Acoustic tracking of pitch, modal and subharmonic vibrations of vocal folds in Parkinson’s disease and Parkinsonism

Jan Hlavnička1,2, Roman Čmejla1, Jiří Klempíř2, Evžen Růžička2, and Jan Rusz1,2
1Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
2Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine, Charles University, Prague, Czech Republic

Corresponding author: Jan Hlavnička (e-mail: hlavniczka@gmail.com).

This study was supported by the by the Czech Ministry of Health grant no. NV19-04-00120, the OP VVV MEYS grant no. CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics", and the Czech Ministry of Education grant no. PROGRES-Q27/LF1.

ABSTRACT The prominent and early presence of dysphonia is considered a valuable marker for differentiation of idiopathic Parkinson’s disease and parkinsonian syndromes. Objective quantification of vibrational regimes represented by modal and subharmonic vibrations may thus be vital for improving accuracy of diagnostic decision. The rationale for analyzing vibrational regimes is that abnormal subharmonic vibrations might be the key factor causing dysphonia in parkinsonian syndromes. This study introduces a new fully automated methodology based on robust pitch tracker for decoupling vibrations controlled by laryngeal muscles from the effect of subharmonics that provides distinguishing features of Parkinson’s disease and atypical parkinsonian syndromes. We tested the method on resynthesized signals with known parameters and demonstrated that vibrations controlled by laryngeal muscles as well as subharmonics can be detected reliably with a precision that outperforms available technologies. We analysed 337 sustained vowels of 22 patients with PD, 21 patients with multiple system atrophy, 18 patients with progressive supranuclear palsy and 22 healthy controls. Our results showed that subharmonics are more prominent in atypical parkinsonian syndromes compared to Parkinson’s disease. Also, increased modulation by laryngeal muscles appears to be a distinctive symptom of multiple system atrophy. Developed algorithm and proposed resynthesized voice signals provide further critical step to understanding and evaluation of dysphonia in Parkinsonism.


I. INTRODUCTION

A. Parkinson’s Disease and Atypical Parkinsonian Syndromes

Idiopathic Parkinson’s disease (PD) represents the most common neurodegenerative disorder affecting the basal ganglia. PD is characterized by the progressive loss of dopaminergic neurons, primarily in the substantia nigra pars compacta [1]. The cardinal motor manifestations of PD are tremor at rest, rigidity, bradykinesia and postural instability. In addition, PD patients manifest poor control and coordination of the speech motor system. Hypokinetic dysarthria is the specific speech disorder affecting PD and can be characterized by the presence of monopitch, monoloudness, imprecise articulation of consonants and vowels, variable speaking rate with rushes of fast speech, reduced vocal loudness, and harsh or breathy vocal quality [2].

Atypical parkinsonian syndromes (APS) such as multiple system atrophy (MSA) and progressive supranuclear palsy (PSP) differ from PD by more widespread neurodegeneration, rapid progression and poor response to dopaminergic medication. MSA manifests with akinetic-rigid syndrome, cerebellar ataxia and autonomic dysfunction. PSP patients suffer from axial rigidity, postural instability, vertigo, gait difficulty, changes in personality and dementia. Due to the overlap of clinical features, PD, MSA and PSP are frequently misdiagnosed until more advanced disease stages [3]. Dysarthria in APS is commonly more severe than in PD. PSP typically manifests mixed
hypokinetic-spastic dysarthria characterized by stuttering-like behaviour, reduced speech rate, decreased intonation variability, prolonged pauses, articulation imprecision and poor quality of voice, whereas MSA patients show mixed ataxic-hypokinetic dysarthria characterized by excess pitch fluctuations, excess intensity variations, increased vocal pitch, reduced speech rate, prolonged phonemes, vocal tremor, voice perturbations and slow variable alternating motion rates [4].

B. DETECTION OF THE FUNDAMENTAL FREQUENCY IN DYSARTHRIA

Any or all of the basic motor processes of speech including prosody, articulation, phonation, resonance or respiration can be disrupted in dysarthria [2]. Nevertheless, phonatory dysfunction represents one of the most frequent manifestations across all types of dysarthria [5-7] and has been reported as one of the earliest speech manifestations in neurodegeneration [8]. Although phonation is traditionally evaluated by auditory-perception, clinicians have routinely begun to use additional acoustic analysis to objectify the diagnosis as computational power and systems become more available. Phonation is commonly analysed by a sustained vowel paradigm with well-established fundamental frequency (F₀) characteristics such as mean and variation of F₀, and perturbation analysis represented by jitter, shimmer and harmonic-to-noise ratio. However, these measures substantially rely on the precise detection of F₀ [9].

Detection of F₀ is widely applied in speech transmission and synthesis. It is estimated that more than 3000 bibliographic entries concern about detection of F₀ [10]. Generally, no single method operates perfectly in any situation, and most methods represent a trade-off for intelligible speech transmission. Conventional detectors show respectable robustness to environmental noise. However, heavy perturbations in pathological speech are introduced by the speaker and are the subject of great interest. Unlike speech transmission, where noise suppression is desired and the output is forced to be regular, medical applications must prevent these corrections to provide reliable insight into the pathophysiology of speech production.

Dysarthria introduces difficulties in the detection of F₀ due to elevated levels of glottal noise and cycle-to-cycle aperiodicity, abnormal resonance, extensive pitch variations and pitch breaks. Phonation is pre-eminently examined on sustained vowels to ensure stable airflow through the vocal folds. Majority of available speech analysers are based on autocorrelation or cross-correlation functions. Although these methods can provide very precise results, they suffer from false octave jumps. Manual correction of pitch trace by limiting the range of analysis is a standard feature of many speech analysers and is highly recommended by manufacturers. Evaluation of available software [11,12], standard algorithms [13,14] or their fusion [15] has shown that the results of automated detection should be carefully interpreted and specific detectors are necessary for application to pathological speech. Moreover, to the best of our knowledge, the robustness of F₀ detection has only been tested in PD [15], which does not typically involve ataxic pitch fluctuations or spastic pitch breaks. In this regard, MSA and PSP represent the best model for testing the robustness of F₀ detection, as the severity of dysarthria is typically greater than in PD and all types of voice abnormalities are present.

C. SUBHARMONIC VIBRATIONS

Subharmonic vibrations of the vocal folds represent a specific oscillatory pattern due to period and/or amplitude alternation of the glottal cycle. Alternating cycles enrich the spectrum with additional subharmonics at an integer fraction of F₀, typically F₀/2. The pathophysiological mechanism of subharmonic vibrations is not well known. Švec [16] suggested that subharmonics result from a combination of two distinct vibrations with a frequency ratio of 3:2. Titze [17] hypothesized that left-right asymmetry of vocal fold parameters may be the cause. Depending on the depth of alternation, the voice can be perceived either as rough or sounding one octave lower [18,19]. The ambiguous nature of perception introduces discrepancies into current clinical terminology that categorizes subharmonics either as diplophonia, due to the presence of concurrent subharmonic periodicities, harsh voice, a rough quality of voice caused by subharmonics, or sudden shifts in pitch called pitch break [20].

Although healthy vocal folds can be forced to vibrate with subharmonics by certain vocal manoeuvres [16] and trained individuals such as singers are capable of producing subharmonics consciously [22], extensive spontaneous subharmonic vibrations are typically associated with phonatory dysfunction [23]. Subharmonic vibration of the vocal folds can result from tension, stiffness or mass lesion [23]. Vocal cord paralysis, scarring, inflammation, cyst or polyps are common causes of subharmonic vibrations [23]. Additionally, subharmonics can arise from poor control and coordination of the speech motor system. Signals activating laryngeal muscles may induce subharmonic vibrations when transmitted and/or produced inappropriately by damaged nervous system structures such as the basal ganglia.

