Automatic Assessment of Parkinson's Disease Using Speech Representations of Phonation and Articulation

Yuanyuan Liu, Mittapalle Kiran Reddy, Nelly Penttilä, Tiina Ihalainen, Paavo Alku, and Okko Räsänen

***Abstract*—Speech from people with Parkinson's disease (PD) are likely to be degraded on phonation, articulation, and prosody. Motivated to describe articulation deficits comprehensively, we investigated 1) the universal phonological features that model articulation manner and place, also known as speech attributes, and 2) glottal features capturing phonation characteristics. These were further supplemented by, and compared with, prosodic features using a popular compact feature set and standard MFCC. Temporal characteristics of these features were modeled by convolutional neural networks. Besides the features, we were also interested in the speech tasks for collecting data for automatic PD speech assessment, like sustained vowels, text reading, and spontaneous monologue. For this, we utilized a recently collected Finnish PD corpus (PDSTU) as well as a Spanish database (PC-GITA). The experiments were formulated as regression problems against expert ratings of PD-related symptoms, including ratings of speech intelligibility, voice impairment, overall severity of communication disorder on PDSTU, as well as on the Unified Parkinson's Disease Rating Scale (UPDRS) on PC-GITA. The experimental results show: 1) the speech attribute features can well indicate the severity of pathologies in parkinsonian speech; 2) combining phonation features with articulatory features improves the PD assessment performance, but requires high-quality recordings to be applicable; 3) read speech leads to more accurate automatic ratings than the use of sustained vowels, but not if the amount of speech is limited to correspond to the sustained vowels in duration; and 4) jointly using data from several speech tasks can further improve the automatic PD assessment performance.**

***Index Terms*—Parkinson's disease, phonological features, speech attributes, glottal features, automatic speech assessment.**

I. Introduction

PARKINSON'S disease (PD) is the second most common neurodegenerative disease and has a high incidence among the aging population [1]. Since speech production is in general affected by the integrity of the underlying neural system, PD is likely to lead to voice and speech disorders in the forms of dysphonia and dysarthria [2]. Dysphonia refers to abnormal functioning of the voice and is therefore related to phonation, while dysarthria corresponds to articulation abnormalities. It has been reported that up to 90% of people with PD suffer from dysarthria, while the prevalence of dysphonia can be around 65.5% [3], [4]. However, it has been reported that only $3-4\%$ of people with PD receive speech treatment [5]. The most common perceptual speech pathologies in PD include impaired phonation, imprecise articulation, reduced variability of pitch and loudness, and other prosodic disturbances on speech rate, stress, and pauses. These pathologies result in degraded intelligibility of speech produced by speakers with PD [3], [6], [7].

Speech signal has been demonstrated to be a valuable indicator of disease progression and treatment efficacy in PD [7]. There is a large body of research on automatic PD assessment using speech, which employs acoustic analysis and pattern recognition techniques and aims at objective, non-invasive, and cost-efficient health care technology for the benefit of clinical practice [8], [9], [10], [11], [12], [13], [14]. Most studies have investigated PD detection, which is typically formulated as a binary classification problem. Only a few studies have, however, investigated the prediction of the PD severity level (e.g., [12], [15]), which would be more suitable for clinical purposes in monitoring disease progression, the impact of medication, or speech rehabilitation interventions where relative degrees and over-time changes of motor/speech impairment are of interest.

Given the basic mechanisms of speech production, it would be useful to analyze speech in PD using speech representations that model phonation, and articulation. However, although there is a large body of studies on phonation and articulation in parkinsonian speech demonstrating the relevance of both aspects in speech assessment (see [2] for a review), only a few studies have investigated the *joint* use of both aspects (e.g., [16]). In addition, the most commonly used features for articulatory analysis, such as vowel working space area (VSA), vowel articulation index (VAI), and formant centralization ratio (FCR) [17], [18], focus on vowel formants despite the fact that PD can affect the articulation of both vowels and consonants [19], [20].

Besides the different speech representations, another issue that affects automatic PD assessment from speech signals is the selection of the appropriate speech task used in the collection of data in the system training and testing phases. Speech tasks used

Manuscript received 3 March 2022; revised 21 June 2022 and 16 August 2022; accepted 26 September 2022. Date of publication 17 November 2022; date of current version 2 December 2022. This work was supported by the Academy of Finland under Grants 314602 and 330139. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Juan Ignacio Godino-Llorente. *(Corresponding author: Yuanyuan Liu.)*

Yuanyuan Liu is with the Unit of Computing Sciences, Tampere University, 33100 Tampere, Finland (e-mail: yuanyuan.liu@tuni.fi).

Mittapalle Kiran Reddy and Paavo Alku are with the Department of Signal Processing and Acoustics, Aalto University, 02150 Espoo, Finland (e-mail: kiran.r.mittapalle@aalto.fi; paavo.alku@aalto.fi).

Nelly Penttilä and Tiina Ihalainen are with the Faculty of Social Sciences, Tampere University, 33100 Tampere, Finland (e-mail: nelly.penttila@tuni.fi; tiina.ihalainen@tuni.fi).

Okko Räsänen is with the Unit of Computing Sciences, Tampere University, 33100 Tampere, Finland, and also with the Department of Signal Processing, Acoustics, Aalto University, 02150 Espoo, Finland (e-mail: okko.rasanen@tuni.fi).

Digital Object Identifier 10.1109/TASLP.2022.3212829

in the study area include the production of sustained vowels and short sentences, text reading, and spontaneous speech [2], [21], [22]. Sustained vowels are more stable for feature computation and can be collected and analyzed similarly in different language populations, while continuous speech, such as text reading or monologue, are more representative of daily speech communication [22]. In [2], the authors reported that the selection of the best speech task depends on the analysis target and also on the features to compute. In [12], PD detection and UPDRS regression results were reported for different speech tasks in the Spanish PD database PC-GITA [23]. However, the selection of speech tasks in PD assessment was not considered in that work. In investigating the Frenchay Dysarthria Assessment (m-FDA) score regression for the speakers in PC-GITA, the authors of [24] studied the contributions of different speech tasks and recommended using three tasks (the oral diadochokinetic (DDK[1]) task, the text reading task, and the monologue task) in clinical practice. In many other studies, data from different speech tasks have been used together without any distinction, such as PD disease detection in [25] and phonation impairments prediction in [26].

In connection to the above-mentioned issues in automatic PD assessment, the aims of the present study were to 1) investigate the combination of phonation and articulatory features in automatic assessment, and 2) to study the effects of different speech tasks in assessment. Moreover, we 3) carried out these investigations in the context of regression tasks instead of classification in order to study how accurately these features and speech tasks can capture subject-level variability in speech intelligibility and pathology in PD patients.

A. Related Work

A comprehensive review of phonatonary and articulatory aspects for automatic PD assessment using speech signals was published by Moro-Velazquez et al. [2]. They concluded that both phonation and articulation are associated with the severity of PD. As described in [2], the commonly used phonation features include, for instance, jitter, shimmer, harmonic-to-noise ratio (HNR), noise-to-harmonic ratio (NHR), pitch, and Mel-frequency cepstral coefficients (MFCCs). The most popular articulation features include MFCCs, features based on linear predictive coding (LPC), and perceptual linear prediction (PLP), as well as formant-based features like FCR, VSA, and VAI to measure vowel articulation characteristics. From the above articulation features, MFCCs have been particularly popular in many areas of speech research as a method to parameterize vocal tract information. In [25], MFCCs were applied in the classification of PD and healthy control (HC) speakers from the Spanish PC-GITA database. The authors trained support vector machines (SVMs) with statistical functionals of MFCCs, which were computed on the Bark bands and extracted from speech consisting of individual sentences, DDK, text reading, and monologue. As a result, an accuracy of 73.7% was reported based on K-fold ($K = 10$) cross-validation [25]. In [27], a comparative study was conducted on classification of PD and HC using sustained vowel utterances of /a/ and several features, including jitter, shimmer, MFCC, tunable Q-factor wavelet transform (TQWT) features among others. Of all the features, TQWT and MFCC were shown to outperform others in terms of classification accuracies.

As a paralinguistic feature set of low dimension, eGeMAPS [28] is also popular in pathological speech analysis (see Section II-B for details). In [29], statistics (mean, maximum, and range) were computed over frame-level eGeMAPS features to discriminate PD and HC speakers in PC-GITA. The study used logistic regression with leave-one-speaker-out cross-validation, where classification accuracies of 72.0% and 80.0% were obtained on text reading and monologue speech, respectively.

Besides the traditional acoustic features mentioned above, phonological features (also known as $speech\ attributes$), such as articulation manner and place, can be considered as "universal" descriptors across all spoken languages [30]. These articulation-related phonological features have been successfully utilized in different domains, such as spoken language recognition, speech recognition, as well as pathological speech analysis [30], [31], [32]. In [32], phonological features (on vocal source, manner, place-consonant, and vowel) together with phoneme features were employed for the prediction of speech intelligibility scores. In their work, transcriptions were used for forced-alignment in feature computation. In [33], the effect of articulation manner was studied in PD detection using several Spanish PD speech corpora, including PC-GITA. In the study, speech frames were represented by Rasta-PLP and then grouped according to the articulation manners of the corresponding forced-aligned phonemes (affricate, fricative, liquid, nasal, plosive, and vowel). The force-alignment was conducted using a language-matched ASR system and the transcripts of the speech under analysis. This phonemic grouping was combined with GMM-UBM classifiers and K-fold ($K = 11$) cross-validation. The best accuracy for classifying the 50 PD speakers and 50 HC speakers of PC-GITA was $85 \pm 7\%$ (95% confidence interval), obtained on read speech utterances with a UBM trained for plosives. In [15], phonological posteriors from non-silence frames were used for manually designed features, which were utilized to measure the similarity between non-modal phonation and pathological speech. In the study, a read voice quality database (including speech of modal and non-modal phonation) and data from healthy speakers in PC-GITA were utilized to normalize the phonological posterior features. The proposed features were used to predict the larynx-related score in the m-FDA score for the 50 PD speakers in PC-GITA, which improved the Spearman correlation coefficient on read speech when using baseline features on articulation and prosody [15]. However, the extraction of the above-mentioned phonological features required manual efforts, transcripts for test speech, or a language-matched ASR system, which limits the flexibility of using phonological features for completely automatic processing.

