

Received April 26, 2020, accepted May 13, 2020, date of publication May 19, 2020, date of current version June 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995737

Detection of Speech Impairments Using Cepstrum, Auditory Spectrogram and Wavelet Time Scattering Domain Features

ANDRIUS LAURAITIS¹, RYTIS MASKELIŪNAS², (Member, IEEE),
ROBERTAS DAMAŠEVIČIUS^{1,2,3}, (Member, IEEE), AND TOMAS KRILAVIČIUS^{2,4}

¹Department of Multimedia Engineering, Kaunas University of Technology, LT-44249 Kaunas, Lithuania

²Department of Applied Informatics, Vytautas Magnus University, LT-44404 Kaunas, Lithuania

³Faculty of Applied Mathematics, Silesian University of Technology, 44-100 Gliwice, Poland

⁴Baltic Institute of Advanced Technology, LT-01403 Vilnius, Lithuania

Corresponding author: Tomas Krilavičius (tomas.krilavicius@bpti.lt)

ABSTRACT We adopt Bidirectional Long Short-Term Memory (BiLSTM) neural network and Wavelet Scattering Transform with Support Vector Machine (WST-SVM) classifier for detecting speech impairments of patients at the early stage of central nervous system disorders (CNSD). The study includes 339 voice samples collected from 15 subjects: 7 patients with early stage CNSD (3 Huntington, 1 Parkinson, 1 cerebral palsy, 1 post stroke, 1 early dementia), other 8 subjects were healthy. Speech data is collected using voice recorder from Neural Impairment Test Suite (NITS) mobile app. Features are extracted from pitch contours, Mel-frequency cepstral coefficients (MFCC), Gammatone cepstral coefficients (GTCC), Gabor (analytic Morlet) wavelet and auditory spectrograms. 94.50% (BiLSTM) and 96.3% (WST-SVM) accuracy is achieved for solving healthy vs. impaired classification problem. The developed method can be applied for automated CNSD patient health state monitoring and clinical decision support systems as well as a part of Internet of Medical Things (IoMT).

INDEX TERMS Neural impairment, mobile app, deep learning, wavelet scattering, decision support, speech processing, digital health, Internet of Medical Things.


I. INTRODUCTION

Central nervous system disorders (CNSD) include Huntington Disease (HD), Parkinson Disease (PD), Alzheimer Disease (AD), mild cognitive impairment (MCI) and dementia. These diseases cover a broad range of symptoms, in particular, tremor (muscle stagnancy, body balance disorders, involuntary movements, etc.), cognitive (decision-making difficulties, behavioral disorders, attention problems, memory loss, etc.), speech (lack of pronounced words, use of shorter phrases, pauses) and energy expenditure (weight loss, negative energy balance) impairments [1].

Speech impairments are long known to be one of the most common symptoms in HD [2] and PD [3]. Although, HD and PD have many different symptoms, which are related only to that one specific disease, they present a similar set of deficits expressed in speech e.g. slow, weak, imprecise,

uncoordinated speech (dysarthria) [4], swallowing difficulties (dysphagia) [5], trouble sequencing the sounds in syllables and words (apraxia) [6], difficulty to express thoughts orally (aphasia) [7]. Such circumstances (also combined with cognitive impairments) lead to the need of specialized assessment and speech treatment for people with HD or PD. Usually, this treatment is provided by a speech-language pathologist (SLP) who checks for speech dysfunctions. SLP gives guidelines for maintaining safe swallowing, evaluates speech acceptance criteria i.e. pitch (degree of voice highness or lowness), loudness (ability for patient to project his own voice), articulation (ability to pronounce sounds), voice quality (ability to hold pitch properly), respiration (coordination of speech with breathing), resonance (quality of voice that is determined by the balance of sound vibration during speech), prosody (rhythm, stress and intonation during speaking) [5]–[7].

Current research in the computer science field focuses on replicating the analysis of SLP with assistive

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei .

devices [8]–[11], adapting heuristic algorithms [12], [13] and deep learning [14]–[16] for monitoring change in speech patterns, speech recognition and classification [17]–[20]. In addition, wavelet transforms (discrete, continuous, tunable-Q) are successfully utilized for speech impairment monitoring based on voice signal analysis [21], [22]. Here we focus on adopting bidirectional recurrent neural network (BiRNN) with long short-term memory (LSTM) [23], [24] and wavelet scattering transform–Gabor [25]) methods for solving healthy vs. impaired test subject classification problem based on speech signals. Our approach is based on the digitized data collection using the extended self-administered gerocognitive examination (SAGE) [26] methodology via non-invasive interface of a smart device (mobile phone or tablet) adapted for early stage patients with the CNSD disorders.