For generations, the existence of subharmonic vibrations was known to speech processing engineers only as a marginal problem in F₀ detection. Subharmonic detection or subharmonic tracking have been seen as a misjudgement of F₀ favouring F₀/2 for shorter or longer periods of time, respectively. The majority of authors have focused on subharmonics only to overcome their delusional quality, whereas clinical aspects of their detection have been overlooked. The clinical assessment of subharmonics relies only on auditory perception or additionally on visual inspection of a spectrogram or detecting octave jumps in the pitch trace. The fundamental frequency can easily be identified in intervals when subharmonics are not present, i.e., intervals of regular vibrations [24]. However, when
subharmonics are present, decisions regarding $F_0$ become ambiguous as both the original $F_0$ and its subharmonic fraction coexist during subharmonic vibratory patterns \cite{9,20}.

Generally, detection of subharmonics requires exquisite integration of subharmonic measurement and robust identification of $F_0$ corresponding to the modal register, hereby modal $F_0$. Modal $F_0$ represents modulation conducted by phonatory muscles unlike subharmonics that arise from a constellation of conditions and physical properties of the vocal folds. Tracking modal $F_0$ thus could provide better insight into the pathophysiology of speech disorder than detected pitch and allows the detection subharmonics as a special event.

Detected subharmonics allow to describe the quality of voice with a simple metric that is not biased by the ambiguity of perception. We hypothesize that subharmonics are the key factors causing the specific manifestations of dysphonia in APS such as harsh voice, roughness of voice and pitch breaks. Therefore, detection of subharmonics may contribute to improved accuracy of diagnostic decision. Moreover, the detection of subharmonics can help to resolve the ambiguity of pitch by describing the pitch as a continuous trace of modal $F_0$ that can be modified by halving modal $F_0$ within subharmonic intervals or described directly by the depth of alternation; all can be done without necessity to construct complicated criteria for pitch jumps.

No such algorithm for acoustic analysis of modal and subharmonic vibrations in dysarthria is available to date. Demand for robust acoustic analysis of regular and subharmonic vibrations is magnified by a lack of clarity in descriptive terminology and unexplored mechanisms leading to aberrant vibrations of the vocal folds affected by dysarthria.

D. AIMS AND OBJECTIVES

We examined subharmonic vibrations in dysarthria with several aims:

(i) To develop a robust method for the detection of modal $F_0$ and intervals of subharmonic vibrations.

(ii) To design a method for tracking the reliability of $F_0$ detection from sustained vowels based on resynthesized prolonged vowels.

(iii) To compare the accuracy of the current method with a large set of publicly available state-of-the-art pitch detectors based upon resynthesized vowels and 337 phonations from 83 speakers along a spectrum of healthy to severe dysarthria,

(iv) To develop acoustic features allowing the examination of subharmonic vibrations and apply them to voice data in PD, MSA and PSP, to provide more insight into the pathophysiology of subharmonic vibrations due to basal ganglia dysfunction.

II. METHOD

A. SUBJECTS

A total of 83 Czech participants were included in the study. The majority of participants were originally recruited for the previous study focused on the detailed assessment of severity and patterns of dysarthria in PSP and MSA; characteristics related to subharmonic vibrations were not reported. The database consisted of 22 patients with PD (10 men, 12 women), 21 patients with MSA (9 men, 12 women) and 18 patients with PSP (12 men, 6 women). A cohort of 22 subjects (11 men, 11 women) with no history of communication or neurological disorder was included as a healthy control (HC) group. PD was diagnosed by Parkinson’s disease Society Bank Criteria \cite{25}, MSA by the consensus diagnostic criteria for MSA \cite{26}, and PSP by NINDS-PSP clinical diagnosis criteria \cite{27}. Disease duration was estimated based on the self-reported occurrence of first motor symptoms. A neurologist examined all patients and scored their motor ability accordingly to Unified Parkinson’s Disease Rating Scale motor subscore \cite{28} (UPDRS III) for PD and natural history and neuroprotection in Parkinson plus syndromes–Parkinson plus scale \cite{29} (NNIPPS) for MSA and PSP. Perceptual severity of speech impairment was enumerated by speech dysarthria items of relevant clinical scales. PD subjects were investigated in the on-medication state after at least 4 weeks of stable medication. PSP and MSA patients were medicated with various doses of dopamine agonists and/or amantadine. Characteristics of participants are summarized in Table 1. All participants provided written, informed consent. The study was approved by the Ethics Committee of the General University Hospital in Prague (approval number 67/14 Grant VES AZV 1.LFUK). The study was carried out in accordance with the approval guidelines.

B. RECORDINGS

All participants were recorded in a room with low ambient noise using headset condenser microphone Opus 55 (Beyerdynamic, Heilbronn, Germany) positioned approximately 5 cm from their lips. Recordings were digitalized with 16-bits resolution and 48 kHz sampling frequency. Every participant was instructed by a trained speech-language pathologist to produce prolonged vowel /A/ and /I/ (International Phonetic Alphabet) as long and steadily as possible using a modal voice. Each vowel was recorded at least two times. A total of 337 recordings were captured.

C. ACOUSTIC ANALYSIS

The presented acoustic analysis was adopted from first author’s doctoral thesis \cite{30}. The acoustic analysis was carried out in four steps. First, the signal was segmented to voiced and voiceless intervals. Then, modal $F_0$ was tracked using statistical modelling of modal $F_0$, which improves the selection of candidates when detection is corrupted by alternating vibrations, increased perturbation, or strong higher-order harmonics emphasize by formants. Values of modal $F_0$ were modelled by a normal distribution described by the mean and SD. A special algorithm for estimating distribution of speaker’s modal $F_0$ was used to initialize the
model. Values of modal $F_0$ were measured with regard to the speaker’s, and subharmonics were subsequently analyzed. The model was then adapted by predicting new $F_0$ value (i.e., mean of the model) with known uncertainty (i.e., the SD of the model) from measured modal $F_0$ using a Kalman filter. Finally, acoustic features were calculated from detected values of modal $F_0$ and time labels of subharmonic intervals.

The segmentation, initialization of the speaker’s model of modal $F_0$ and detection of $F_0$ are independent algorithms. Therefore the presented processing is linked only by the outcome such as labels of voiced signal and model of modal $F_0$. The recorded signal is also processed without linking to other algorithms.

1) SEGMENTATION

The recorded signal was decimated to a sampling frequency of 8 kHz and analysed in a sliding window of 50 ms duration, 5 ms step, and the Gaussian window with a length of 6 standard deviations (SD). The length of Gaussian window of 6 SD means that one SD has duration of 50/6 ms. The sampling frequency of 8 kHz was chosen as the frequencies below the Nyquist frequency of 4 kHz are the most important for differentiating voiced and voiceless intervals using given parameters. Window duration of 50 ms represents a trade-off between time-resolution of the segmentation and a number of glottal pulses covered. Power of the signal (PWR), maximal peak in normalized autocorrelation function (MPA) and zero-crossings rate of the autocorrelation function (ZCR) were computed for each position of the window. PWR was normalized by the maximal value of the recording. The first 10% of the recording was not included in the calculation of maxima as the human voice may produce exaggerated loudness in the initial setting. MPA was determined as the maximal peak in lags corresponding to a range of frequencies from 50 to 500 Hz of autocorrelation function. The values of autocorrelation function were corrected by the autocorrelation function of the Hamming window [31]. ZCR was computed as the zero-crossing rate of the autocorrelation function within a range of frequencies from 50 to 500 Hz, and was rescaled to Hz.

Each position of the sliding window was labelled as voiced if PWR was higher than -50 dB, MPA exceeded a value of 0.24 or the value of ZCR was in the range of 50–500 Hz. Voiced position was rejected when PWR was less than -80 dB. Since the PWR was previously normalized to be 0 dB for the maximal PWR value of the signal, the PWR threshold corresponds to the typical dynamic vocal range. MPA threshold assumed that the harmonic-to-noise ratio of the voiced signal exceeded -10 dB. The acceptable range of ZCR corresponded to the range of modal voice. Threshold values were adjusted to be more tolerant to perturbations of pathological voice. Only voiced intervals longer than 100 milliseconds were accepted.