Besides articulation, phonation features have also been utilized in several previous studies on pathological speech. In PD patients, abnormalities in vocal fold closure patterns have

[1]DDK task: rapid repetition of /pa-ta-ka/, /pe-ta-ka/, /pa-ka-ta/, /pa/, /ta/, and /ka/.

been observed through laryngeal videoscopic examinations [34]. Therefore, glottal features representing the mode of the vibration of the vocal folds are effective in the analysis and detection of PD. In [35], glottal features were used to capture dysphonia in the sustained phonation of PD speakers. Their study showed that glottal features outperformed the traditional perturbation and cepstral approaches in the assessment of PD-related dysphonia. In [36], glottal features were shown to have a fairly good performance in discriminating between dysarthric and healthy speech on utterances of non-words, words, and sentences. Moreover, glottal features have also been shown to be effective in clinical research topics like the detection of depression [37], specific language impairment in children [38], and heart failure [39] using speech signals.

In this work, we go beyond the existing language-specific or manual-work intensive methods in the usage of articulatory features by investigating the usage of speech attribute scores automatically extracted from a universal (language-independent) automatic speech attribute estimation system. Moreover, we combine the articulatory features with automatically extracted phonation features to study their complementarity in PD assessment, also comparing and combining them with the more commonly used MFCC and eGeMAPS feature sets. We also take a systematic and cross-lingual stance to the issue of speech task, and investigate the usefulness of the features in sustained vowels, read speech, and spontaneous speech in Finnish and Spanish data. Finally, we carry out these investigations in the context of regression tasks, which are rarely utilized (but see [12], [15]) in pathological speech analysis. We argue that a more detailed picture of disease stage or progression is required for the assessment systems to have practical clinical value, and hence performance assessment should also be conducted in terms of regression instead of (binary) classification.

This article is organized as follows: Section II introduces the workflow, speech representations, and neural regressors we used for PD assessment. Section III describes the PD speech corpora. The experimental setup is introduced in IV. The experimental results are presented and discussed in Section V. This work is concluded in Section VI.

II. Methods

A. Overview of Workflow

The workflow for our automatic PD speech assessment system is illustrated in Fig. 1. The input speech can be from different speech tasks, such as sustained vowels, text reading, or spontaneous monologue. After pre-processing (e.g., resampling), frame-level features are extracted from the signals. In this work, we investigated the efficacy of four different speech representations: automatically extracted speech attribute scores (SAS), glottal features, MFCCs, and low-level descriptors from a well-known compact feature set for automatic speech analysis, eGeMAPS. These features will be introduced in the following section. The frame-level features were grouped into one-second clips with a temporal overlap of 20%. As a result, each 1-s segment was represented by a fixed-dimensional feature matrix, which served as the input to a dedicated neural network model

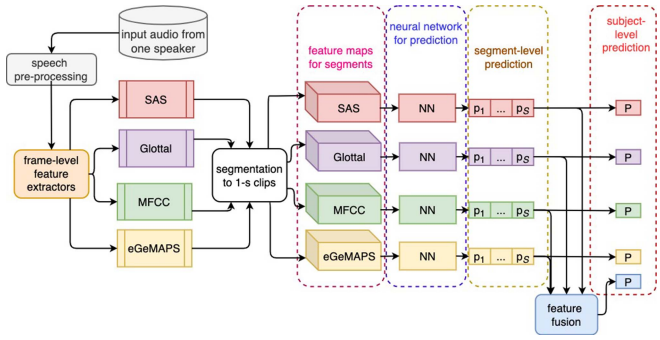

Fig. 1. Workflow of the automatic Parkinson's disease assessment experiments. SAS, Glottal, MFCC, and eGeMAPS stand for different feature types introduced in Section II-B. NN denotes neural network. Here, *S* stands for the total input segment count.

TABLE I
SPEECH ATTRIBUTES INCLUDED IN THE ASAT SYSTEM [40], [41] AND USED AS SAS FEATURES IN THE EXPERIMENTS

Type	Attributes
Manner	fricative, glides, nasal, stop, voiced, vowel
Place	coronal, dental, glottal, high, labial, low, mid, palatal, velar

for a specific target to be predicted, such as the rating of speech intelligibility or the severity level of voice impairment. The neural network gave as its output segment-level predictions for the targets. Finally, a subject-level prediction was computed as the median from the segment-level predictions.

As PD can affect various aspects of speech, a straightforward feature fusion strategy was adopted in this work. Feature fusion is investigated by averaging the predicted segment-level scores across the features to be combined. From these averaged segment-level scores, the prediction regarding the test speaker was generated by taking the median across all the segments for the given speaker.

B. Speech Representations

1) Speech Attribute Scores: In [40], [41], a universal automatic speech attribute transcription (ASAT) system was developed and implemented in Kaldi [42]. The ASAT system was trained with speech of six languages from the OGI Multilanguage Telephone Speech corpus [43]: English, German, Hindi, Japanese, Mandarin, and Spanish. The time-aligned phoneme labels in the OGI corpus were converted to corresponding speech attribute labels according to a phonological table. Totally there are 17 articulatory attributes of interest modeled in [40], [41], including 6 manner and 9 place attributes as shown in Table I, together with "silence" and "other" for silence and unlabeled frames.

In ASAT, input audio is first resampled to 8 kHz and 40 log Mel-filter bank coefficients together with 3 fundamental frequency features are computed for every 25-ms frame using a shift of 10 ms. Then the 43-dimensional frame-level features are normalized and input to a DNN, from which attribute label

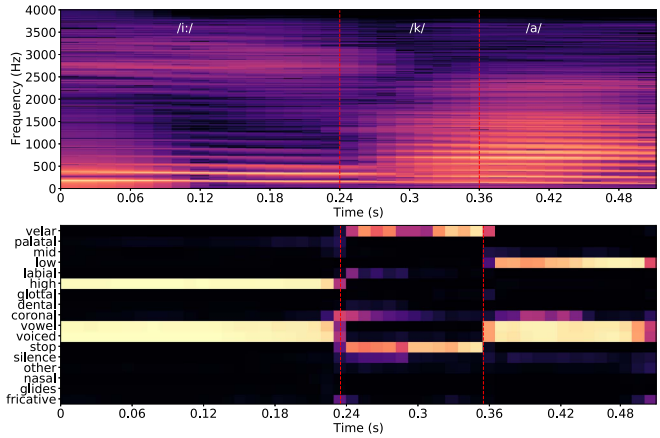

Fig. 2. Spectrogram (top) and speech attribute scores (bottom) for speech segment of uttering "/i: k a/". Brighter shades denote higher posterior scores for the particular speech attributes and black stands for zero score.

hypotheses and the underlying 17-dimensional sigmoid posterior scores for the modeled speech attributes are generated for each input frame. The 17-dimensional attribute posteriors ($p_{attribute} \in [0, 1]$) and their first- and second-order temporal derivatives with a context of 9 frames (as computed with Librosa [44]) were used as frame-level features in our experiments, resulting in a total of 51 SAS features.

As an example, the SAS for the speech segment /i: k a/ in the Finnish word "*siika*" are illustrated in Fig. 2 (time-derivatives not shown). For the high vowel /i:/, attributes "voiced," "vowel," and "high" are detected with large scores. The low vowel /a/ is labeled with attributes "voiced," "vowel," and "low". For the stop consonant /k/, "stop" and "velar" are recognized. The example shows that one phone can be represented well by one or a few speech attributes despite a language mismatch between training and testing.

2) Glottal Features: Glottal features represent the phonation information of speech. They are computed by first estimating the acoustic excitation of voiced speech, the glottal flow, from speech microphone recordings using glottal inverse filtering (GIF) and then expressing the estimated glottal flow with selected parameters [45]. In the current study, we used the quasi-closed phase (QCP) algorithm [46] as the GIF method. QCP was selected because it was shown in [46] to be more accurate in the estimation of the glottal flow than four state-of-the-art GIF methods. The glottal features used in the current study (listed in Table II) consist of 12 time- and frequency-domain parameters [47], [48]. These parameters were computed in 25-ms frames with a shift of 10 ms using the APARAT toolbox [49]. Two frequency domain parameters (the harmonic richness factor (HRF) and the difference between the first two glottal harmonics (H1H2)) were computed pitch asynchronously once per frame and all the other parameters were computed pitch synchronously once per glottal cycle and then averaged over the frame. H1H2 and HRF were expressed using the dB scale, whereas all the 9 time-domain parameters and the parabolic spectral parameter (PSP) were expressed using a linear scale. The glottal parameters were computed from voiced speech

TABLE II
TIME- AND FREQUENCY-DOMAIN GLOTTAL PARAMETERS USED AS FEATURES IN THE EXPERIMENTS

Time-domain glottal parameters	
OQ1	Open quotient, calculated from the primary glottal opening
OQ2	Open quotient, calculated from the secondary glottal opening
NAQ	Normalized amplitude quotient
AQ	Amplitude quotient
CIQ	Closing quotient
OQa	Open quotient, derived from the LF model
QOQ	Quasi-open quotient
SQ1	Speed quotient, calculated from the primary glottal opening
SQ2	Speed quotient, calculated from the secondary glottal opening
Frequency-domain glottal parameters	
H1H2	Difference between the first two glottal harmonics
PSP	Parabolic spectral parameter
HRF	Harmonic richness factor

For more details, see [47] and [48].

frames, which were detected using a straightforward method based on the frame's log energy. The first- and second-order derivatives of the 12 glottal parameters were calculated using Librosa with 9 contextual frames for each frame, resulting finally in a 36-dimensional vector (from now on referred to as Glottal).