The structural organization of the paper is as follows. Section II analyses and compares the related work by showing limitations of existing solutions. Section III covers materials and methods used, i.e. the implementation and working principle of voice recorder task as part of proposed neural impairment screening software, audio file collection procedure, test subjects involved and definition and formularization of feature extraction methods for analysis of speech signal. Section IV describes two experiments for solving binary classification problem (healthy vs. impaired). Section V contains discussion, conclusion and future works.

II. RELATED WORK

There are many studies being conducted in detection of speech impairments in central nervous system disorder patients (CNSD). Gillivan-Murphy *et al.* [27] use voice recordings (collected in sound-treated laboratory with ambient noise measured at 50 dB level by using a AKG-C420 head mounted microphone) to detect speech tremors in PD. Acoustic analysis was performed with a Voice and Tremor Protocol (VTP), i.e., amplitude of voice, periodicity, rate, and magnitude of frequency signal features. Gaballah *et al.* [28] investigate subjective and objective assessment of the PD speech quality. The analyzed features are derived from the speech recordings (collected with 7 amplification devices) based on cepstral, spectral, and/or temporal parametrization (mel-frequency cepstral coefficients (MFCC) [29], gammatone frequency cepstral coefficients (GTCC) [30], discrete cosine transform (DCT) [31], speech-to-reverberation masking ratio (SRMR) [32], modulation area (ModA) [33], Low Complexity Quality Assessment (LCQA) [34]. Support vector regression (SVR), Gaussian process regression, machine learning methods and correlation analysis were used achieving an accuracy of 0.85.

Identification of acoustic and spectral features in PD is analyzed in [35] with data recording at 44.1 kHz, 16bits per sample by using the same microphone. MFCC, linear prediction coefficients (LPC) [36], discrete wavelet transform (DWT) [36], Gaussian mixture model (GMM) [37], time domain entropy (ET) [38], spectral entropy (ES) [39] features

were used for the evaluation (77.2% accuracy was achieved by using SVM with linear kernel).

Wu *et al.* in [40] target learning acoustic features (MFCC, spherical K-means, pooling method) to detect PD. All data was captured in a soundproof room and then resampled at a 16 kHz rate. Random Forest (RF) and SVM methods were used for the evaluation of detection accuracy (best achieved result is 96.37% with RF classifier). Perez *et al.* [41] differentiate between healthy controls and HD patients) based on acoustic and lexical features (MFCC, GMM, pause, speech rate, goodness of Pronunciation (GoP) [42]). The results were evaluated with k-Nearest Neighbours (k-NN) and Long-Short-Term Memory Recurrent Neural Networks (LSTM-RNN) algorithms (0.87 correlation).

Sakar *et al.* in [43] provide a comparative analysis of speech processing algorithms for PD recognition using detrended fluctuation analysis (DFA), pitch period entropy (PPE), recurrence period density entropy (RPDE), MFCCs, wavelet transform (WT) methods for feature extraction. The results were validated with a set of supervised classifiers (Logistic Regression, Multilayer Perceptron, Naive Bayes, Random Forest, SVMs with linear and RBF kernels, and k-NN algorithms) (0.86 best achieved correlation).

The classification of PD severity is introduced by Oung *et al.* in [44]. The data for speech signals were acquired by using a Sennheiser DW Pro2 headset positioned in 5 cm distance from the mouth of a subject. The researchers for feature extraction in speech adapted wavelet energy (WE), Shannon wavelet entropy (ShWE), Renyi wavelet entropy (ReWe), Tsallis wavelet entropy (TsWe), permutation entropy (Pe) and fuzzy entropy (Fe). The classifiers used were extreme learning machine (ELM), K-nearest neighbour (KNN), probabilistic neural network (PNN) and (best accuracy 91.11%).

Ali *et al.* [45] use Parkinson speech-based dataset from the UCI repository to investigate the classification of early diagnosis of PD. 15 acoustic features were considered in research: jitter, number of pulses, number of periods, mean period, standard deviation of period, number of voice breaks, degree of voice breaks, mean pitch, standard deviation, minimum pitch, autocorrelation, noise-to-harmonic ratio and harmonic-to-noise ratio. Four classifiers were examined: Bayes Net, Random Forests, Decision Stump and SVM (95.6% best accuracy).

The fusion of wavelet packet transform (WPT) and MFCC methods were applied for the diagnosis of PD from recorded speech signal by using Hidden Markov Models (HMM) and SVM classifiers [46]. Burk *et al.* in [47] analysed acoustic recordings (special software and hardware were used for the data collection) from PD patients based on the cepstral peak prominence (CPP) and aerodynamic measures of transglottal airflow (TAF) features in order to distinguish between speakers with no tremor and tremor (correlation 0.96).

