverse aspects of the speech signal, therefore influencing in a different way the encoding component.

Figure 2 reports the precision of each trained classifier. It is interesting to observe how the layers closer to the input provide the higher precision, while the overall tendency, getting closer to the decoding component, is a constant reduction, which is likely related to the specific objective of the pre-trained models, namely, to provide a clean transcription. However, the model that seems to provide the most informative and stable latent representation seems to be the Conformer-CTC (orange line) in

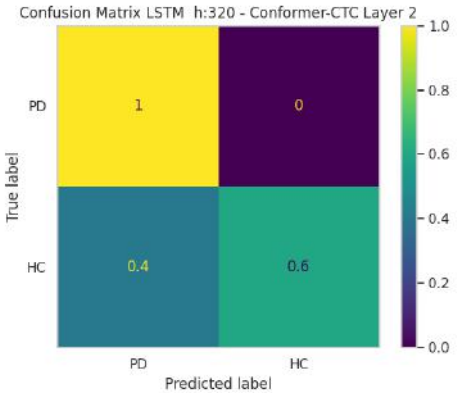
layers from 2 to 6, showing a constant precision over all the considered configurations.

To enable a more nuanced and phenotypically informed classification, we considered two types of Recurrent Neural Networks (RNN) for our classifier, namely Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM). Both architectures are designed to preserve memory of previously observed sequences. However, BiLSTM offers a crucial advantage by enabling the re-evaluation of past observations in light of subsequent inputs. For instance, the quality of a vowel sound's realization at a given point can alter the assessment of the entire preceding segment of the speech signal, a capability effectively captured by the bidirectional nature of BiLSTM.

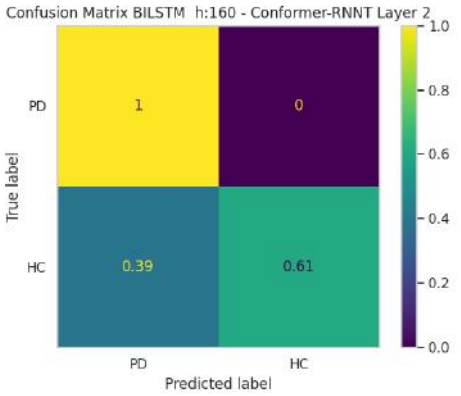
Consistent with the literature, the earliest layers proved to be sensitive to speech-related information, allowing them to distinguish between speakers with PD and those without PD. Likewise, the earliest layaers appear to contain sufficient information to discriminate between disfluent and non-disfluent segments, letting performance in line with the literature [22].

To gain insight into the reliability of both the approach and the latent representation space, figure 3 reports the confusion matrices of the best-performing classifiers for the two considered RNN architectures (i.e., LSTM, BiLSTM), showing that in both cases the most critical choice, namely identifying a PD speaker, is correctly addressed, whereas we observe some false positive predictions, where healthy speakers' productions were misclassified as PD productions (in both cases about 0.4). This may be explained by considering that similar sets of phonetic cues in speech may index different information.

The qualitative exploratory inspection showed that pitch, loudness, voice quality-related features, including shimmer and Harmonics-to-Noise Ratio (HNR), and Spectral flux were most relevant for distinguishing between PD and HC speech. This observation is in line with previous findings described in [7] where features concerning the spectral distribution, energy and frequency emerged as the most relevant to discriminate between PD and HC speech.



(a) LSTM with hidden size $h = 320$ trained on the dataset represented in the latent space of Conformer-CTC's encoding layer #2.



(b) BiLSTM with hidden size $h = 160$ trained on the dataset represented in the latent space of Conformer-CTC's encoding layer #2.

Figure 3: Confusion matrix of the best-performing classifier for LSTMA (top) and BiLSTM (bottom).

5. Conclusion

The main findings highlight that focusing on speech correlates such as disfluency phenomena provides a convenient choice to enhance the development of decision support systems. The latent representation from the intermediate encoding layer was shown to be highly informative and quite reliable for the classification of PD and HC speakers. In addition, the CTC decoder seemed to provide slightly more stable performance in this task, probably due to its non-autoregressive nature. Future steps will involve a comparison of performance with different pre-trained models and classification architectures. Indeed, since disfluency phenomena encompass different types of phenomena (i.e. textual phenomena, like lexical fillers and repetitions, and phonetic phenomena, like non-lexical fillers and prolongations), different approaches may perform better on specific types.

Also, the analysis led to observing (not yet noticeable) alterations of acoustic features revealing the onset of PD-related motor impairment. It is worth noticing that some of the considered disfluency phenomena, namely prolongations and filled particles, consist in prolonged vocalisations, which are similar to the sustained vowel traditionally used in highly controlled studies on PD

speech. Thus they provide a nice integration of data efficacy and ecology.

References

[1] W. A. Rocca, The burden of parkinson's disease: a worldwide perspective, The Lancet Neurology 17 (2018) 928–929.

[2] J. D. Steinmetz, K. M. Seeher, N. Schiess, E. Nichols, B. Cao, C. Servili, V. Cavallera, E. Cousin, H. Hagins, M. E. Moberg, Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021, The Lancet Neurology 23 (2024) 344–381.