3) Mel-Frequency Cepstral Coefficients: MFCCs were used as standard baseline features capturing the main characteristics of the vocal tract, and hence carrying articulatory information. MFCCs were computed for each 25-ms speech frame using a shift of 10 ms. The first 13 coefficients were extracted and the first- and second-order derivatives were calculated with Librosa with 9 contextual frames. As a result, each speech frame was represented by a 39-dimensional MFCC feature vector.

4) Low-Level Descriptors in eGeMAPS: eGeMAPS is a compact and knowledge-based acoustic feature set designed for automatic voice analysis that has been used in paralinguistic and clinical studies [28]. In general, the eGeMAPS features consist of statistical functionals computed from a set of low-level acoustic descriptors (LLDs) extracted at the frame-level. Here we adopted the most up-to-date ("eGeMAPSv02") 25-dimensional LLD version of eGeMAPS, extracted with the openSMILE toolkit [50], using a frame length of 25 ms with a shift of 10 ms. Among these LLDs, there are prosodic parameters on pitch and intensity and voice quality parameters like jitter, shimmer, and HNR. Specific measurements of formants and certain MFCC coefficients are also included. Together with the first- and second-order dynamic features calculated with Librosa with 9 temporal contextual frames, each speech frame was represented by a 75-dimensional eGeMAPS vector.

C. Neural Networks for Feature Modeling

In order to model the features for the PD assessment, our preliminary experiments investigated a number of neural architectures and SVMs for regression analysis. Based on the average performance across different features, datasets, and conditions, and for the conciseness of the paper, only the results for the best performing CNN architecture are reported.

Input to the CNN consists of features of a 1-s speech segment and the output consists of one scalar number depending on the regression task. The CNN includes three convolutional ReLu

TABLE III
STATISTICS OF PARTICIPANT INFORMATION IN THE FINNISH PDSTU CORPUS (MEAN, (MIN, MAX))

	Control speakers	**PD speakers**
Female	9	21
Male	6	14
Age	57.7 (51, 67)	65.6 (48, 82)
Years after diagnosis	N/A	5.5 (1, 18)
H&Y	N/A	1.8 (1, 2.5)
Speech intelligibility	97.3 (76.0, 114.2)	78.3 (45.0, 101.7)
Voice impairment	12.4 (2.5, 26.1)	28.3 (5.7, 75.2)
Overall severity	9.5 (1.2, 20.8)	23.2 (4.0, 66.9)
Speech task	**Duration** (in seconds)	
VOWEL	9.6 (7.2, 13.6)	12.8 (6.2, 23.7)
READ	38.5 (28.5, 50.0)	38.8 (25.7, 55.1)
SPON	56.7 (12.6, 170.4)	52.7 (21.5, 164.0)

layers with a filter size of $(8, 1)$ and filter count of 128, separated by two temporal maxpooling layers with pooling sizes of (5,1) and (2,1), respectively, followed by one temporal average-pooling layer that integrates information across the entire 1-s segment, ultimately feeding into a dense layer with 32 units that merges the features, followed by a one-unit dense ReLu layer that performs the regression to targets. Unlike typical approach of convolving across the feature dimensions, the CNN only performs convolutions along the temporal dimension and independently for each channel. Kernel weights were shared across the feature channels and the outputs of the filters were not interacting between the channels before the dense layer. In contrast, the combination of feature channels only occurs in the dense layers following the hierarchical feature-specific temporal analysis. We found that focusing on the temporal modeling of the individual features led to the best performance in the analysis tasks, likely due to the substantially reduced parameter count due to the sharing of kernel weights across each channel dimension. The input of CNN was a feature map with a dimension of $100 \times feat\_dim$, where 100 is the frame count in one speech segment and $feat\_dim$ is the frame-level feature dimension described in the previous section.

III. SPEECH CORPORA

A. Pdstu

A subset of a Finnish PD speech corpus, PDSTU [51], was used for the PD assessment experiments in this work. PDSTU speech has been recorded in the mono channel with 32 bits and a sampling rate of 44.1 kHz with a close-talking microphone. This subset contains recordings of speech in terms of vowels-in-words (VOWEL), passage reading (READ), and spontaneous monologue (SPON) from 35 PD speakers as well as 15 HC speakers and their associated expert ratings (see Table III for statistics).

The VOWEL task contained utterances of 5 Finnish words with long vowels, which were *siika*, *laatta*, *kookos*, *tuuri*, and *veeti* for [i:], [a:], [o:], [u:], and [e:], respectively. The speakers were instructed to elongate the long vowels in the words as much as possible. For example, duration of [i:] in *siika* was on average of 1.59 s (SD = 0.86 s; min = 0.48 s, max = 3.58 s) across all the 50 speakers. For the READ task, a passage from "*Pohjantuuli ja aurinko*" ("*North Wind and the Sun*"), which contains 77 Finnish words and has been commonly used in clinical studies in Finland (e.g., in [52]), was used:

- *Read passage:* Pohjantuuli ja aurinko väittelivät, kummalla olisi enemmän voimaa, kun he samalla näkivät kulkijan, jolla oli yllään lämmin takki. Silloin he sopivat, että se on voimakkaampi, joka nopeammin saa kulkijan riisumaan takkinsa. Pohjantuuli alkoi puhaltaa niin, että viuhui, mutta mitä kovempaa se puhalsi, sitä tarkemmin kääri mies takin ympärilleen, ja viimein tuuli luopui koko hommasta. Silloin alkoi aurinko loistaa lämpimästi, eikä aikaakaan, niin kulkija riisui manttelinsa. Niin oli tuulen pakko myöntää, että aurinko oli kuin olikin heistä vahvempi.

The SPON task consisted of a monologue to describe the plot of a one-page cartoon using one's own words (controls) or to describe what the speaker did last summer (PD subjects). Hence, the spoken content in the VOWEL and READ tasks was the same for each speaker, whereas SPON varied between each talker.

The speech of each talker in this PDSTU subset was rated by 3 external logopedics experts recruited through public advertisement. The raters had on average 23 years (at least 16 years) of working experience in speech therapy. The rating was performed on samples of read speech, and the rated dimensions included speech intelligibility, voice impairment, and overall severity of communication disorder. The rating was conducted in a quiet room using high-quality headphones (Sennheiser HD598). The recordings were randomly presented to the raters and participant information was hidden. The average of the 3 experts' ratings of each rated perspective was assigned to the speaker under assessment.

For speech intelligibility, each rater was asked to compare the intelligibility of the given sample with that of a standard sample, where the standard was selected from a HC speaker with a defined intelligibility score of 100. A score larger (smaller) than 100 means it is more (less) intelligible than the standard sample. To reduce the effects of familiarization, 3 short phrases were randomly selected from the reading passage, and the selected phrases were different for each speaker. The rater could only listen to the presented sample once.

A scale from 0 (normal) to 100 (most severe) was used in the ratings of voice impairment and overall severity of communication disorder. For these two measures, the raters could listen to the presented samples as many times as they required.

As a reference of PD severity in this patient group, the Hoehn and Yahr scale (H&Y) [53] was used to measure the severity of PD for each speaker. The rating was determined by the medical doctor responsible for the patient. The H&Y value ranges from 1 to 5 in increments of 0.5. The larger the value, the more severe the movement disorder.

Table III summarizes the statistics of the dataset. On average, PD speakers in PDSTU were at a mild stage of PD with a mean H&Y of 1.8. However, PD-related symptoms of speech and voice disorders can be seen from the expert ratings, where on average the PD patient has a lower score for intelligibility

TABLE IV
STATISTICS OF PD SPEAKER INFORMATION IN THE SPANISH PC-GITA CORPUS (MEAN, (MIN, MAX))

	Male	Female	Total
Number of speakers	25	25	50
Age	61.3 (33, 81)	60.7 (49, 75)	61.0
Years after diagnosis	8.7 (0.4, 20)	13.8 (1, 43)	11.2
H&Y	2.1 (1, 4)	2.2 (1, 3)	2.2
UPDRS	37.8 (6, 93)	37.6 (19, 71)	37.7
Speech tasks	**Duration** (in seconds)		
VOWEL	18.9 (4.6, 81.3)	18.2 (6.1, 49.6)	18.6
READ	19.0 (13.0, 37.3)	18.2 (10.3, 45.3)	18.6
SPON	44.2 (19.7, 111.0)	47.4 (14.1, 104.1)	45.8

and higher scores for voice impairment and overall severity than those of the controls. Average recording duration for each speech task is also shown in the table.

Permission for PDSTU data collection and analysis was obtained from the Ethics Committee of Tampere University. All participants provided written informed consent according to the Declaration of Helsinki.