Burk *et al.* [47] target vocal impairment detection for early prediction in PD. They applied MFCC and GMM for feature extraction. The data (96 kHz audio samples) was collected with a professional head mounted omnidirectional condenser

TABLE 1. Comparison of related work findings for speech impairment detection.

Work ref.	Target group	Dataset	Evaluation metrics (features)	Hardware	Software	Audio samples	Experiment supervision	Statistical analysis (methods)
[27]		30 PD and 28 healthy	Voice and Tremor Protocol (VTP), i.e., adapts rate, periodicity, magnitude of frequency and amplitude of voice signal features	AKG-C420 head-mounted microphone	SPSS	50 kHz (Kay data file)	No	regression analysis ($p < 0.05$) on acoustic features
[28]		10 PD and 10 healthy	MFCC, DCT, GTCC, speech-to-reverberation masking ratio (SRMR), modulation area (ModA), Low Complexity Quality Assessment (LCQA)	7 different amplification devices	Malcolm Slaney's auditory toolbox	16 bits, 16 kHz	Yes	correlation analysis (0.85)
[35]	PD	12 PD and 12 healthy	MFCC, linear prediction coefficients (LPC), discrete wavelet transform (DWT), Gaussian mixture model (GMM), time domain entropy (ET), spectral entropy (ES)	Microphone	Not provided	16 bits, 44 kHz	Yes	77.2% accuracy (SVM with linear kernel)
[36]		27 PD and 446 healthy	MFCC, spherical K-means, pooling method	Soundproof room (equipment details not provided)	Matlab (Voice Analysis Toolbox)	16 kHz	Yes	96.37% (Random Forest) classifier
[37]	HD	31 HD and 31 healthy	Acoustic and lexical features, MFCC, GMM, pause, speech rate, goodness of pronunciation (GoP)	Not provided	Computerized Language Analysis (CLAN)	Not provided	Yes	correlation analysis (0.87)
[43]		188 PD and 64 healthy	Recurrence period density entropy (RPDE), detrended fluctuation analysis (DFA), pitch period entropy (PPE), MFCCs, wavelet transform (WT)	Microphone	Praat	44.1 kHz	No	correlation analysis (0.86)
[44]		65 PD	Wavelet energy (WE), Shannon wavelet entropy (ShWE), entropy (Pe), fuzzy entropy (Fe)	Headset (Sennheiser DW Pro2)	Matlab, Simulink	16 bits, 44 kHz	Yes	91.11% (extreme learning machine)
[45]	PD	20 PD and 20 healthy (from UCI repository)	Period, number of voice breaks, degree of voice breaks, mean pitch, standard deviation, minimum pitch	Not provided	Audacity, Praat and Weka	NA	No	97.6% (Random Forest)
[46]		NA	Wavelet packet transform (WPT) and MFCC	Not provided	Not provided	NA	Not provided	not provided
[47]		34 PD and 11 healthy	Cepstral peak prominence (CPP) and aerodynamic measures	Head-mounted condenser microphone	SPSS, analysis of Dysphonia in Speech and Voice (ADSV)	144 kHz	Yes	correlation analysis (0.96)
[48]		75 PD and 54 healthy	MFCC and Gaussian Mixture Model (GMM)	Professional head mounted microphone	Praat	24 bits, 96 kHz	Yes	83%, bootstrap aggregation

microphone that was placed by 10 cm from the mouth of PD patient. The classification results were validated with bootstrap aggregation approach from log-likelihood on each frame (83% best accuracy).

Refer to Table 1 for comparison of related work to track speech impairments in PD and HD.

To sum up, speech impairments are very intensively analysed by the other computer scientists. The majority of related works involve the PD patients. However, very different approaches are adapted for the collection of voice recordings, i.e., most solutions require custom hardware (special microphones, amplification devices or headsets) and audio signal

processing software (Matlab, Praat, Audacity, SPSS) for test supervision.

In addition, the evaluation metrics (features) that are used for detecting speech impairments cover a wide range of choices, i.e., from acoustic features (jitter, number of pulses, voice breaks etc.), Gaussian Mixture Model (GMM), Mel-Frequency Cepstral Coefficients (MFCC), spectrum (kurtosis, spread, entropy etc.), wavelet transforms (WT) to strategies for combining these features.