[3] S. Skodda, Analysis of voice and speech performance in parkinson's disease: a promising tool for the monitoring of disease progression and differential diagnosis, Neurodegenerative Disease Management 2 (2012) 535–545.

[4] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, E. Ruzicka, Imprecise vowel articulation as a potential early marker of parkinson's disease: Effect of speaking task, The Journal of the Acoustical Society of America 134 (2013) 2171–2181.

[5] A. Favaro, L. Moro-Velázquez, A. Butala, C. Motley, T. Cao, R. D. Stevens, J. Villalba, N. Dehak, Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in parkinson's disease, Frontiers in Neurology 14 (2023) 1142642.

[6] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results, Movement disorders: official journal of the Movement Disorder Society 23 (2008) 2129–2170.

[7] M. Maffia, L. Schettino, V. N. Vitale, Automatic detection of parkinson's disease with connected speech acoustic features: towards a linguistically interpretable approach, in: Proceedings of the 9th Italian Conference on Computational Linguistics. CEUR WORKSHOP PROCEEDINGS, 2023.

[8] L. O. Ramig, C. Fox, S. Sapir, Speech treatment for parkinson's disease, Expert review of neurotherapeutics 8 (2008) 297–309.

[9] F. L. Darley, A. E. Aronson, J. R. Brown, Clusters of deviant speech dimensions in the dysarthrias, Journal of speech and hearing research 12 (1969) 462–496.

[10] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, E. Nöth, Automatic detection of parkinson's disease from words uttered in three different languages, in: Fifteenth annual conference of the international speech communication association, 2014.

[11] H. Ackermann, W. Ziegler, Articulatory deficits in parkinsonian dysarthria: an acoustic analysis., Journal of Neurology, Neurosurgery & Psychiatry 54 (1991) 1093–1098.

[12] M. Maffia, R. De Micco, M. Pettorino, M. Siciliano, A. Tessitore, A. De Meo, Speech rhythm variation in early-stage parkinson's disease: a study on different speaking tasks, Frontiers in Psychology 12 (2021) 668291.

[13] A. M. Goberman, M. Blomgren, Parkinsonian speech disfluencies: effects of l-dopa-related fluctuations, Journal of fluency disorders 28 (2003) 55–70.

[14] A. M. Goberman, M. Blomgren, E. Metzger, Characteristics of speech disfluency in parkinson disease, Journal of Neurolinguistics 23 (2010) 470–478.

[15] F. S. Juste, F. C. Sassi, J. B. Costa, C. R. F. de Andrade, Frequency of speech disruptions in parkinson's disease and developmental stuttering: a comparison among speech tasks, Plos one 13 (2018) e0199054.

[16] T. Benke, C. Hohenstein, W. Poewe, B. Butterworth, Repetitive speech phenomena in parkinson's disease, Journal of Neurology, Neurosurgery & Psychiatry 69 (2000) 319–324.

[17] L. Schettino, M. Maffia, R. De Micco, A. Tessitore, Disfluency and speech management in italian patients with early-stage parkinson's disease, in: Proceedings of Disfluency in Spontaneous Speech (DiSS) Workshop 2023, 2023.

[18] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, J. I. Godino-Llorente, Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects, Biomedical Signal Processing and Control 66 (2021) 102418.

[19] A. Favaro, Y.-T. Tsai, A. Butala, T. Thebaud, J. Villalba, N. Dehak, L. Moro-Velázquez, Interpretable speech features vs. dnn embeddings: What to use in the automatic assessment of parkinson's disease in multi-lingual scenarios, Computers in Biology and Medicine 166 (2023) 107559.

[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, Conformer: Convolution-augmented transformer for speech recognition, Interspeech 2020 (2020).

[21] A. Pasad, J.-C. Chou, K. Livescu, Layer-wise analysis of a self-supervised speech representation model, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2021, pp. 914–921.

[22] N. Vitale, L. Schettino, F. Cutugno, Rich speech signal: exploring and exploiting end-to-end auto-

matic speech recognizers' ability to model hesitation phenomena, in: 25th Annual Conference of the International Speech Communication Association (INTERSPEECH 2024), ISCA, 2024, pp. 222–226.
[23] W. J. Levelt, Speaking: From intention to articulation, volume 1, Cambridge/London: MIT press, 1993.
[24] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.
[25] A. Prasad, P. Jyothi, How accents confound: Probing for accent information in end-to-end speech recognition systems, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3739–3753.
[26] V. N. Vitale, F. Cutugno, A. Origlia, G. Coro, Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique, Neural Computing and Applications (2024) 1–27.
[27] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.
[28] A. Graves, Sequence transduction with recurrent neural networks, arXiv preprint arXiv:1211.3711 (2012).
[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research 15 (2014) 1929–1958.
[30] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE transactions on affective computing 7 (2015) 190–202.
[31] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.
[32] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, B. Zupan, Orange: Data mining toolbox in python, Journal of Machine Learning Research 14 (2013) 2349–2353. URL: http://jmlr.org/papers/v14/demsar13a.html.