B. Pc-Gita

The popular Spanish PD speech database PC-GITA [23] was used as the second database in the current study for automatic PD assessment. PC-GITA contains speech recordings from 50 speakers diagnosed with PD and 50 HC speakers matched by age and gender. All the speakers are native Spanish speakers. Each PD speaker is associated with a Unified Parkinson's Disease Rating Scale (UPDRS) rating of PD severity. The UPDRS was originally developed in the 1980 s as a core assessment tool for PD-associated symptoms from diverse perspectives, including mentation, behavior and mood, activities of daily living, motor, and complications [54]. The scale can be used in a clinical setting as well as in research. The higher the UPDRS score, the more severe the PD symptoms, which makes the scale a suitable target for regression analysis. Since only the PD speakers have UPDRS ratings in PC-GITA, the 50 PD speakers' speech from sustained vowels (VOWEL), text reading (READ), and spontaneous monologue (SPON) speech tasks were used in the experiments.

In our experiments, the VOWEL audio sample for each speaker was a concatenation of sustained utterances of five Spanish vowels (/a/, /e/, /i/, /o/, and /u/). The spoken content of the READ task consisted of predefined sentences read by all speakers. In the SPON task, each speaker was asked to talk about what they commonly do in a normal day. The demographic information, expert ratings on UPDRS, and time durations for each speech task are summarized in Table IV.

IV. EXPERIMENTAL SETUP

In this work, we conducted regression tasks on PDSTU and PC-GITA, training and testing our models on each corpus separately. For PDSTU, the prediction targets included expert ratings of speech intelligibility, voice impairment, and overall severity of communication disorder for all the speakers of the database (35 PD patients, 15 controls). For PC-GITA, we estimated the UPDRS for the 50 PD patients of the database. The regression experiments were conducted on each speech task (VOWEL, READ, and SPON) using the different features (SAS, Glottal, MFCC, and eGeMAPS) and for each prediction target separately.

As shown in Fig. 1, the regression was carried out at the level of 1-s segments. The segment counts for the VOWEL, READ, and SPON tasks in PDSTU were 2,734, 9,449, and 13,239, respectively. For the 50 PD participants in PC-GITA, there were a total of 4,405, 4,421, and 11,216 speech segments in the VOWEL, READ, and SPON task, respectively. The CNN model was separately trained with different speech representations (SAS, Glottal, MFCC, eGeMAPS) for each prediction target. Loss function of mean absolute error was used together with Adam optimizer and a learning rate of 0.001. Early stopping based on validation data and patience of 10 epochs was adopted in all experiments. Cross-validation of leave-one-speaker-out was used for model testing. At the test time, the median value of the predicted 1-s segment-level ratings was used as the final subject-level prediction for the given regression task. The Pearson correlation coefficient (PCC) between the predicted and true expert rating for each subject was used to measure regression performance.

A. Feature Fusion

We also investigated feature fusion due to the potential complementarity of the investigated features. Feature fusion was conducted by averaging the model-based segment-level predictions in each regression task. Then the median value of the averaged segment-level predictions was taken as the final score for the speaker under test. We expected that the combination of the articulation-related SAS features and the phonation-related glottal features would improve the performance of parkinsonian speech assessment. We were also specifically interested in the combination of MFCC and eGeMAPS, as they were expected to simultaneously capture the characteristics of the vocal tract and prosody in PD speech. Besides the two combinations mentioned above, we also tested all other possible combinations among the four speech representations.

V. RESULTS

The regression performance scores for speech intelligibility, voice impairment, overall severity of communication disorder (on PDSTU), and UPDRS (on PC-GITA) are shown in Table V. The table shows PCC for the different studied features and for each of the involved speech tasks. In addition, fusion performance of the features motivated by speech production (SAS + Glottal) and fusion performance of the more commonly used eGeMAPS + MFCC are shown. More detailed observations of the results are presented in the following sections.

A. Comparison of Speech Features

To compare the efficacy of different speech features, Table VI shows the performance for each feature type for each regression

TABLE V
REGRESSION RESULTS FOR EACH REGRESSION TARGET, SPEECH TASK, AND FOR INDIVIDUAL FEATURES AND SAS+GLOTTAL AND MFCC+EGEMAPS COMBINATIONS

Corpus	Target	Speech task	SAS	Glottal	SAS + Glottal	MFCC	eGeMAPS	MFCC + eGeMAPS
PDSTU	Speech intelligibility	VOWEL	0.63^{***}	0.66^{***}	0.72^{***} ↑	0.45^{***}	0.59^{***}	0.57^{***}
		READ	0.72^{***}	0.63^{***}	$\mathbf{0.73}^{***}$ ↑	0.59^{***}	0.66^{***}	0.66^{***}
		SPON	0.61^{***}	0.64^{***}	0.71^{***} ↑	0.53^{***}	0.56^{***}	0.57^{***} ↑
	Voice impairment	VOWEL	0.42^{**}	0.41^{**}	0.51^{***} ↑	0.46^{**}	0.63^{***}	0.57^{***}
		READ	$\mathbf{0.64}^{***}$	0.55^{***}	$\mathbf{0.64}^{***}$	0.40^{**}	0.37^{**}	0.43^{**} ↑
		SPON	0.42^{**}	0.45^{***}	0.49^{***} ↑	0.44^{**}	0.57^{***}	0.55^{***}
	Overall severity	VOWEL	0.55^{***}	0.53^{***}	0.62^{***} ↑	0.43^{**}	0.46^{***}	0.50^{***} ↑
		READ	0.67^{***}	0.62^{***}	$\mathbf{0.72}^{***}$ ↑	0.55^{***}	0.39^{**}	0.50^{***}
		SPON	0.69^{***}	0.45^{***}	0.62^{***}	0.50^{***}	0.53^{***}	0.55^{***} ↑
PC-GITA	UPDRS	VOWEL	0.10	−0.27	−0.10	0.10	0.16	0.18 ↑
		READ	0.19	−0.14	0.10	0.33^{*}	0.11	0.27
		SPON	$\mathbf{0.41}^{**}$	−0.04	0.24	0.38^{**}	0.32^{*}	0.37^{**}

Performance is measured using pearson correlation coefficient (PCC) between the subject-level predictions and ground truths. Boldface marks the highest score for each target and ↑ marks the cases where feature fusion improves over the individual features. For all correlations, * stands for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

TABLE VI
REGRESSION PERFORMANCE (PCC) FOR DIFFERENT SPEECH FEATURES AND REGRESSION TARGET COMBINATIONS

Corpus	Target	Best performance				Average performance			
		SAS	Glottal	MFCC	eGeMAPS	SAS	Glottal	MFCC	eGeMAPS
PDSTU	Speech intelligibility	**0.72**	0.66	0.59	0.59	**0.66**	0.64	0.52	0.60
	Voice impairment	**0.64**	0.55	0.46	0.63	0.49	0.47	0.43	**0.52**
	Overall severity	**0.69**	0.62	0.55	0.53	**0.64**	0.54	0.49	0.46
PC-GITA	UPDRS	**0.41**	−0.04	0.38	0.32	0.23	−0.15	**0.27**	0.20

Maximum and average across the different speech tasks is shown. The highest PCC for each prediction target is marked in bold case.

target when the average and max across the different speech tasks is reported. Individual subject-level ground truth and predicted values are exemplified in Fig. 4 of Appendix for speech intelligibility.

As can be observed from the table, SAS consistently outperformed the other three speech representations in PD assessment. Glottal features ranked as the second best in predicting speech intelligibility and overall severity, but failed in the prediction of UPDRS. Post-hoc analysis revealed that the decreased performance of the glottal features in the prediction of UPDRS was due to the recording quality of the speech data used in this task. This task was namely computed using the speech signals of the PC-GITA database, which were recorded as reported in [23] using a dynamic microphone. The non-linear phase response of dynamic microphones is known to be a source of distortion in GIF analysis [45]. The use of microphones with non-linear phase responses in GIF results in particular in the degradation of the time-domain waveform of the estimated glottal flow. Therefore, the use of a low-quality microphone in the speech recordings of the PC-GITA database resulted in the distortion of the time-domain glottal parameters (listed in Table II), which in turn resulted in the poor performance of the glottal features in the UPDRS prediction.

When considering the average feature performance across three different speech tasks in terms of which is a more robust descriptor of the general usefulness of the features independently of the exact speech data, SAS always ranked as the best or second best on the four prediction targets. eGeMAPS and MFCC ranked as the best in predicting voice impairment and UPDRS, respectively. As noted, glottal features suffer from estimation issues in UPDRS, but they ranked as the second or third best for the other three prediction targets, highlighting the relevance of glottal features computed from high-quality speech recordings.

B. Fusion of Different Speech Representations

Table V shows the results of the feature combinations of SAS+Glottal and MFCC+eGeMAPS, while Table VII shows the results for the best possible performance for each regression target and speech task across all possible feature combinations. As can be observed from Table V, the combination of SAS and Glottal achieved a notable improvement in overall severity assessment over the use of SAS only. In addition, performance of the feature combination is comparable to that of SAS only on speech intelligibility and voice impairment. Slight improvements are observed for the four prediction targets when using combinations of MFCC and eGeMAPS with certain speech tasks. However, the combination of MFCC and eGeMAPS never improves over the best performance obtained for a given regression target using only MFCC or eGeMAPS alone, when using the optimal speech task and feature combination for the given regression target.