2) ESTIMATION OF SPEAKERS’S MODEL OF MODAL $F_0$

The model was initialized by the following process (see Fig. 1A). The recorded signal within voiced intervals was decimated to a sampling frequency of 3 kHz, as higher frequency components are unnecessary for further analyses. Signal was analysed in a sliding window of 75 ms length and 7.5 ms step. The sampling frequency was determined as six times higher than the maximal detected frequency of 500 Hz, and thus the subsequent cepstral analysis is carried out of sufficient number of harmonics. Signal inside the sliding window was weighted by the Gaussian window with a length of 6 SD, zero-padded to a length of 512, and transformed to real cepstrum. A high-resolution Gaussian window was preferred as it suppresses undesired noise. The duration of the sliding window was preferably longer with regard to the temporal selectivity of the weighting window. Cepstrum corresponding to a frequency range of 50–500 Hz was further analysed. Maximal peak was detected in every position of the window. All peaks with a gain below the 90th percentile of all detected gamnitudes were discarded. The remaining peaks represented high-quality measurements of modal voice. Parameters of the $F_0$ model were estimated as the median and SD of frequencies corresponding to selected high-quality peaks (see Fig. 1B). The SD of the model was always kept above 10 Hz as too narrow a model can yield zero probabilities for outlying frequencies due to the limited numerical representation of small numbers in computers.

3) DETECTION OF MODAL $F_0$

The recorded signal was decimated to a sampling frequency 3 kHz to decrease computational burden resulting from the processing of higher frequencies that are redundant for the subsequent analysis. Voiced intervals of the signal were processed in windows with a duration corresponding to 10 periods of modal $F_0$ provided by initial estimation of modal $F_0$. The analysing window step was set to 10% of its duration. A Gaussian window with a length of 6 SD was preferred weighting for application of Gaussian interpolation in frequency measurement [32]. Note that the tapering function is much steeper than other conventional windows such as a hamming window. Thus, increased length of the analysing window is preferable.

Fast Fourier transformation was computed for each position of the sliding window zero-padded to a length of 1024 samples to decrease the spacing between frequency bins. The single-sided spectrum was normalized to unity sum. Local minimums and maximums were determined. When the interval between local extremes of the same category, i.e., minimum or maximum, was less than 25 Hz, less extreme values were discarded. Resolution of measured local extremes was refined using Gaussian interpolation [32].

Candidates of $F_0$ were selected from local maximums that showed prominence higher than 0 dB measured against the mean magnitude of the surrounding local minimums (see Fig. 1C). Subsequently, minimal magnitude of the surrounding local minimums was subtracted from the magnitude of each candidate in order to reduce influence of noise. A weighted frequency histogram was built from 7 harmonics accordingly to Schroeder [33]. Harmonics were determined by matching
candidates of \( F_0 \) to a corresponding harmonic series with a tolerance less than 1 semitone. The contribution of each harmonic was weighted by the inverse of its harmonic number, e.g., the weight of the fifth harmonic was 1/5. Each entry of the frequency histogram was calculated as a sum of contributions of all matched harmonics. Frequencies of frequency histogram were refined by averaging frequencies of all matched harmonics. Only entries with a total weight higher than 0.75 were accepted as an entry. Candidates of \( F_0 \) were extended by entry only when more than 30\% of its constituent harmonics were odd, and their distance to any other candidate was higher than half an octave. These conditions prevented the inflation of fractions of even harmonics. The frequency histogram allows the reconstruction of candidate \( F_0 \) from higher harmonics even when the \( F_0 \) was noisy or completely missing. Candidates with values lower than 50 Hz or higher than 500 Hz were rejected. These limits represent extended range of fundamental frequencies in adult speakers inspired by the default settings of range in PRAAT (75–500 Hz).

Local extremes were matched to the harmonic series of all \( F_0 \) candidates up to the 7th harmonics with a tolerance of less than 1 semitone. The amplitude of local maxima was accounted for with a positive sign, whereas local minima were accounted for with a negative sign. This approach emphasized the natural harmonic structure of maximum on the frequencies of the harmonic series and minima between. The probability of each candidate was determined by the mean amplitude of accounted extremes. All candidates were compared with the probability model of the modal register (see Fig. 1D). \( F_0 \) was identified as the candidate with maximum likelihood.

4) DETECTION SUBHARMONICS

The detection of subharmonics was performed in each position of the analysing window subsequent to \( F_0 \) detection. The harmonic series of \( F_0/2 \) was analysed in terms of subharmonic-to-harmonic ratio (SHR) defined as the ratio of even multiples of \( F_0/2 \) and odd multiples of \( F_0/2 \) [19] by the following equation:

\[
\text{SHR} = \frac{\sum_{i=1}^{N} A_i \left( \frac{f_0}{f_0 - 2} \right)^i}{\sum_{i=1}^{N} A_i (f_0/2)^i},
\]

where \( A \) is the amplitude of a given frequency obtained by the Fourier transform, \( f_0 \) refers to detected modal \( F_0 \), and \( N \) is a maximal number of harmonics. Frequencies of harmonic series up to 7 harmonics were matched with local extrema of the amplitude spectrum refined using Gaussian interpolation [32]. Amplitudes of frequencies outside the tolerance of 1 semitone were set to zero and amplitudes of local minima were negated to compensate for the influence of perturbation. The position of the window was labeled as subharmonic when SHR exceeded a threshold value of 0.1 (see Fig. 1F), which was adopted from perceptual experiments [18].

5) ADAPTATION OF THE STATISTICAL MODEL OF MODAL \( F_0 \)

Generally, any estimation of a model of modal \( F_0 \) calculated directly from previous measurements will be sensitive to misdetections that make the scope of the model wider (i.e., increase SD) causing inflation of more and more errors. A Kalman filter is preferable for this task as its error covariance is bounded. \( F_0 \) was modelled by a Kalman filter as a first order linear system corrupted by stationary Gaussian noise. We assume that fluctuating \( F_0 \) denoted as \( x \) evolves from previous state \( x_{t-1} \) according to the following equation:

\[
x_t = F \cdot x_{t-1} + v_t,
\]

where \( F \) is the state transition model to update the previous state and \( v_t \) is the process noise, which is assumed to be normally distributed with zero mean and covariance \( Q \). The true state \( x_t \) is observed as \( z_t \) in the form:

\[
z_t = H \cdot x_t + w_t,
\]

where \( H \) maps the true state into observed space and \( w_t \) is the observation noise assumed to be normally distributed with zero mean and covariance \( R \).

\( F_0 \) was tracked by a constant velocity model, where the state is defined by position (i.e., \( F_0 \) ) and velocity (i.e., melodic change):

\[
x_t = \left( \begin{array}{c} f_0 \\ \frac{d f_0}{dt} \end{array} \right),
\]

where \( f_0 \) represents \( F_0 \) and \( \frac{d f_0}{dt} \) its derivation. The state transition matrix \( F \) was then defined as:

\[
F = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix},
\]

where \( T \) is a step of the sliding window. The control matrix \( H \) was set to:

\[
H = \begin{pmatrix} 1 \\ 0 \end{pmatrix},
\]

because only the \( F_0 \) is measured. We assumed no relationship between \( F_0 \) and melody, the noise-free state of \( f_0 \) and imperfect estimation of melody:

\[
Q = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_{f_0}^2 \end{pmatrix},
\]

where \( \sigma_{f_0}^2 \) is variance of the initial \( F_0 \) model. The scope of our adapted model can be conveniently controlled by adjusting the covariance of measurement noise. The value of \( R \) was set to the error of measurement, which in our case strongly depends on the uncertainty of our initial estimation:

\[
R = \sigma_{f_0}^2.
\]

The error covariance matrix was initialized by the error of the initial measurement, which is the variance of the initial \( F_0 \) model:

\[
P = \begin{pmatrix} \sigma_{f_0}^2 & 0 \\ 0 & \sigma_{f_0}^2 \end{pmatrix}.
\]

The prediction step of Kalman filtering was realized by the following equations:

\[
\hat{x}_{t|t-1} = F \cdot \hat{x}_{t-1|t-1},
\]

\[
P_{t|t-1} = F \cdot P_{t-1|t-1} \cdot F^T + Q,
\]

where \( \hat{x}_{t|t-1} \) represents the predicted state of modal \( F_0 \) and \( P_{t|t-1} \) predicted error covariance. The Kalman filter was updated by following equations:

\[
K_t = P_{t|t-1} \cdot H_t^T \cdot (H_t \cdot P_{t|t-1} \cdot H_t^T + R_t)^{-1},
\]

\[
P_{t|t} = P_{t|t-1} - K_t \cdot H_t \cdot P_{t|t-1},
\]

\[
\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t (z_t - H \cdot \hat{x}_{t|t-1}).
\]
where $K_t$ represent Kalman gain, $P_{0t}$ is updated error covariance, and $x_{0t}$ updated state of modal $F_0$.