Statistically, the proposed related work models for speech impairment detection were evaluated by using regression analysis (Spearman correlation coefficient = 0.0156) and

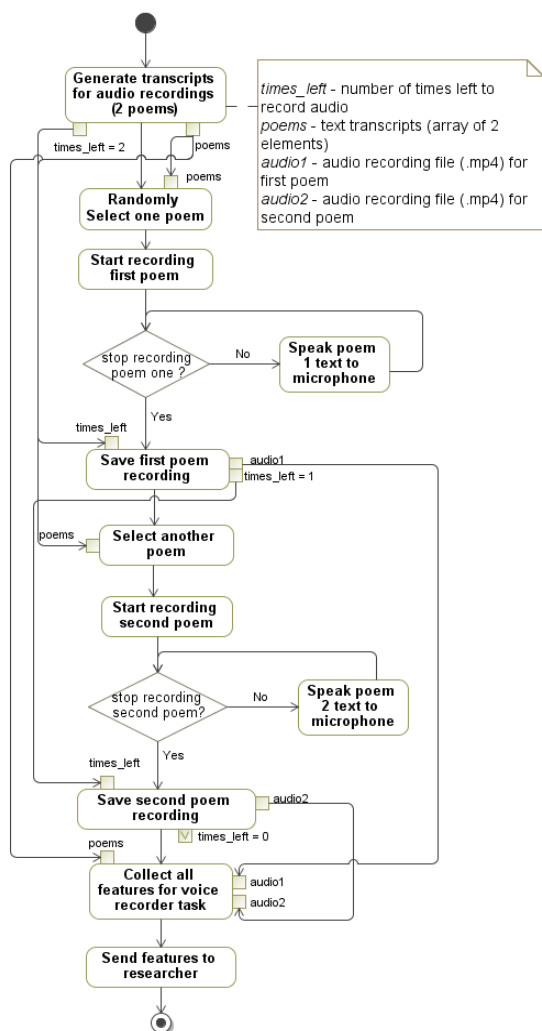


FIGURE 1. Algorithmic implementation of task T14: Voice Recorder.

classification methods (97.6% achieved with the Random Forest algorithm).

III. MATERIALS AND METHODS

A. TASK: VOICE RECORDER

The task is a part of Neural Impairment Test Suite (NITS) [49] system (mobile app) proposed by authors of this paper. NITS is a framework for collecting various features of data from tremor, cognitive, speech and energy expenditure tasks e.g. finger motion tracking, duration, distance evaluation of geometrical shapes, graph similarity evaluation, image collection from clock drawing (CDT) task, voice recordings, daily calorie balances, etc.) It is developed for Android OS with core supported software development kit (SDK), and includes third party libraries and custom algorithms for required functionality.

NITS supports different screen sizes and was tested with Lenovo YOGA YT3-X50L tablet (10.1” screen, with a resolution of 1280 x 800 px), SAMSUNG S7 smartphone (5.1”, 2560 x 1440 px) and OnePlus 5 (5.5”, 1920 x 1080 px).

Voice recorder task is named T14 in the NITS framework. Patient is instructed to read a short text of predefined poems into the mobile device microphone (Figure 1).

Predefined transcripts can be provided in English and Lithuanian languages if needed. The process is repeated two times, i.e., first, a poem is selected randomly; then, the remaining one is displayed. The recording begins when a patient is ready and presses the button ‘Start Recording’. Single poem recording finishes by pressing ‘Stop Recording’ button (a patient can make a pause if needed before the second recording). The T14 is executed two times as a precaution measure for more reliable test execution. In case a test subject (a CNSD patient) did not understand or follow the T14 task properly for the first time, a chance for repeating the procedure was given. Such approach allows collecting more voice recordings from each patient (damaged audio recordings were excluded), thus resulting in a larger dataset. After T14 is completed, two audio files (compressed .mpeg4 format, 44.1 kHz sample rate and AAC audio codec), together with the associated transcripts, are stored in the external storage of a mobile device. Defined parameters for audio files were chosen based on compatibility recommendations with the latest Android devices [50]. Audio codec MPEG-4 supports standard sampling rates from 8 to 48 kHz (mono or stereo channels). In addition, there is no significant effect in the quality of collected audio files for the analysis as all the recordings were collected with direct supervision of T14 execution by author of this paper. In such setup, isolating surrounding environment for audio data acquisition without external interference was ensured and distance from mobile device microphone to speaker’s mouth was adjusted accordingly.

B. TEST SUBJECTS, PROCEDURE, DATASET

A total number of 15 test subjects were involved in the audio file collection process. 7 patients with neurological disorders (3 Huntington (one of them juvenile of 18 years), 1 Parkinson, 1 cerebral palsy, 1 post stroke, 1 early dementia), other 8 were healthy subjects. Health state of neurological patients were in their early stage, e.g., the HD patients had the early (I or II) clinical form of HD according to the Shoulson-Fahn functional capacity rating scale [51]. All participants were asked to perform T14 task.