As for the best possible feature combinations across the regression targets and speech tasks (Table VII), there is a clear trend of SAS being nearly always included in the optimal feature set (in 9 out of 12 cases). For voice impairment prediction, none of the feature combinations outperformed SAS when used in combination with READ speech. Glottal features are included in the best feature sets for intelligibility and overall severity, while MFCCs are the most useful ones in UPDRS assessment. Notably, eGeMAPS is never part of the best feature set for any of the

TABLE VII
BEST-PERFORMING FEATURE COMBINATIONS FOR EACH OF THE REGRESSION TARGETS AND SPEECH TASKS

Corpus	Target	Speech task	Feature type				PCC
			SAS	Glottal	MFCC	eGeMAPS	
PDSTU	Speech intelligibility	VOWEL	1	1	0	1	0.72^{***}
		READ	1	1	1	0	$\mathbf{0.76}^{***}$
		SPON	1	1	0	0	0.71^{***}
	Voice impairment	VOWEL	0	0	0	1	0.63^{***}
		READ	1	0	0	0	$\mathbf{0.64}^{***}$
		SPON	0	0	0	1	0.57^{***}
	Overall severity	VOWEL	1	1	0	0	0.62^{***}
		READ	1	1	0	0	$\mathbf{0.72}^{***}$
		SPON	1	0	0	0	0.69^{***}
PC-GITA	UPDRS	VOWEL	0	0	1	1	0.18
		READ	1	0	1	0	0.34^{*}
		SPON	1	0	1	0	$\mathbf{0.44}^{**}$

1 in each column denotes that the corresponding feature was included in the best feature set. Bolding marks the highest score for each prediction target. For all correlations, stands *** for $p < 0{:}001$, ** for $p < 0{:}01$, and for $^{*}p < 0{:}05$. otherwise, $p > 0{:}05$.

TABLE VIII
CORRELATIONS BETWEEN SUBJECT-LEVEL PREDICTIONS AND GROUND TRUTHS FOR EACH SPEECH TASK AND REGRESSION TARGET WHEN AVERAGED ACROSS THE FEATURE TYPES

Target	Speech task		
	VOWEL	READ	SPON
Speech intelligibility	0.58	**0.65**	0.59
Voice impairment	0.48	**0.49**	0.47
Overall severity	0.49	**0.56**	0.55
UPDRS	0.02	0.12	**0.27**

The highest PCC for each prediction target is marked in bold case. Updrs resultsare for PC-GITA and the other three for PDSTU.

regression targets when the best performing speech task is only taken into account for each target. Scatter plots for predictions and corresponding ground truth ratings in PDSTU are shown in Fig. 5 of the article Appendix.

In general, for all the regression targets, except for voice impairment assessment, a combination of features improves the performance compared to the use of individual feature sets. As supplementary, the results measured in root mean squared error (RMSE) are reported in Table X and Table XI in Appendix. As can be observed from these two tables, the range of system output is slightly compressed compared to the original human ratings, yet internally consistent, as characterized by the reported correlations (i.e., an increase in human ratings leads to an increase in automatic ratings). Therefore, direct comparisons of absolute values from humans and the automatic system should be carried out with caution. However, if human and automated ratings are required to be directly comparable, it is easy to adopt a linear post-correction function to match the range of the automatic ratings to the range of human ratings. By definition, such a scaling would not affect the correlations reported here for our main results.

C. Comparison Between Speech Tasks

Based on the results in Table V, we calculated the average performance across the four features for each target/task combination. These are shown in Table VIII.

As seen from the table, READ is superior to VOWEL and SPON in the speech intelligibility assessment, while SPON is superior to others for UPDRS rating. For voice impairment, all three speech tasks lead to a similar average performance across the feature sets, and READ and SPON achieve a similar performance for automatic assessment of overall severity of communication disorder.

When considering the best feature/speech task/regression target combinations (Table VII) instead of averages across features, the advantage of READ can be observed for intelligibility and overall severity assessment. SPON is again the best for UPDRS, and VOWEL and READ are comparable for voice impairment.

However, caution should be used in interpreting small differences between the performance figures for different speech tasks. In fact, the differences in correlations obtained for the optimal feature sets (Table VII) are not statistically different from each other ($p > 0.05$; William's test for comparing dependent correlations [55]). In fact, a substantially larger number of speakers would be needed to extract fine-grained differences in the performance measures—a requirement that very few existing pathological speech corpora satisfy at the time of writing.

As the final step of our speech task analysis, we wanted to investigate the complementarity of the different speech tasks. This was obtained by weighted summation of the subject-level predictions obtained by each speech task as a means to create the final predictions for the individual speakers. For each task and regression target, the best feature set for the given combination was used in the fusion. The weight for each speech task was searched from 0.0 to 1.0 with a step of 0.1. As our final test, we tested the fusion of all speech representations and all speech tasks.

The weights for each task/target combination and the corresponding PCC scores are shown in Table IX. As a result of optimal task fusion, the PCC was further improved to 0.78 in the prediction of speech intelligibility, to 0.74 for voice impairment, and to 0.76 for overall severity. No improvement for UPDRS was observed. As these are better than or equal to any of the results reported in Tables V– VIII, this further demonstrates the benefit of multiple alternative ways to 1) collect speech data from PD patients, and 2) analyze it in terms of complementary features.

TABLE IX
RESULTS FOR SPEECH TASK FUSION WHEN OPTIMAL WEIGHTS ARE USED FOR TASK-SPECIFIC REGRESSION SCORES FROM EACH TASK AND FOR EACH REGRESSION TARGET

Target	Weights			PCC
	VOWEL	READ	SPON	
Speech intelligibility	0.6	1.0	0.4	0.78^{***}
Voice impairment	0.7	1.0	0.1	0.74^{***}
Overall severity	0.1	0.7	0.7	0.76^{***}
UPDRS	0.0	0.3	1.0	0.44^{**}

For all correlations, ** stands for $p < 0.01$ and *** for $p < 0.001$. UPDRS results are for PC-GITA and the other three for PDSTU.

D. Comparison Between Regression Targets

Besides comparing the speech tasks and technical solutions to assess speech from these tasks, one central question is the accuracy at which different aspects of pathological speech can be estimated automatically. By looking at Table VII, one can observe that speech intelligibility can be most accurately estimated from the speech data (PCC = 0.76 with the best feature and speech task combination), followed by overall communication disorder severity (PCC = 0.72), voice impairment (PCC = 0.64), and UPDRS (PCC = 0.44) in the descending order. If optimal weighing across speech tasks is used (Table IX), intelligibility, impairment, and overall severity reach a similar overall accuracy, while UPDRS is again substantially lower.

The above findings are not surprising: speech intelligibility is largely determined by the most prominent properties of speech, namely suitable temporal modulations of the spectral envelope and adequate prosody (rhythm, intonation, stress) for the given linguistic content—all properties that are strongly reflected in many of the studied features. The other end of the performance spectrum, UPDRS, is a general measure of PD severity that involves a variety of non-motor and motor factors, and where potential difficulties in speech production represents only one of the many factors contributing to the overall score. Therefore it was predictable that an automatic speech-based assessment of a general health measure such as UPDRS is much more difficult than predicting expert ratings directly related to characteristics of speech.

As for the exact UPDRS prediction scores on the PC-GITA data, the best PCC we obtained in this work was 0.44, which is lower than the PCC of 0.79 obtained in [12]. However, it should be noted that a different cross-validation method (K-Fold, $K = 10$) was used in [12], and the model hyperparameters used in [12] were optimized on the held-out set each time, which could lead to optimistic results.

This also raises another topic of generalizing the PD assessment model to different speech corpora, which is beyond the aims of this work.

E. The Effect of the Amount of Training and Testing Data on Speech Tasks

Compared with VOWEL, READ contains richer dynamic information on prosody, articulation of various phonemes, and transitions between different linguistic units, which is unarguably more representative of the daily use of speech communication [56]. However, it should be noted that the data amount for

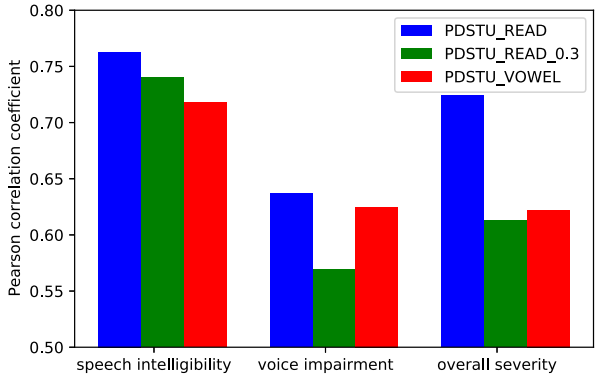

Fig. 3. Results for the READ versus VOWEL tasks when matching the duration of read speech with that of vowels (denoted by PDSTU_READ_0.3). Only the best results across the alternative feature combinations are shown.

READ in PDSTU was substantially larger than that for VOWEL. Given that the same machine learning model architecture was used across all the tasks, it is also possible that READ and SPON resulted in more accurate regression due to their larger number of training samples to optimize the parameters and the number of testing segments to derive the subject-level predictions, compared to the VOWEL data. To check how much the data amount affects the prediction performance, we ran the assessment experiments on PDSTU using only 30% of READ speech data from each speaker, making it comparable to VOWEL in size (2,800 READ segments versus 2,734 VOWEL segments in PDSTU).

Fig. 3 shows the PCC for each prediction target on PDSTU when using VOWEL, READ, and 30% of READ speech, when only the best feature configuration for the given setup is shown. The figure shows that the performance decreased when only 30% of READ speech was used. When READ speech had the same amount of data as VOWEL, the performance of speech intelligibility prediction was slightly better than that of VOWEL, while the performance on voice impairment prediction was worse than that of VOWEL. On overall severity prediction, the performance when using VOWEL and 30% READ was similar. However, these correlation differences obtained by VOWEL and 30% of READ are not significant ($p > 0.05$ for William's test).