$F_0$ model was determined by updated state $x_{0t}$ representing the mean and the first element of error covariance representing the variance of the model (see Fig. 1E). We assume that the $F_0$ is continuous as long as the voicing itself is continuous. When the voicing suddenly stopped in the middle of phonation causing a silence and started the phonation again, the previous tracking of $F_0$ was stopped, the Kalman filter was reset and a new track of Kalman filter was initialized using the initial $F_0$ model.

6) ALGORITHM OUTCOME

The following rules were applied to the measurement of modal $F_0$ to compensate for imprecisions in segmentation and short-term erroneous decisions. When only one glottal pulse is in the scope of a sliding window, which can happen typically on the border of the voiced interval, extreme values of $F_0$ can be detected. Therefore, all values of $F_0$ distributed 8 SD from the respective $F_0$ model mean were discarded and the position was labeled as unvoiced. This rule was applied after all values were measured, which allows adaptation of the detector to all possible scenarios even when the $F_0$ model was not properly initiated. The resulting $F_0$ envelope was smoothed by the median filter of the third order, which efficiently reduces secluded values while keeping outstanding time resolution.

Detected modal $F_0$ represents perceived pitch only during intervals of regular vibrations. The perceived pitch during subharmonic intervals corresponds to modal $F_0$ divided by two at intervals of subharmonic vibrations (see Fig. 1G).

Intervals of subharmonic vibrations were identified from detected subharmonics using the following rules. The decision about subharmonics was smoothed by a sliding median filter of seventh order. All neighboring subharmonics were fused when the interval between them was shorter than 300 ms and no pause interval was present. Only subharmonic intervals longer than 50 ms were accepted. Each subharmonic interval was described by the time label.

In summary, the algorithm provides labels of regular and subharmonic intervals, values of modal $F_0$, and measured values of SHR.

7) DESCRIPTIVE SPEECH FEATURES

Variation of melody was assessed as the SD of modal $F_0$ expressed in semitones (stdF0). The SD was estimated using median absolute deviation rescaled to the quantile of SD.

The proportion of subharmonic intervals (PSI) was calculated as the ratio of the total duration of subharmonic intervals per total duration of all voiced segments including regular and subharmonic intervals.

The depth of subharmonic alternation (DSA) was estimated as mean SHR calculated within detected subharmonic intervals.

The location of initial subharmonic (LIS) helps to explain whether the subharmonics arose as a result of fatigue or weakening respiratory support at the end of the phonation, inappropriate initial settings of the vocal folds, or other causes.

LIS was determined by the initial time of the first subharmonic interval occurring in the course of phonation.

D. TRACKING DETECTION ACCURACY

1) REFERENCE SIGNALS

The evaluation of detection accuracy requires precise reference of $F_0$ and subharmonics. An electroglottogram is a common reference of $F_0$, but no standard for subharmonics is currently established. We thus considered synthesizing reference signals to control for the presence of subharmonics. However, no synthetic data can mimic the natural manifestations of various dysarthrias. Therefore, we replicated each acoustic signal based on supervised parameterization to obtain authentic reference signals with the known time course of $F_0$ (see Fig. 2). The envelope of $F_0$ and related perturbation metrics represented by jitter, shimmer and harmonic-to-noise ratio were measured using PRAAT (http://www.fon.hum.uva.nl/praat/). First, intervals of subharmonic vibrations were labeled based on visual analysis of the pitch trace detected by PRAAT. Then, the time course of $F_0$ was inspected and octave jumps were manually corrected in the PRAAT pitch editor by selecting pitch candidates corresponding to the modal register to obtain the most consistent and smooth estimation of $F_0$ with no octave jumps. Perturbation represented by jitter, shimmer and HNR was measured in thoroughly-selected voiced intervals with no evidence of subharmonic vibrations (see Fig. 2B).

Glottal pulse shape was additionally measured by fitting the Liljencrants-Fant model of the glottal pulse to inversely filtered signal using a manually-controlled TKK apparat [34] (https://sourceforge.net/projects/aparat/; see Fig. 2B). A Kronecker delta series was generated in the period derived from the measured time course of $F_0$. The amplitude and position of impulses were randomly altered to match measured shimmer and jitter (see Fig. 2C). Subharmonic vibrations were simulated by damping every second impulse. The difference between amplitudes of two adjacent impulses was defined by the amplitude modulation index estimated as the median SHR using subharmonic summation [21] in all intervals of subharmonic vibrations. A sample of the Liljencrants-Fant derivative pulse (LF) was computed from measured parameters and convoluted with the series of impulses (see Fig. 2D). The spectrum of each signal was parameterized by linear predictive coding (LPC) in a sliding window with 75 milliseconds duration and 10 milliseconds step (Fig. 2E). The order of the model in each position of the sliding window was selected in the range of 20–60 using minimum description length criterion (MDL) computed from the error signal. The glottal pulse series was then filtered by the series of LPC coefficients obtained from the original signal (Fig. 2F). Glottal noise was generated as a normally distributed random vector modulated by the unfiltered glottal pulse series. Amplitude modulation was performed to emulate the pitch–synchronous character of natural glottal noise. Subsequently, glottal noise was filtered by the LPC coefficients estimated from the
original signal (Fig. 2G) and added to the filtered series of glottal pulses in a ratio corresponding to the measured harmonic-to-noise ratio (Fig. 2H). Glottal pulses and glottal noise were filtered separately as it is not possible to accurately match harmonic-to-noise ratio prior to filtering by varying LPC coefficients. The resulting replicates of the original signal served as a precise reference for detection accuracy evaluation.

2) ACCURACY ANALYSIS

The accuracy of \( F_0 \) estimation and detection of subharmonic vibrations was computed by comparison with reference signals. The performance of \( F_0 \) detection was scaled in semitones to reduce the difference between the variability of lower and higher pitch. The initial estimation of modal \( F_0 \) was evaluated using Pearson’s correlation coefficient between the initial modal \( F_0 \) and median \( F_0 \) of reference signals. Pearson’s correlation was preferred because the initial estimation of \( F_0 \) does not have to be perfectly precise but must be stable over all possible values of various speakers. Tracking of modal \( F_0 \) was analyzed in terms of errors between detected modal \( F_0 \) and reference in corresponding time. Mean semitone error (ME), the SD of error in semitones (SDES), root means square error in semitones (RMSE), and median absolute semitone error (MAE) were defined as follows:

\[
e_{n} = 12 \cdot \log_{2} \hat{y}_{n} - 12 \cdot \log_{2} y_{n},
\]

\[
ME = \frac{1}{N} \sum_{n=1}^{N} e_{n},
\]

\[
SDE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (e_{n} - \bar{e})^2},
\]

\[
RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (e_{n})^2},
\]

\[
MAE = \text{median}|e|,
\]

where \( e_{n} \) is difference between the \( n \)-th estimation of the \( F_0 \) value \( y_{n} \) and corresponding reference \( y_{n} \). The symbol \( N \) denotes a total number of measurements. Only intervals labeled as voiced by the reference and algorithm at the same time were evaluated in order to avoid resulting from the systematic failure of the voiced/unvoiced decision made by an algorithm.