Dataset was collected during 5 rounds i.e. face-to-face patient visitations. All tests were supervised by author of this paper to explain working principle of T14. Moreover, such approach was chosen to ensure the fair execution of the test, i.e., without any cheating or faking the results. In some cases test subject were asked to perform T14 task multiple times, because CNSD patients tended to lose focus, thus resulting in interrupted audio recording process. In total, 339 samples (audio files) (including healthy and impaired test subjects) were collected in the dataset. The collected data was labelled using a healthy vs. impaired (0 or 1) objective assessment criteria for the health status of each subject, where 0 indicates that the subject is healthy, whereas 1 means that a subject

has a neurological disorder (e.g. Huntington Disease). In the process of data collection, the class label is specified in the mobile application before starting the actual testing procedure.

C. AUDIO SIGNAL FEATURE EXTRACTION METHODS

Stored .mpeg4 files (Figure 1, audio1 and audio2) are used as inputs for further audio signal processing. The authors considers the following methods for speech feature extraction.

The pitch for the calculation of the fundamental frequency (Hz) of an audio signal with sampling rate f_s can be 44.1 kHz. The pitch function estimates the fundamental frequency as determined by WindowLength (default size is $round(0.052 \cdot f_s)$) and OverlapLength (default size is $round(0.042 \cdot f_s)$) name-value pairs, WindowLength < OverlapLength. Pitch contours (WindowLength and OverlapLength ranges) can be estimated by 5 methods: Normalized Correlation Function (NCF) [52], Pitch Estimation Filter (PEF) [53], Cepstrum Pitch Determination (CEP) [54], Log-Harmonic Summation (LHS) [55], Summation of Residual Harmonics (SRH) [56].

The PEF method models signal Y at time t in the spectral domain with frequency f as defined in formula (1):

$$Y_t(f) = \sum_{k=1}^K a_{k,t} \delta(f_s - kf) + N_t(f) \quad (1)$$

here K is the number of peaks in the audio signal, $N_t(f)$ is the power spectral density of unwanted noise, $a_{k,t}$ is the power of the k -th harmonic at time t .

In the LHS method, the signal is modelled by (2):

$$H(s) = \sum_{n=1}^N h_n P(s + \log_2 nc) \quad (2)$$

here nc - compression factor, $s = \log_2 f$, $h_{nc} = 0.84^{nc-1}$ is a decreasing sequence implying that higher harmonics contribute less to the pitch than lower harmonics to the noise, $P(s) = W(s) \cdot A(s)$, $W(s)$ - spectral window function, $A(s)$ - logarithmic frequency abscissa, $N = 15$ (the number of harmonics considered).

The SRH method tracks the pitch by using calculations in (3) formula. Please consider the provided references for extra information of NCF and CEP methods.

$$SRH(f) = E(f) + \sum_{k=2}^N \left[E(k \cdot f) - E\left(\left(k - \frac{1}{2}\right) \cdot f\right) \right] \quad (3)$$

here $E(f)$ - amplitude spectrum signal (f - frequency in the range of $[F_{min}, F_{max}]$, computed for each Hanning-windowed frame, covering several cycles of the resulting residual signal) of the k -th harmonic, N - number of harmonics that are taken into account.

The additional considered method is Mel-frequency cepstral coefficients (MFCC) [29]. MFCC returns the coefficients sampled at a frequency of f_s as well as the change in coefficients (delta) and the change in delta values deltaDelta). WindowLength and OverlapLength default configuration setup is the same as in the Pitch method.

MFCC computes a frequency analysis based on a filter bank. A short-time Fourier analysis results in a discrete Fourier transform (DFT) for signal $X_t[k]$ in time t . DFT values are grouped together in critical bands and weighted by a triangular function. The (4), (5) and (6) formulas are used for MFCC calculations ($R = 22$, m -th signal sample, the number of MFCC coefficients is usually 13):

$$MF_t[r] = \frac{1}{A_r} \sum_{k=L_r}^{U_r} |V_r[k] \cdot X_t[k]|^2 \quad (4)$$

here $MF_t[r]$ - Mel-frequency spectrum at analysis time t for $r = 1, 2, \dots, R$. $V_r[k]$ is the triangular weighting function for the r -th filter, ranging from DFT index L_r to U_r .

$$A_r = \sum_{k=L_r}^{U_r} |V_r[k]|^2 \quad (5)$$

A_r - is a normalizing factor for the r -th Mel-filter.

$$mfcc_t[m] = \frac{1}{R} \sum_{r=1}^R \log(MF_t[r]) \cdot \cos\left[\frac{2\pi}{R} \left(r + \frac{1}{2}\right) m\right] \quad (6)$$

Having calculated MFCC as defined in (6), it uses the least-squares approximation of the local slope over a region around the current time sample method to determine delta (passing MFCC) and deltaDelta (passing delta). The same rule applies for GTCC.