While Fig. 3 shows that the reduction of read speech duration to be comparable with carefully collected sustained vowels leads to a similar performance, we cannot conclude that sustained vowels are similar in informativeness to that of read speech. While larger amounts of read speech can be (and already is) collected in a clinical setting and can be utilized in machine learning, it is not so clear whether collection of larger amounts of sustained vowel data would add to a comparable boost in performance. In addition, we showed that the combination of several speech tasks improves performance beyond the individual tasks, indicating complementarity in the data and models trained for the speech tasks. However, the result does show that continuous read speech with standardized contents, when available in comparable amounts to that of sustained vowels, is not necessarily superior in automatic speech pathology assessment compared to the vowel data. These observations complement the previous discussions on appropriate speech tasks used for pathological speech analysis [21], [22], [24], [57].

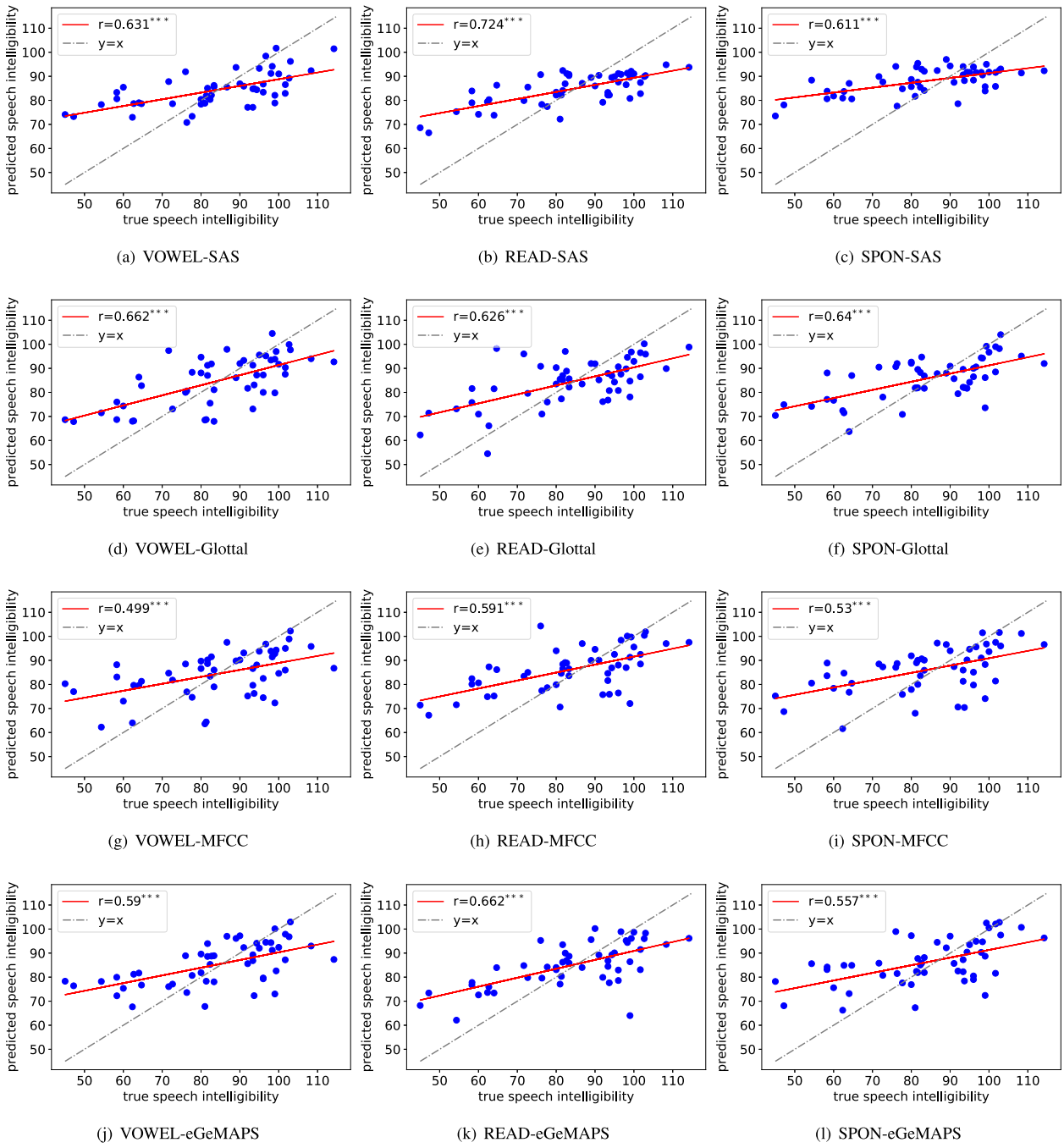

Fig. 4. Subject-level ground truth and leave-one-speaker-out predictions of speech intelligibility using different feature types (SAS, Glottal, MFCC, eGeMAPS) on each speech task (VOWEL, READ, SPON) on the PDSTU corpus. Each dot corresponds to one speaker. In the legend, 'r' refers to Pearson correlation and *** stands for $p < 0.001$.

VI. Conclusion

This work studied the use of phonation and articulation features in the automatic assessment of PD speech, also comparing and combining these features with commonly used MFCCs and eGeMAPS features. We also systematically explored how the type of speech material affects the assessment outcomes. Notably, all the experiments were conducted in the context of regression tasks using speech from two very distinct languages (Finnish and Spanish). Even though regression is not often utilized in automatic pathological speech assessment, we found its use central to our research purposes. This is because our primary long-term interest is in the automatic evaluation and monitoring of disease progression, including following the effects of treatment and therapeutic interventions, making gradual subject-level changes important to track. In contrast, hard classification of subjects into "health categories," such as healthy, mild, or severe, produces information that is difficult to utilize in practical healthcare settings, at least as long as we

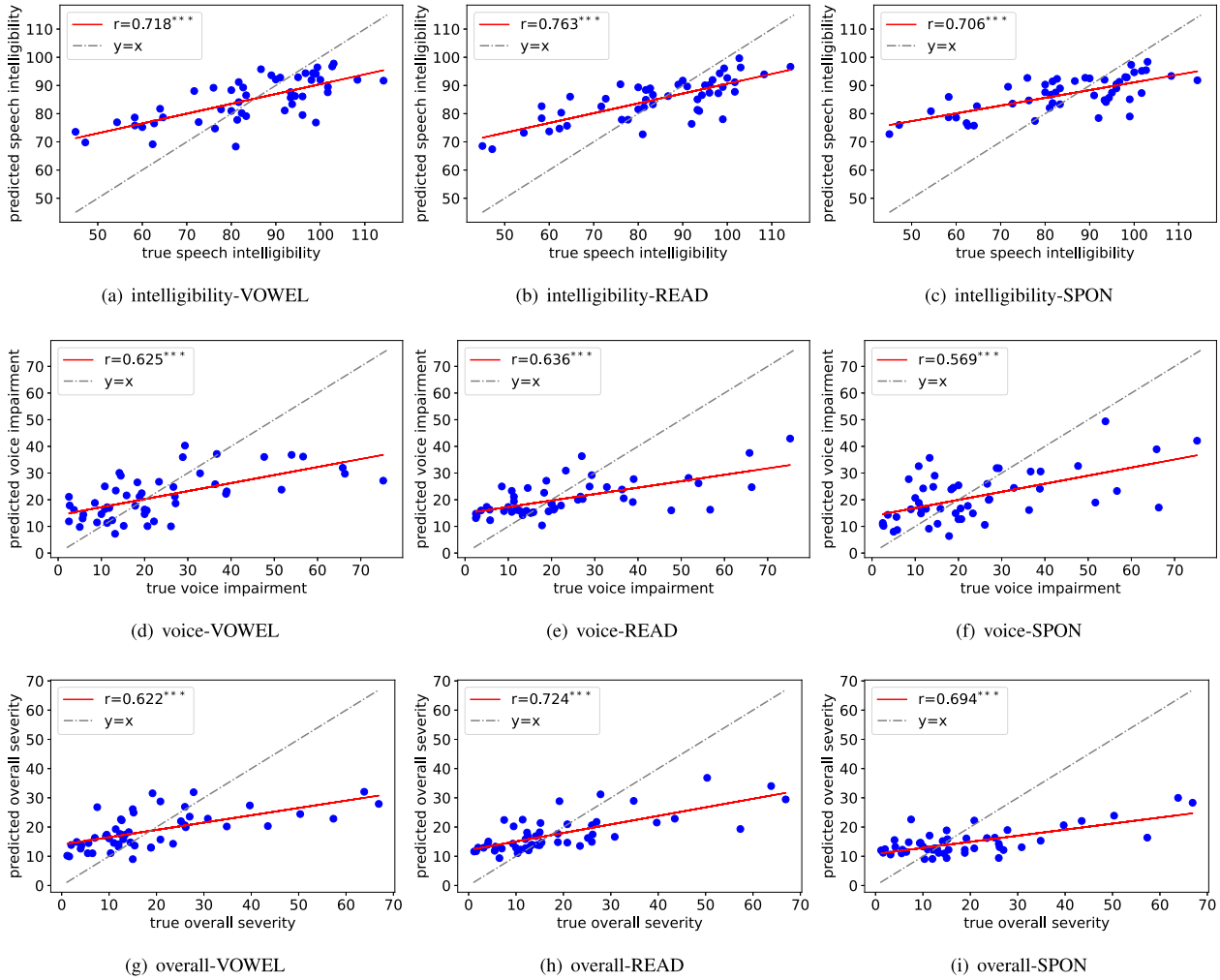

Fig. 5. Scatter plots for predictions vs. ground truth of speech intelligibility (intelligibility), voice impairment (voice), and overall severity (overall) of the PDSTU data from the VOWEL, READ, and SPON tasks while using the best feature combination for each (as shown in Table VII). Each dot corresponds to one speaker. In the legend, 'r' refers to Pearson correlation and *** stands for $p < 0.001$.

lack extremely accurate classifiers and standardized (clinical) criteria for the categorical groupings.