We searched English literature in the Web of Science, Google Scholar, and IEEE Xplore for single-track pitch detectors suitable for the analysis of sustained vowels. The criteria were noise-robustness and availability of code or executables. The accuracy of the proposed \( F_0 \) detector was compared with a set of 19 publicly-available, state-of-the-art detectors. Speech Filing System (SFS; http://www.phon.ucl.ac.uk/resource/sfs/) provided \( F_0 \) detectors based on autocorrelation of the cubed signal without tracking (FXAC) and with a tracking algorithm (FXANAL) designed by Secrest and Doddington [35]. SFS also offers a robust algorithm for pitch tracking applied on normalized cross-correlation (FXRAPT). An autocorrelation-based detector with comprehensive post-processing (PRAAT_AC) is implemented in PRAAT. The cross-correlation version of the PRAAT_AC algorithm (PRAAT_CC) was also analysed. An autocorrelation-based detector with a robust algorithm for pitch tracking (RAPT) was proposed by Talkin [36] and is incorporated in VOICEBOX (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.htm). The sophisticated approach developed by David Talkin at Google (https://github.com/google/REAPER) employs dynamic programing to track the normalized cross-correlation of glottal closure instants obtained from the residuals of linear predictive coding (REAPER). An intriguing multiband summary correlogram [37] (MBSC) method was tested using the shareware code (http://www.ee.ucla.edu/~spapl/code/MBSC_matlab.zip). A sophisticated algorithm from Yet Another Glottal source analysis framework (YANGsaf; https://github.com/google/yang_vocoder) developed by Kawahara [38] at Google provides the \( F_0 \) trajectory refined by analysis of the harmonic component and \( F_0 \) adaptive time warping of an initial estimate based on instantaneous frequency and outputs of band-pass filters. The famous Noll’s cepstrum-based detector [39] was implemented in SFS (FXCEP). Spectral compression techniques were represented by computationally efficient subharmonic summation on logarithmic frequency mantissa [40] provided by PRAAT (PRAAT_SHS) and utilized in the SHRP algorithm (https://www.mathworks.com/matlabcentral/fileexchange/1230-pitch-determination-algorithm) by Sun [21]. A sawtooth waveform-inspired pitch estimator for speech and music (SWIPE) by Camacho et al. [41] provided in the author’s dissertation [42] optimises the estimation of \( F_0 \) as the best match between the spectrum of a periodic sawtooth and analysed spectrum. \( F_0 \) derived from positions of glottal closure instants was estimated by the dynamic Programming Projected Phase-Slope Algorithm [43] (DYPSCA) released in VOICEBOX. Another detector from a spectral domain called Pitch Estimation Filter with Amplitude Compression [44] (PEFAC) available at VOICEBOX applies dynamic programming for selection of the optimal \( F_0 \) candidates determined from the outcome of a special harmonic summation filter that integrates broadened harmonic peaks and rejects additive noise. The harmonic modelling and tracking is represented by the Robust Bayesian Pitch Tracking Based on the Harmonic Model (BFONLS) [45]. DIO [46] incorporated in the WORLD vocoder (https://github.com/mmorise/World/) performs a selection of \( F_0 \) candidates obtained from zero-crossings of the low-pass filtered signal. HARVEST [47] available in the WORLD vocoder calculates smoothed \( F_0 \) contour based on refined \( F_0 \) candidates estimated from the signal filtered by a bank of band-pass filters.

Some algorithms may be more robust against subharmonic vibrations while some may be susceptible. The following two types of evaluation were introduced to ensure equal opportunities for both types of algorithms. First, the detected \( F_0 \) was compared to the modal \( F_0 \), i.e., subharmonic intervals were not accounted for as octave jumps. Second, the detected
F₀ was compared to the F₀ derived from the modal F₀ by halving the modal F₀ frequencies during intervals of subharmonic vibrations. Performance of an algorithm was then described by the evaluation with the lowest RMSE.

For all tested algorithms, the range of fundamental frequencies was set from 50–500 Hz to make the performances comparable to each other and the proposed method. All algorithms were applied using default settings except for SHRP, which deserved deeper analysis of subharmonic correction. Testing the algorithms in their default setting was motivated by the fact that algorithms are based on different principles and all their parameters suppose to be tuned for the task, so the standardization of setting such as window duration could cause a bias of their performance. In order to examine SHRP’s ability to detect modal F₀ and F₀ regardless of subharmonics, SHRP was tested two times with a SHR correction threshold set to 0.1 and 1.

To evaluate the initial estimation of the modal F₀ presented in this study, we compared the values of the estimated modal F₀ with the median reference modal F₀ values using Pearson’s correlation coefficient. The robustness of subharmonic detection was evaluated by comparing the position of detected and true edges of subharmonic intervals. Only subharmonics longer than 100 milliseconds were analyzed. Each recording was scored by an F-score calculated by the following equations:

\[
\text{precision} = \frac{TP}{TP + FP},
\]

(20)

\[
\text{recall} = \frac{TP}{TP + FN},
\]

(21)

\[
F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},
\]

(22)

where TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives, and F is the resulting score. Only the edges of detected subharmonic intervals closer than 100 ms from the corresponding true subharmonic intervals were accepted as true positives. All detected edges outside the tolerance interval were considered false positives. All edges of reference intervals that were not matched with detected intervals were labeled as false negatives. Average and SD of scores were computed across all recordings.

Further evaluation was focused only detecting the presence or absence of subharmonics in a recording irrespective of temporal localization of the detected subharmonic intervals. When any subharmonic interval was present in a recording, the recording was labeled as positive when no subharmonic interval was present anywhere in a recording, the recording was labeled as negative. Overall accuracy, sensitivity and specificity were evaluated based on comparison of negativity or positivity of the reference with detection outcome.

No subharmonic detectors have been published to date allowing comparison. The possibility of detection without tracking the modal register was analyzed on a simple subharmonics detector based on thresholding the SHR parameter computed by the SHRP detector with the same decision smoothing as the proposed method applies.

Correlation of the estimated acoustic feature DSA to the reference values of amplitude modulation index was quantified using Pearson’s correlation coefficient.

E. STATISTICS

Group differences in the features stdF₀, LIS, PSI, SHI and DSA were tested by one-way ANOVA with post hoc Tukey's honestly significant difference. The one sample Kolmogorov-Smirnov test and Bartlett test were applied to test assumptions of ANOVA regarding normality and homogeneity of variance, respectively. Logarithmic transform was applied to the data prior to ANOVA to treat log-normality and/or heteroscedasticity. Effect of the task on the presence of subharmonics was estimated by comparing proportions of speakers that presented subharmonics in all repetitions across individual tasks /A/ and /I/ using McNemar’s test. Proportions of subjects were compared using the χ² test for goodness-of-fit followed by post hoc z-test with Bonferroni correction. Relationships between speech features and clinical scales were evaluated by Spearman’s correlation coefficient. The level of significance was set to 0.05.

Only subjects that showed subharmonics in at least one repetition were included into group comparison of LIS, DSA and PSI since LIS and DSA could be measured only on subharmonic intervals and comparison of PSI could be biased by disproportionality of subjects that manifested no subharmonics.

III. RESULTS

A. DETECTION ACCURACY

Figure 3 compares the performance of the presented method and publicly available F₀ detectors. Table 2 summarizes all measurement accuracies including the reference of modal F₀ and reference regarding subharmonics. The presented method outperformed other detectors in all terms. Only DIO, DYPsA, HARVEST, REAPER, YANGsaf, and the presented method showed lower RMSE of modal F₀ than RMSE of F₀ regarding subharmonics. The initial estimation of mean modal F₀ was strongly correlated with reference (r = 0.99, p < 0.001). The presented method detected individual intervals of subharmonic vibrations with a mean F-score of 91.25% ± 21.85 SD, mean precision 92.06% ± 22.35 SD, and mean recall 90.21% ± 23.56 SD. The presence or absence of subharmonics in the recording was detected with 97.03% accuracy, 98.70% sensitivity and 93.46% specificity. Automatically estimated DSA values showed strong correlation (r = 0.91, p < 0.001) with reference values.