Similarly, as MFCC, Gammatone cepstral coefficients (GTCC), including delta and deltaDelta, can be used for audio signal feature extraction. GTCC is a bio-inspired adaptation of the MFCC that employs Gammatone (GT) filters [30]. The GT filter with its properties is defined in formula (7):

$$g(t) = K t^{(n-1)} e^{-2\pi B t} \cos(2\pi f_c t + \varphi), \quad t > 0 \quad (7)$$

here n is the filter order, K is the amplitude factor, B is impulse response, f_c is the central frequency, φ is phase shift.

In GTCC extraction, the audio signal is first sliced into short frames, usually about 10–50 ms (same as in MFCC). This allows signal to remain stationary, thus allowing for the signal analysis. Afterwards, GT filter bank is applied to the signal's FFT, highlighting the perceptually meaningful voice frequencies. Lastly, DCT is applied to model the human perception of sound and to decorrelate the filter outputs, therefore achieving better energy compaction:

$$gtcc_t[m] = \sqrt{\frac{2}{R}} \sum_{r=1}^R \log(X_t[r]) \cdot \cos\left[\frac{\pi r}{R} \left(m - \frac{1}{2}\right) m\right] \quad (8)$$

here R is the number of GT filters, $X_t[r]$ is the energy of the signal in the r -th spectral band, $1 \leq m \leq M(5)$ is the number of outputs.

Another considered method for speech feature extraction is called wavelet scattering transform (WST). It defines a representation which is resistant to time-warping deformations. WST extends MFCC by calculating modulation spectrum coefficients through wavelet convolutions and modulus operators [57]. In addition, WST overcomes MFCC in audio

representations for the classification problems at time scales more than 25 ms.

Scattering transform restores the information lost by a Mel-frequency averaging by employing a cascade of wavelet decompositions and modulus operators. The constant-Q filter banks calculate a wavelet transform. A wavelet $\varphi(t)$ is band-pass filter with $\check{\varphi}(0) = 0$ and is written in the centre frequency ω form as defined in formula (9):

$$\varphi_{\omega}(t) = \omega \cdot \varphi(\omega t), \quad \check{\varphi}_{\omega}(s) = \check{\varphi}\left(\frac{s}{\omega}\right) \quad (9)$$

Here the centre frequency of $\check{\varphi}$ is normalized to 1. $\omega = 2^{k/Q}$, Q are the wavelets per octave, $k \in \mathbb{Z}$. $\check{\varphi}$ is of the order of Q^{-1} .

Finally, 10 methods for feature extraction in audio signals and auditory spectrograms are defined. These are spectralSlope, spectralSkewness, spectralSpread, spectralCentroid, spectralDecrease, spectralKurtosis [58], spectralFlux & spectralRolloff [59], spectralFlatness [60], spectralEntropy [39]. Term ‘bin’, referred in (10) – (19) equations is a segment, e.g., [fl, fh] of the frequency axis that collect the amplitude, magnitude or energy from a small range of frequencies.

SpectralSlope evaluates the spectral shape slope by using a linear approximation of the magnitude spectrum. A linear function is modelled from the magnitude spectrum as defined in (10) equation.

$$slope = \left(\sum_{k=b_1}^{b_2} (f_k - \mu_f)(s_k - \mu_s) \right) / \left(\sum_{k=b_1}^{b_2} (f_k - \mu_f)^2 \right) \quad (10)$$

here f_k - is the frequency in Hz corresponding to bin k , μ_f is the mean frequency, s_k is the spectral value at bin k , μ_s is the mean spectral value, b_1 and b_2 are the band edges, in bins, over which to calculate the spectral method (e.g., slope), μ_f is the spectral centroid, Spectral skewness evaluates the symmetry of the spectral magnitude distribution around their arithmetic mean (11).

$$skewness = \left(\sum_{k=b_1}^{b_2} (f_k - \mu_1)^3 s_k \right) / (\mu_2)^3 \left(\sum_{k=b_1}^{b_2} s_k \right) \quad (11)$$

here μ_2 is the spectral spread.