Our experimental results demonstrate that 1) the speech attribute scores can adequately capture the abnormalities in parkinsonian speech, 2) and combining SAS with different speech representations is likely to improve the PD assessment performance. To be more specific, the combination of articulation-related speech attribute scores and phonation-related glottal features achieved a better performance than that of using either feature set alone in several of our test scenarios, which has not been the case in previous work [16]. However, the benefit of glottal features was also highly dependent on the quality of the available audio recordings. In the case of PC-GITA, the use of glottal features was found to be problematic due to the following reasons: 1) the use of dynamic microphones to collect the data (resulting in a non-linear phase response), 2) uncertainties associated with signal polarity (that was potentially changing between different recordings), and 3) mild but noticeable reverberation in the audio. As a result, the obtained glottal waveforms and the consequent glottal features were of poor quality compared those extracted from the PDSTU data.

The experimental results on different speech tasks show that, in general, text reading speech achieved better performance than sustained vowels. Our results also showed that a combination of model-based predictions from all three speech tasks can improve the prediction performance even further. Based on all the discussions, we recommend the employment of audio recordings from diverse speech tasks for parkinsonian speech analysis.

In general, the efficacy of speech attribute scores and phonation features used for PD assessment was demonstrated in this work. One of our next aims is to study the visualization and sensitivity analysis of the features in the context of pathological speech assessment. The aim would be to connect the mechanisms of speech production with the obtained summary scores on intelligibility and pathology. Given that the articulatory and phonation features have a relatively straightforward interpretation in terms of speech production, if successful, such an approach would help to identify what type of targeted changes (interventions) in speech production could lead to the largest improvements in the quality of produced speech for a given talker. Besides the phonation, articulation, and prosody studied

TABLE X
REGRESSION RESULTS FOR EACH REGRESSION TARGET, SPEECH TASK, AND FOR INDIVIDUAL FEATURES AND SAS+GLOTTAL AND MFCC+EGEMAPS COMBINATIONS

Corpus	Target	Speech task	SAS	Glottal	SAS + Glottal	MFCC	eGeMAPS	MFCC + eGeMAPS
PDSTU	Speech intelligibility	VOWEL	13.0	12.2	**11.9** ↓	14.1	13.2	13.4
		READ	12.3	12.7	12.0 ↓	13.3	12.3	12.5
		SPON	14.2	12.7	12.9	13.9	13.7	13.5 ↓
	Voice impairment	VOWEL	16.1	16.3	15.3 ↓	15.8	**14.2**	14.7
		READ	14.6	14.8	**14.2** ↓	16.4	16.9	16.1 ↓
		SPON	17.0	16.2	16.0 ↓	16.2	14.7	15.0
	Overall severity	VOWEL	13.3	13.1	12.5 ↓	14.0	13.7	13.4 ↓
		READ	12.8	12.2	**11.8** ↓	13.3	14.6	13.6
		SPON	13.3	14.2	13.2 ↓	13.8	13.4	13.1 ↓
PC-GITA	UPDRS	VOWEL	18.3	20.2	18.8	19.5	18.7	18.1 ↓
		READ	18.1	19.9	18.3	17.6	18.9	17.6
		SPON	**16.6**	19.1	17.6	16.9	17.3	16.9

Performance is measured using root mean squared error (RMSE) between the subject-level predictions and ground truths. Boldface marks the lowest RMSE for each target and ↓ marks the cases where feature fusion improves over the individual features.

TABLE XI
BEST-PERFORMING FEATURE COMBINATIONS FOR EACH OF THE REGRESSION TARGETS AND SPEECH TASKS, MEASURED IN ROOT MEAN SQUARED ERROR (RMSE)

Corpus	Target	Speech task	Feature type				RMSE
			SAS	Glottal	MFCC	eGeMAPS	
PDSTU	Speech intelligibility	VOWEL	1	1	0	0	11.9
		READ	1	1	1	0	**11.7**
		SPON	0	1	0	0	12.7
	Voice impairment	VOWEL	0	0	0	1	**14.2**
		READ	1	1	0	0	**14.2**
		SPON	0	0	0	1	14.7
	Overall severity	VOWEL	1	1	0	1	12.6
		READ	1	1	0	0	**11.8**
		SPON	1	0	1	0	12.9
PC-GITA	UPDRS	VOWEL	1	0	0	1	18.1
		READ	1	0	1	0	17.1
		SPON	1	0	1	0	**16.4**

1 in each column denotes that the corresponding feature was included in the best feature set. Bolding marks the highest score for each prediction target.

in this work, we plan to explore representations of the cognitive-linguistic aspect of parkinsonian speech with the employment of text features computed from transcripts of monologue speech. Finally, we would like to point out that the PDSTU data used in the current study is relatively limited in size and not completely balanced between the PD and HC groups in terms of the number of speakers, their ages, or in terms of the monologue task instructions. Although the effect of potential group biases is not critical in present type of regression tasks where within-group variance also needs to be captured by the features and models, it would be useful to further verify the studied methods on larger pathological speech databases with balanced enrollment of participants.

APPENDIX

Scatter plots in Fig. 4 illustrate the relationship between the PDSTU subject-level expert ratings (x-axes) and the automatically predicted ratings (y-axes) in case of different speech tasks and speech features. Fig. 5 shows the corresponding plots as a function of PDSTU regression target and speech task when using the best feature combination for each (as listed in Table VII of the manuscript).

Table X and Table XI are parallel to Table V and Table VII respectively, but only measured in root mean squared errors (RMSE). The advantages of SAS features and feature fusion can be observed from RMSE results as well.

ACKNOWLEDGMENT

The authors would like to thank Rachel Convey for collecting the expert ratings for the PDSTU corpus.

REFERENCES

[1] M. Rijk et al., "Prevalence of parkinson's disease in Europe: A collaborative study of population-based cohorts. neurologic diseases in the elderly research group," *Neurol.*, vol. 54, pp. S21–3, 2000.
[2] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, "Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects," *Biomed. Signal Process. Control*, vol. 66, 2021, Art. no. 102418.
[3] Joseph R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis and Management*, 4th ed. Elsevier Health Sciences, 2019.
[4] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with parkinson's disease," *Behav. Neurol.*, vol. 11, no. 3, pp. 131–137, 1998.
[5] L. O. Ramig, C. Fox, and S. Sapir, "Speech treatment for parkinson's disease," *Expert Rev. Neurotherapeutics*, vol. 8, no. 2, pp. 297–309, Jan. 2014.
[6] L. O. Ramig, C. Fox, and S. Sapir, "Speech treatment for parkinson's disease," *Expert Rev. Neurotherapeutics*, vol. 8, no. 2, pp. 297–309, Feb. 2008.
[7] J. Rusz et al., "Evaluation of speech impairment in early stages of parkinson's disease: A prospective study with the role of pharmacotherapy," *J. Neural Transmiss.*, vol. 120, no. 2, pp. 319–329, Feb. 2013.

[8] G. An, D. G. Brizan, M. Ma, M. Morales, A. R. Syed, and A. Rosenberg, "Automatic recognition of unified parkinson's disease rating from speech with acoustic, i-vector and phonotactic features," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 508–512.

[9] B. Schuller et al., "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 478–482.

[10] D. Sztahó, G. Kiss, and K. Vicsi, "Estimating the severity of parkinson's disease from speech using linear regression and database partitioning," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 498–502.

[11] S. Hahm and J. Wang, "Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 513–517.

[12] J. R. Orozco-Arroyave, Analysis of speech of people with parkinson's disease (doctoral thesis), *Logos Verlag Berlin GmbH*, 2016, vol. 41.

[13] A. Zlotnik, J. M. Montero, R. San-Segundo, and A. Gallardo-Antolín, "Random forest-based prediction of parkinson's disease progression using acoustic, ASR and intelligibility features," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 503–507.

[14] J. R. Williamson et al., "Segment-dependent dynamics in predicting parkinson's disease," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 518–522.

[15] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vásquez-Correa, and E. Nöth, "Characterisation of voice quality of parkinson's disease using differential phonological posterior features," *Comput. Speech Lang.*, vol. 46, pp. 196–208, 2017.

[16] T. Arias-Vergara, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Parkinson's disease and aging: Analysis of their effect in phonation and articulation of speech," *Cogn. Comput.*, vol. 9, pp. 731–748, Aug. 2017.

[17] S. Sapir, C. Fox, J. Spielman, and L. Ramig, "Acoustic metrics of vowel articulation in parkinson's disease: Vowel space area (VSA) vs. vowel articulation index (VAI)," *MAVEBA - 7th Int. Workshop*, pp. 173–175, 2011.

[18] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," *J. Speech Lang. Hear Res.*, vol. 53, no. 1, pp. 114–125, Feb. 2010.

[19] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *J. Speech Hear. Disord.*, vol. 43, no. 1, pp. 47–57, 1978.

[20] M. Novotnỳ, J. Rusz, R. Čmejla, and E. Růžička, "Automatic evaluation of articulatory disorders in parkinsons disease," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014.

[21] J. Jiménez-Monsalve, J. C. Vsquez-Correa, J. R. Orozco-Arroyave, and P. Gomez-Vilda, "Phonation and articulation analyses in laryngeal pathologies, cleft lip and palate, and parkinsons disease," in *Proc. Int. Work- Conf. Interplay Between Natural Artif. Comput*. Springer, 2017, pp. 424–434.