A detector based on thresholding SHR by SHRP identified subharmonic intervals with a mean F-score of 48.20% ± 46.43 SD, mean precision of 47.39% ± 47.10 SD, and mean recall of 47.67% ± 46.74 SD. The presence or absence of subharmonics was estimated with 55.20% accuracy, 87.96% sensitivity and 39.74% specificity.
B. CLINICAL FINDINGS
Manual inspection of the spectrogram and pitch trace in PRAAT discovered no voice break into upper register in the whole database. More than 72% of HC manifested subharmonics in at least one recording. Subharmonic intervals were observed for all repetitions of the vowel /A/ in 45% of speakers and less frequently for phonation of the vowel /I/ in 25% of subjects ($\chi^2 = 8.60, p = 0.003$). Subharmonics were detected in all repetitions for 9% of HC, 23% of PD, 43% of MSA and 17% of PSP ($\chi^2 = 7.51, p = 0.06$).

The speech feature results are summarized in Fig. 4. Statistically significant differences were found in acoustic features including stdF0 ($F(3,79) = 5.54, p = 0.002, \eta^2 = 0.17$), LSI ($F(3,79) = 8.82, p < 0.001, \eta^2 = 0.3$) and PSI ($F(3,79) = 4.41, p = 0.007, \eta^2 = 0.18$). An omnibus test on DSA was not significant ($F(3,30) = 0.7, p = 0.56, \eta^2 = 0.03$). Post hoc analysis showed significantly increased stdF0 only in the MSA group. The features LSI and PSI were abnormal for both MSA and PSP.

Acoustic feature stdF0 showed mild correlation to overall NNIPPS ($r = 0.36, p = 0.02$). Severity of speech was reflected moderately by LIS ($r = -0.45, p < 0.001$) and stdF0 ($r = 0.48, p < 0.001$). We identified mild correlation between stdF0 and maximum phonation time ($r = -0.42, p < 0.001$), PSI and stdF0 ($r = 0.38, p < 0.002$), LSI and stdF0 ($r = -0.39, p = 0.001$) and moderate correlation between LSI and maximum phonation time ($r = 0.64, p < 0.001$), LSI and PSI ($r = -0.54, p < 0.001$) and PSI and DSA ($r = 0.64, p < 0.001$).

IV. DISCUSSION
The present study introduces a new perspective on the acoustic analysis of prolonged phonation in dysarthria by tracking the fundamental frequency in the modal register followed by analysis of subharmonic vibrations. This study describes a method focused on the accurate detection of subharmonic intervals. By comparing the algorithm accuracy to reference labels based on synthetic replicas, our method was capable of detecting not only the presence or absence of subharmonics in speech (accuracy 90.03%) but also the precise position of individual subharmonic intervals (F-score 91.25%). Successful performance of the method indicates that tracking the modal register is not only pivotal in the detection of subharmonic vibrations but also very convenient for the analysis of voice modulation. Automated analysis showed excess modulation of melody in MSA. PSP and MSA manifested subharmonic onset early in the course of phonation as well as an increased incidence of subharmonics indicating the distinct impact of widespread neurodegeneration on control of the vocal folds. Despite common the occurrence of subharmonic intervals in healthy speech, decreased stability of the natural vibrational regime of the vocal folds can indicate cortical atrophy due to neurodegeneration. The presented acoustic analysis may help speech pathologists to objectively quantify excess modulation of melody and the phenomena of subharmonic vibrations using non-invasive acoustic measures that may be potentially applicable in a wider variety of diseases. We believe that the proposed technology will raise awareness of subharmonic vibrations in neurodegeneration among clinicians and boost further development of objective acoustic measures.

We provide a comprehensive comparison of publicly available pitch detectors that demonstrated the need for a specialised approach in the analysis of vocal fold vibrations in prolonged vowels. With regard to the measured performance of pitch detectors based on autocorrelation or cross-correlation functions such as FXAC, FXRAPT, PRAAT_AC, PRAAT_CC, RAPT and MBSC, we found that waveform-matching detectors tend to be sensitive to subharmonics and recommend that any metric based on these detectors should be interpreted with regard to possible subharmonic bias. Alternation of glottal pulses propagates directly to measured patterns, which makes tracking modal $F_0$ almost impossible when the depth of amplitude modulation approaches 100%. Subharmonic tracking is a well-known weakness of these detectors, and their decision process is usually tuned to be less prone to subharmonics or to follow the perceptual pitch. The robustness of waveform-matching techniques to subharmonic tracking can be dramatically improved when the matching waveform is determined from time domain events such as instants of glottal closures as it is implemented in REAPER. Generally, time domain algorithms such as DYPESA and DIO can be very successful in tracking modal $F_0$ as long as time markers can be successfully captured. Spectral and cepstral domain algorithms such as YANGsaf and PRAAT_ShS are more advantageous in tracking modal vibrations of vocal folds as both modal $F_0$ and subharmonics manifest in the spectrum at the same time, and the main challenge of these detectors is to consider which frequency corresponds to pitch. Unfortunately, only YANGsaf and HARVEST showed lower RMSE in tracking modal $F_0$ compared to RMSE of $F_0$ regarding subharmonics. Nevertheless, differences between both types of RMSE for other spectral domain detectors such as SWIPE, SHRP, BF0NILS, PRAAT_ShS and PEFAC were lower than 1 semitone, which indicates that subharmonic tracking was inconsistent and imprecision was not related specifically to the presence of subharmonics. All detectors showed positive ME, which suggests a tendency to confuse higher frequencies with true $F_0$. Finally, subharmonic tracking represents a substantial, but not the only, obstacle in the assessment of melodic variation. Misjudgments higher than several semitones were substantially more important source of errors than small inaccuracies due to local perturbation of $F_0$. Thus, the fact that the processed frames were not exactly the same for all algorithms due to use of their different default windowing did not influence the results of evaluations.

We propose measuring melodic variation on modal $F_0$ rather than pitch for two reasons. First, using modal $F_0$ allows the measurement $F_0$ variability independent of subharmonics. Tracked subharmonics can significantly distort examination results by increasing variability and/or range of $F_0$. A potential
pittal in such an analysis may be illustrated on a hypothetical speaker manifesting simultaneously subharmonics and monopitch. The estimation of melodic variation from $F_0$ that tracks subharmonics can, however, infer the opposite, i.e., excess variability of the fundamental frequency. On the contrary, abnormal variability of $F_0$ in ataxic dysarthria cannot be distinguished from subharmonics without intricate analysis of the pitch contour. Although the trained ear of a speech specialist can distinguish between pitch breaks and abnormal variability of melody, a simple metric based on pitch detection can hardly be that responsive. Given the above, only precise tracking of modal $F_0$ could provide reliable metrics in agreement with established perceptual pitch characteristics, although pitch as the perceptual feature is not the subject of measurement. Second, measurement of subharmonics requires targeting the frequency to which subharmonics are measured. The measurement of subharmonics with regard to other frequencies than the modal $F_0$ can yield disastrous results. This was the cause of low accuracy in the detection of subharmonic intervals derived from the SHRP algorithm. Interestingly, SHRP showed higher errors for modal $F_0$ when no subharmonic correction was applied compared to errors of modal $F_0$ when the subharmonic correction was applied. These results indicate that SHRP frequently picked a higher octave than the modal $F_0$ and measured modal $F_0$ as if it were a subharmonic. Subharmonic correction of SHRP compensated for the higher octave misjudgement error, which increased the precision of $F_0$ detection but resulted in inflation of false positive subharmonic decisions (SHRP precision of 47.39%). Additionally, when a higher octave is misjudged without correction, the recall (SHRP recall of 47.67%) and overall F-score is affected as well. Since precision and recall were balanced in SHRP (F-score 48.20%), we may conclude that the problems of subharmonic detection were related to both types of errors.