Spectral spread measures the concentration of the power spectrum around the spectral centroid (Eq. 12).

$$spread = \sqrt{\left(\left(\sum_{k=b_1}^{b_2} (f_k - \mu_1)^2 s_k \right) / \left(\sum_{k=b_1}^{b_2} s_k \right) \right)} \quad (12)$$

Spectral centroid represents the centre of gravity (COG) of spectral energy. It is defined as the frequency-weighted sum of the power spectrum normalized by its unweighted sum (13).

$$centroid = \left(\sum_{k=b_1}^{b_2} (f_k s_k) \right) / \left(\sum_{k=b_1}^{b_2} s_k \right) \quad (13)$$

Spectral decrease assesses the steepness of the decrease of the spectral envelope. The result of the spectral decrease is a value less than 1. The spectral decrease is not defined for audio blocks with no spectral energy (silence) (Eq. 14).

$$decrease = \left(\sum_{k=b_1+1}^{b_2} \frac{s_k - s_{b_1}}{k - 1} \right) / \left(\sum_{k=b_1+1}^{b_2} s_k \right) \quad (14)$$

Spectral kurtosis evaluates the shape of the spectral magnitude value distribution as compared to the Gaussian distribution (Eq. 15).

$$kurtosis = \left(\sum_{k=b_1+1}^{b_2} \frac{s_k - s_{b_1}}{k - 1} \right) / \left(\sum_{k=b_1+1}^{b_2} s_k \right) \quad (15)$$

Spectral flux is the change of the spectral shape calculated as the mean difference between neighboring Short Time Fourier Transform (STFT) frames (Eq. 16).

$$flux(t) = \left(\sum_{k=b_1}^{b_2} |s_k(t) - s_k(t-1)|^p \right)^{\frac{1}{p}} \quad (16)$$

Spectral rolloff is a measure of the bandwidth of the analyzed block n of audio samples and is specified as the bin of frequency below which the cumulative magnitudes of the STFT reach a certain percentage K of the overall sum of magnitudes (Eq. 17).

$$rolloff(i) = \sum_{k=b_1}^i s_k = K \sum_{k=b_1}^{b_2} s_k \quad (17)$$

Spectral flatness is the ratio of geometric and arithmetic means of the magnitude spectrum (Eq. 18).

$$flatness = \left(\prod_{k=b_1}^{b_2} s_k \right)^{\frac{1}{b_2-b_1}} / \left(\frac{1}{b_2-b_1} \left(\sum_{k=b_1}^{b_2} s_k \right) \right) \quad (18)$$

Entropy evaluates the ‘‘peakiness’’ of a probability mass function (PMF) as follows: (Eq. 19).

$$entropy = \left(- \sum_{k=b_1}^{b_2} s_k \log(s_k) \right) / \log(b_2 - b_1) \quad (19)$$

IV. EXPERIMENTAL RESULTS

Two supervised learning approaches (wavelets with SVM and deep learning neural networks) are considered in experimental research for classifying test subjects into health and impaired instances (2 target classes): 1) Wavelet scattering transform (WST) with SVM; and 2) Bidirectional recurrent neural network (RNN) with Long short-term memory (BiLSTM). Both methods apply percentage split resampling technique for the original data i.e. 70% training set and 30% testing set. For the collected dataset of voice recordings, this corresponds to 207 samples for training and 88 for testing, including 29 samples (healthy test subjects) and 15 samples (impaired test subjects) for predictions on new and unseen data.

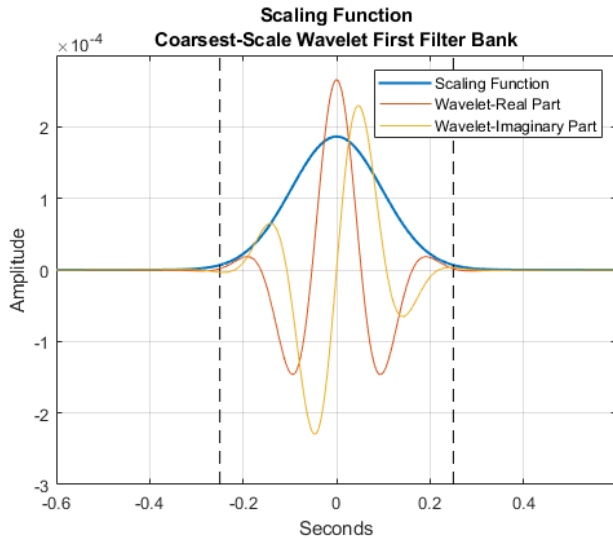


FIGURE 2. Filter bank for scattering transform in WST (invariance is 0.5 seconds).

A. SPEECH IMPAIRMENT DETECTION WITH WST AND SVM

Experiment is based based on voice recordings, collected from T14 task. WST method applies Gabor (analytic Morlet) wavelet. Such wavelets use low pass scaling function to produce low-variance representations of voice.