[22] Y. Liu, T. Lee, T. Law, K. Lee, and P. C. Ching, "Prediction of voice disorder severity: Contributions from sustained vowels and continuous speech," in *Proc. 11th Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 290–294.

[23] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, 2014, pp. 342–347.

[24] J. C. Vásquez-Correa, J. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease," *J. Commun. Disord.*, vol. 76, pp. 21–36, 2018.

[25] J. Vasquez et al., "Convolutional neural networks and a transfer learning strategy to classify parkinson's disease from speech in three different languages," in *Proc. Iberoamerican Congr. Pattern Recognit.*, 2019, pp. 697–706.

[26] J. Vásquez-Correa, J. Fritsch, J. Orozco-Arroyave, E. Nöth, and M. Magimai-Doss, "On modeling glottal source information for phonation assessment in parkinson's disease," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 26–30.

[27] C. O. Sakar et al., "A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform," *Appl. Soft Comput.*, vol. 74, pp. 255–263, 2019.

[28] F. Eyben et al., "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

[29] Z. Syed, S. Ali, and A. Latif, "Deep acoustic embeddings for identifying parkinsonian speech," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, pp. 726–734, Nov. 2020.

[30] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *Proc. Int. Conf. Spoken Lang. Process.*, 2002, pp. 2133–2136.

[31] I. Kukanov, T. N. Trong, V. Hautamäki, S. M. Siniscalchi, V. M. Salerno, and K. A. Lee, "Maximal figure-of-merit framework to detect multi-label phonetic features for spoken language recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 8, pp. 682–695, 2020.

[32] C. Middag, J.-P. Martens, G. V. Nuffelen, and M. D. Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP J. Adv. Signal Process.*, vol. 2009, pp. 1–9, 2009.

[33] L. Moro-Velazquez et al., "Phonetic relevance and phonemic grouping of speech in the automatic detection of parkinson's disease," *Sci. Rep.*, vol. 9, no. 1, pp. 1–16, Dec. 2019.

[34] I. Midi, M. Dogan, M. Koseoglu, G. Can, M. Sehitoglu, and D. Gunal, "Voice abnormalities and their relation with motor dysfunction in parkinson's disease," *Acta Neurologica Scandinavica*, vol. 117, no. 1, pp. 26–34, 2008.

[35] M. Novotný, P. Dušek, I. Daly, E. Růžička, and J. Rusz, "Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with parkinson's disease: Correlation between acoustic speech characteristics and non-speech motor performance," *Biomed. Signal Process. Control*, vol. 57, 2020, Art. no. 101818, doi: 10.1016/j.bspc.2019.101818.

[36] N. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3403–3407.

[37] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 96–107, Jan. 2008.

[38] M. K. Reddy, P. Alku, and K. S. Rao, "Detection of specific language impairment in children using glottal source features," *IEEE Access*, vol. 8, pp. 15273–15279, 2020.

[39] M. K. Reddy et al., "The automatic detection of heart failure using speech signals," *Comput. Speech Lang.*, vol. 69, 2021, Art. no. 101205.

[40] I. Kukanov, V. Hautamäki, S. M. Siniscalchi, and K. Li, "Deep learning with maximal figure-of-merit cost to advance multi-label speech attribute detection," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2016, pp. 489–495.

[41] I. Kukanov, T. N. Trong, V. Hautamäki, S. M. Siniscalchi, V. M. Salerno, and K. A. Lee, "Maximal figure-of-merit framework to detect multi-label phonetic features for spoken language recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 682–695, 2020.

[42] D. Povey et al., "The kaldi speech recognition toolkit," in *ASRU*, Dec. 2011.

[43] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proc. Int. Conf. Spoken Lang. Process.*, 1992, pp. 895–898.

[44] B. McFee et al., "Librosa: Audio and music signal analysis in python," in *Proc. 14th Python Sci. Conf.*, vol. 8, 2015, pp. 18–24.

[45] P. Alku, "Glottal inverse filtering analysis of human voice production - a review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 36, no. 5, pp. 623–650, Oct. 2011.

[46] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.

[47] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, Jul. 1991.

[48] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112, no. 2, pp. 701–710, Aug. 2002.

[49] M. Airas, H. Pulakka, T. Bäckström, and P. Alku, "A Toolkit for voice inverse filtering and parametrisation," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 2145–2148.

[50] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 835–838.

[51] Y. Liu, N. Penttilä, T. Ihalainen, J. Lintula, R. Convey, and O. Räsänen, "Language-independent approach for automatic computation of vowel articulation features in dysarthric speech assessment," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2228–2243, 2021.

[52] E. Kankare et al., "The acoustic voice quality index version 02.02 in the finnish-speaking population," *Logopedics Phoniatrics Vocol.*, vol. 45, no. 2, pp. 49–56, Apr. 2020.
[53] C. G. Goetz et al., "Movement disorder society task force report on the hoehn and yahr staging scale: Status and recommendations," *Movement Disord.*, vol. 19, no. 9, pp. 1020–1028, Sep. 2004.
[54] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, "The unified parkinson's disease rating scale (UPDRS): Status and recommendations," *Movement Disorders, vol.* 18, no. 7, pp. 738–750, 2003.
[55] E. J. Williams, "The comparison of regression variables," *J. Roy. Statist. Soc.: Ser. B. (Methodological)*, vol. 21, no. 2, pp. 396–399, 1959.
[56] Y. Liu, T. Lee, T. Law, and K. Y.-S. Lee, "Acoustical assessment of voice disorder with continuous speech using ASR posterior features," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1047–1059, Jun. 2019.
[57] Y. Maryn and N. Roy, "Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity," *J. da Sociedade Brasileira de Fonoaudiologia*, vol. 24, pp. 107–12, 2012.

Yuanyuan Liu was born in China in 1988. She received the B.S. degree in techniques and science of electronics in 2010 from the Central South University, Changsha, China, and the Ph.D. degree in electronic engineering from the Chinese University of Hong Kong, Hong Kong, in 2019. From 2010 to 2014, she was a Full-time Senior Engineer with the Department of Wireless Products in MediaTek Inc. Shenzhen. From 2014 to 2018, she was a Part-time Research Assistant with the Shenzhen Municipal Engineering Laboratory of Speech Rehabilitation Technology, China. From March 2020 to February 2022, she was a Postdoctoral Researcher with the Unit of Computing Sciences, Tampere University, Tampere, Finland. She is currently working in industry as an Audio and DSP Engineer. Her research interests include automatic assessment of neurodegenerative diseases, automatic speech recognition, user identification, and machine learning.

Mittapalle Kiran Reddy received the M.E. degree in communication systems from the SSN College of Engineering, Chennai, India, in 2014, and the Ph.D. degree in speech processing from the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India, in 2019. From October 2014 to March 2018, he was a Senior Scientific Officer with the research project sponsored by the Department of Information Technology, Govt. of India, undertaken by IIT Kharagpur. He is currently a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests include signal processing, speech synthesis, speech recognition, analysis and detection of speech disorders, and machine learning.

Nelly Penttilä was born in Finland in 1989. She received the M.Sc. and Ph.D. degrees in logopedics from Tampere University, Tampere, Finland, in 2012 and 2019, respectively. She is currently a Senior Lecturer with the discipline of logopedics and as a Principal investigator in Kuuluva Ääni -project (carrying voice-project). Her research interests include fluency, fluency disorders, and speech intelligibility. She also works as speech and language pathologist. She was the recipient of the multiple awards from different foundations, e.g. Fulbright Finland Foundation, and Emil Aaltonen Foundation.

Tiina Ihalainen was born in Finland in 1974. She received the M.Sc. and Ph.D. degree in logopedics from Oulu University, Oulu, Finland, in 2008 and 2018, respectively. From 2008 to 2019, she was a speech therapist with the Tampere University Hospital, Finland. She is specially experienced in working with adult people who have neurological diseases affecting in speech and communication (e.g. stroke, brain injury, brain tumor, Parkinson's disease, motor neuron disease, different types of dementia and other neurogenerative diseases.). She is currently a Senior Lecturer with Degree Program in Logopedics with the Faculty of Social Sciences, Tampere University. Her research interests include perceptual and automatic assessment of typical and pathological speech and creating novel user interfaces for augmentative and alternative communication.

Paavo Alku received the M.Sc., Lic.Tech., and Dr.Sc.(Tech) degrees from the Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an Assistant Professor with the Asian Institute of Technology, Khlong Nueng, Thailand, in 1993, and an Assistant Professor and Professor with the University of Turku, Finland, from 1994 to 1999. He is currently a Professor of speech communication technology with Aalto University, Espoo, Finland. His research interests include analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modelling of speech, speech-based biomarking of human health, and cerebral processing of speech. He has authored or coauthored around 230 peer-reviewed journal articles and around 220 peer-reviewed conference papers. He is an Associate Editor for the *Journal of the Acoustical Society of America*. He was an Academy Professor assigned by the Academy of Finland in 2015 to 2019. He is a Fellow of ISCA.

Okko Räsänen was born in Finland in 1984. He received the M.Sc. (Tech.) degree in language technology from the Helsinki University of Technology, Espoo, Finland, in 2007, and the D.Sc. (Tech.) degree in language technology from Aalto University, Espoo, Finland, in 2013. He also holds the Title of Docent (Adjunct Professor) from Aalto University in the area of Spoken Language Processing. He is currently an Associate Professor with the Unit of Computing Sciences, Tampere University, Tampere, Finland, and a Visiting Researcher with Aalto University. In 2015, he was a Visiting Researcher with the Language and Cognition Lab of Stanford University, Stanford, CA, USA. His research interests include computational modeling of language acquisition, cognitive aspects of language processing, and speech analysis and processing in general.