The proposed method achieved excellent accuracy in tracking modal $F_0$ via initial estimation of modal $F_0$ followed by a probability-driven selection of candidates and subsequent adaptation of the probability model. The technical problem asked by Weismer [20], “What is the true fundamental frequency?,” is answered here by the assumption that the true $F_0$ will more likely be the one with higher regularity since vibrations are more regular when the vocal folds operate in their natural settings, i.e., modal register. When we are not able to decide which $F_0$ candidate is the true one, we can look at other measurements that show a more confident decision and build a model. Indeed, our findings suggest that the introduction of a model based on the selection of high-quality measurements can improve the accuracy of $F_0$ detection enormously. Our results emphasize the relevance of measurements in the Fourier spectrum for a medical grade analysis of vocalization and show that selection of $F_0$ candidates directly from the Fourier spectrum could outperform other methods, when a decision about $F_0$ is based on robust $F_0$ modelling. The presented approach upgrades the role of the Kalman filter in $F_0$ tracking that has been limited only to $F_0$ smoothing [47] and fusing measurements of multiple $F_0$ detectors [15]. Regarding relatively high and balanced F-score (91.25%) in the detection of subharmonic intervals, we hypothesize that proposed $F_0$ tracking based on statistical modelling of modal voice provided a reliable baseline for detection of subharmonic intervals. Also, the presented method for estimation of SHR by peak-picking subharmonic and harmonic series directly in the spectrum contributed to good reliability because the process allows the penalization of false peaks that would be accumulated into the resulting SHR. Note that the additive noise is canceled by the division in the definition of SHR, which further increases robustness of the measurement. The window with wide mainlobe also increased accuracy of the measurement because the spectrum is more smoothened. Nevertheless, the principle of SHR determination is applicable using other windowing functions as well. The proposed method achieved excellent performance on a given task and showed the potential to bring new objectivity and clarity into the clinical assessment of prolonged phonation in dysarthria.

Observed excess melody variation corresponded with the incidence of involuntary movements and discoordination of mixed dysarthria in MSA [2]. We assume that $F_0$ instability related to unsteady tension of laryngeal muscles may be another factor that increases the variability of modal $F_0$. Melody variation in PD and PSP showed no significant deviation from the norm, which is in concordance with the nature of the disease. This finding is also in line with a previous study suggesting that MSA patients manifest overall poorer voice control in comparison with PSP [4].

Subharmonics were detected in more than 72% of healthy subjects, highlighting subharmonics as a natural phenomenon of human vocalisation. Indeed, the presence of subharmonics per se can be too ambiguous to indicate pathological voice regarding the weakly insignificant observation of subharmonics in all repetitions for disease groups. Although these findings cannot suggest a link between subharmonics and neurogenic disorder, we have to stress that approximately 9% of HC subjects showed subharmonics consistently; thus, the presence of subharmonics in individual speakers should be interpreted in the wider context of other manifestations and patient history. Furthermore, the significant effect of task we observed suggests that more factors such as more limited airflow in the vowel /I/ than in the vowel /A/ can possibly influence the presence of subharmonics. Interestingly, depth of alternation is not a differentiating factor between pathological and healthy phonation, as one could incorrectly expect based on previous perceptual experiments demonstrating the importance of modulation depth for determination of pitch break [18,19].

The proportion of subharmonic intervals per total time of phonation and the location of subharmonics within the phonation is important for assessment of subharmonics. The proportion of subharmonics was significantly increased in
MSA and PSP. We hypothesize that decreased speech motor control may impair the self-sustaining mechanism of vocal fold vibrations, which causes persistent aberrant vibrations when subharmonics are triggered. The tendency to transition into the subharmonic vibrational regime and persist reflected by PSI could describe severity, whereas inspection of location could clarify origin.

We showed that subharmonics in MSA and PSP subjects are more likely to arise earlier in the course of phonation. As even small disruptions of neuromuscular input can completely reorganise interactions in the nonlinear mechanism of voice production [20,49], we can look to the chaotic nature of subharmonics as a stochastic process and assume that the increased probability of subharmonics simply results in subharmonics manifesting earlier. Also, the correlation between LIS and maximum phonation time may imply that muscle fatigue and/or weak respiration introduce instability to the nonlinear mechanism of voice production, which gradually increases the chance to transition to a subharmonic vibrational regime over time.

Subharmonics were generally abnormal in proportion and location in APS but only a weak trend towards significance was observed in PD. We hypothesize that severe subharmonics in APS are caused by different factors than hypokinetic components in mixed dysarthria. The spastic components may possibly detune functional symmetry of the vocal folds and thus induce a subharmonic regime. We have to note that the overall severity of APS subjects was higher compared to PD, which may also contribute to the increased presence of subharmonics. Finally, although diverse factors participate in the phenomena making the interpretation more complex, these results demonstrate that thorough acoustic analysis of subharmonic vibrations can bring insight into speech motor control affected by dysarthria.

Application of the presented technology is currently limited to subharmonic non-diplophonic voices, i.e., a single F0 with possible subharmonics. We assume that double pitch could be tracked with a similar methodology, but a more advanced initialization of the F0 model would be required. Based on the analysis of accuracy in phonation with excess melody variation, we suggest that the method could be successful in tracking the melody of connected speech. Nevertheless, a thorough evaluation of the method on connected speech with regard to subharmonics as well as other non-modal vibrational regimes such as vocal fry would be required. The method was evaluated on synthetic replicas obtained via LPC, which may cause some distortion when poles of the filter fit into harmonic or subharmonic frequencies that do not correspond to the measured F0 or subharmonics. Subharmonics not captured by manual measurement using PRAAT may especially propagate into synthetic replicas. Nevertheless, no redundant subharmonics were discovered by manual inspection of false positive subharmonic detections. For this reason and because the resonances were the only totally unsupervised measurements used for replication, we hypothesize that inaccuracies of the measurements can make the replicated signal different from the original but do not bias the reference related only to replicated signals. Only F0/2 subharmonics were analysed in the present study, as intervals of F0/3 subharmonics were very rare, always accompanied the more frequent F0/2 subharmonics, and did not dominate any recording in the database. We suggest that the detection of F0/2 is sufficient for the assessment of instability of vibrational regimes in dysarthria. Although the evaluation was limited on the most popular vowels /A/ and /I/, we hypothesize that the methodology could be applied to any other type of vowel because the period histogram proved to be applicable regardless on resonances and can reconstruct F0 even in cases when first harmonic is weak or missing [33]. Nevertheless, we suggest testing the methodology also on other vowels of interest. The presented method for the detection of subharmonics cannot be applied to the detection of a transition from modal to pulse register. Although this may be possibly related to the instability of the modal register, our algorithm was designed to track the change of modal F0 into low frequencies since we assume that subharmonics appear suddenly. We recommend that analysis of vocal range or inspection of F0 contour may be more suitable for the detection of this abnormality. Only freely-available pitch detectors were used for comparison with our method, whereas other promising approaches for harmonic detection and tracking are available (for example see [50-51]). The final limitation is that the proposed method was designed for sustained vowels, which are standard for evaluation of dysphonia. We hypothesize that only minor changes in the principle would be required to apply the proposed detection of modal F0 and subharmonics on the connected speech. Nevertheless, a thorough evaluation using the signals of connected speech resynthesized according to the presented method would be required.

We address a key issue of vibration analysis of prolonged vowels in dysarthria associated with PD and related diseases and provide a solution with superior accuracy. The advantage of the proposed method is that it allows deconvolution of pitch into the modulation of laryngeal muscles and the effect of subharmonic vibrations. Identification of regular and subharmonic intervals could have an immense impact on the development of new technologies for the clinical assessment of dysphonia and have the potential to upgrade perturbation measurements such as jitter, shimmer, harmonic-to-noise ratio, or cepstral peak prominence that can be considerably biased by the presence of subharmonics [9]. The technology for the detection of subharmonics may also influence future development of speech synthesis that is becoming more aware of the necessity of considering subharmonics in more natural-sounding speech models. Finally, the increased occurrence of subharmonics was found even in the prodromal stages of PD and Huntington’s disease [52-54], suggesting that the analysis of subharmonics may be suitable for the early detection of neurodegeneration.
ACKNOWLEDGMENT

J.H. and J.R. conceived and designed the experiments. J.H., R Č, J.K., and J.R. performed the experiments. J.H. developed and tested analysis method used in the project, analysed the data and wrote the paper. All authors were involved in discussion of results and revision of the manuscript.

DATA AVAILABILITY

All characteristics of subjects analysed during this study are included in this published article (and its Supplementary Information files). The reference signals generated during the current study are available in the figshare repository with the identifier https://doi.org/10.6084/m9.figshare.7628819.

REFERENCES


## TABLE I

<table>
<thead>
<tr>
<th>Group</th>
<th>Age (years) Mean/SD (range)</th>
<th>Disease duration (years) Mean/SD (range)</th>
<th>Disease severity # Mean/SD (range)</th>
<th>Speech severity χ Mean/SD (range)</th>
<th>MPT (s) Mean/SD (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD (hypokinetic)</td>
<td>64.4 / 9.6 (48–82)</td>
<td>9.3 / 5.5 (1–24)</td>
<td>15.9 / 7.6 (6–34)</td>
<td>0.7 / 0.7 (0–2)</td>
<td>15.1 / 4.4 (9.1–29.3)</td>
</tr>
<tr>
<td>PSP (hypokinetic-spastic)</td>
<td>66.7 / 6.5 (54–84)</td>
<td>3.7 / 1.4 (2–7)</td>
<td>70.6 / 25.4 (19–116)</td>
<td>3.8 / 1.4 (2–6)</td>
<td>13.5 / 5.0 (7.7–23.7)</td>
</tr>
<tr>
<td>MSA (ataxic-hypokinetic)</td>
<td>61.0 / 6.5 (45–71)</td>
<td>4.1 / 1.3 (2–7)</td>
<td>76.4 / 23.7 (35–123)</td>
<td>3.3 / 1.2 (1–6)</td>
<td>13.2 / 5.4 (5.8–26.1)</td>
</tr>
<tr>
<td>HC (none)</td>
<td>63.6 / 10.0 (41–79)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>20.0 / 7.1 (8.4–34.4)</td>
</tr>
</tbody>
</table>

# Scores on the Unified Parkinson’s Disease Rating Scale III (UPDRS III) for PD (ranging from 0 to 108) and Natural history and neuroprotection on Parkinson (NNIPPS) for MSA and PSP (ranging from 0 to 332) motor subscore (ranging from 0 to 124), higher scores indicate more severe disability.

χ Scores on the UPDRS III 18 speech item for PD and NNIPPS Bulbar-pseudobulbar signs subscale (item 3) for MSA and PSP. All scores represent speech motor examination and range from 0 to 4, where 0 represents normal speech, 1 mildly affected speech, 2 moderately impaired speech (still intelligible), 3 markedly impaired speech (difficult to understand), and 4 unintelligible speech.

Abbreviations: HC = healthy controls, MPT = maximum phonation time, MSA = multiple system atrophy, PD = Parkinson’s disease, PSP = progressive supranuclear palsy, SD = standard deviation.
FIGURE 1. Illustration of proposed acoustic analysis. Sample of the signal is plotted as loudness envelope (A). Statistical model of modal $F_0$ was initiated using cepstral analysis (B). Harmonic analysis of spectrum (C) was performed inside the sliding window to select candidates of modal fundamental frequency accordingly to the probability model of modal voice (D). The probability model of modal voice was updated using a Kalman filter (E). The decision about regularity or subharmonicity of the analysed interval was based on analysis of subharmonic-to-harmonic ratio (F). Graph (G) illustrates resulting track of modal fundamental frequency (solid red line) and track recalculated from the presence or absence of subharmonics (dashed black line). $F_0 = \text{fundamental frequency}$, $SD = \text{standard deviation}$, $SHR = \text{subharmonic to harmonic ratio}$.

FIGURE 2. Process diagram illustrating replication of the signal by supervised parameterization. Original signal (A) was manually parameterized (B). Series of impulses representing the position and amplitude of glottal pulses was generated (C). Sample of glottal pulses was convoluted with impulses to obtain glottal source (D). Coefficients of linear predictive coding were analysed on the original signal in a sliding window (E) and used for filtration of the glottal source and glottal noise (G). Reference signal comprised of filtered glottal source and filtered glottal noise in given harmonic-to-noise ratio (H). $F_0 = \text{fundamental frequency}$, FFT = Fast Fourier Transformation, HNR = harmonic-to-noise ratio, LPC = linear predictive coding, TKK = TKK aparat.
### TABLE II
COMPARISON OF DETECTION ERRORS FOR PUBLICLY AVAILABLE DETECTORS

<table>
<thead>
<tr>
<th>Detector</th>
<th>Reference of modal register</th>
<th>Reference regarding subharmonics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>SDE</td>
</tr>
<tr>
<td>BF0NLS</td>
<td>1.91</td>
<td>1.25</td>
</tr>
<tr>
<td>DIO</td>
<td>2.01</td>
<td>1.33</td>
</tr>
<tr>
<td>DYPSA</td>
<td>1.5</td>
<td>1.11</td>
</tr>
<tr>
<td>FXAC</td>
<td>7.97</td>
<td>5.55</td>
</tr>
<tr>
<td>FXANAL</td>
<td>4.21</td>
<td>2.02</td>
</tr>
<tr>
<td>FXCEP</td>
<td>7.66</td>
<td>5.05</td>
</tr>
<tr>
<td>HARVEST</td>
<td>1.05</td>
<td>0.88</td>
</tr>
<tr>
<td>MBSC</td>
<td>3.98</td>
<td>0.77</td>
</tr>
<tr>
<td>PEFAC</td>
<td>4.61</td>
<td>2.56</td>
</tr>
<tr>
<td>PRAAT_AC</td>
<td>2.07</td>
<td>1.57</td>
</tr>
<tr>
<td>PRAAT_CC</td>
<td>2.32</td>
<td>1.71</td>
</tr>
<tr>
<td>PRAAT_SHS</td>
<td>1.8</td>
<td>1.37</td>
</tr>
<tr>
<td>RAPT</td>
<td>2.33</td>
<td>1.87</td>
</tr>
<tr>
<td>FXRAPT</td>
<td>1.91</td>
<td>1.39</td>
</tr>
<tr>
<td>REAPER</td>
<td>1.04</td>
<td>0.82</td>
</tr>
<tr>
<td>SHRP T=0.1</td>
<td>2.04</td>
<td>1.66</td>
</tr>
<tr>
<td>SHRP (T=1)</td>
<td>4.25</td>
<td>2.29</td>
</tr>
<tr>
<td>SWIPE</td>
<td>1.52</td>
<td>1.28</td>
</tr>
<tr>
<td>YANGsaf</td>
<td>3.73</td>
<td>0.83</td>
</tr>
<tr>
<td>Presented</td>
<td>0.22</td>
<td>0.18</td>
</tr>
</tbody>
</table>

All scores are in semitones. $T =$ threshold of subharmonic correction in SHRP algorithm.
FIGURE 3. Detection accuracy of publicly available algorithms and the proposed method. Algorithms that showed lower RMSE in the tracking of modal $F_0$ are denoted by an asterisk. Presented errors of SHRP were measured using subharmonic correction with threshold = 0.1. RMSE = root mean square error, MAE = median error, SDE = standard deviation of error, ME = mean error.