Wavelet is designed as follows. The signal length is a natural logarithm value of 2^{19} . For WST configuration, only 3 parameters are provided: the duration of the time invariance, the number of wavelet filter banks (band-pass filters that separate voice data into multiple components, each one carrying a sub-band of the original data) and the number of wavelets per octave. Two wavelet filter banks are used: first (fb1) and second (fb2). The first filter bank has 8 wavelets per octave, and the second filter bank has 1 wavelet per octave. The time invariance scale is set to 0.5 seconds. For such setup, invariance scale parameter that is plotted on the coarsest scale [61] (Figure 2) does not exceed the invariant scale of the wavelet scattering decomposition, i.e., is indicator of low variance.

The plot of fb1 and fb2 filter banks using Littlewood-Paley of sums [62] representation is provided in Figure 3.

The audio materials are transferred to a single object in memory ads. Train (Ttrain) and test (Ttest) data are converted to tall arrays. Then, scattering train features (scatteringTrain) and scattering test features (scatteringTest) are created by applying log transformation of each audio file and subsamples, the number of scattering windows by 8. The scattering features are combined together to a matrix by using MATLAB Parallel Pool (Number of Workers = 4) on a single GPU, resulting in the training features and the testing features (each row of the matrix is 1 time window across the $N = 341$ paths in the scattering transform of each audio signal).

The training features and the testing features are used to fit the data for support vector machine (SVM) model with

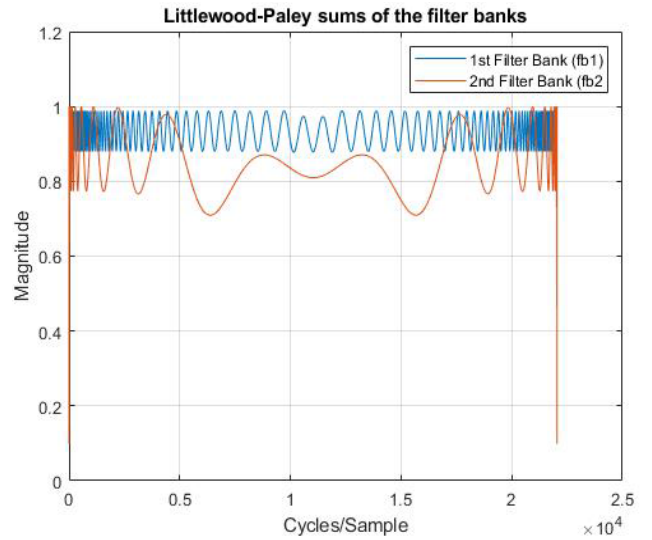


FIGURE 3. Littlewood-Paley sums of 1-st and 2-nd filter banks of wavelet.



FIGURE 4. WST accuracy of health vs impaired classification.

polynomial kernel (order = 3). SVM tuning was applied by using the Majority Vote method, which achieved 96.3% accuracy of the supplied test data, as shown in confusion matrix (Figure 4). The model build time is 369.83 seconds.

B. SPEECH IMPAIRMENT DETECTION WITH BiLSTM

Similarly, as in the WST approach, this deep learning experiment also analyses voice recordings collected from the T14 task. First stage is the pre-processing of the original audio signal i.e. removing silence segments. To eliminate not useful information that is pertaining to the health status indicator of the speaker, the isolation of the speech segment method is applied. This method uses the thresholding approach. First, 2 features (signalEnergy, centroid) over non-overlapping frames of the audio data are calculated. Next, the energy and spectral centroid for each frame is evaluated;

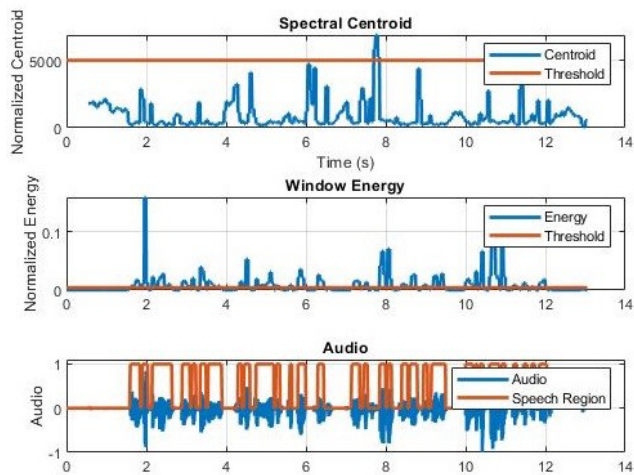


FIGURE 5. Detection of active voice regions in speech segment.

centroid threshold ($T_C = 5000$ Hz) and energy threshold (T_E) are calculated afterwards. The speech regions where the feature values fall below or above their respective thresholds are disregarded (Figure 5). On the contrast, the speech region is active in cases as shown in (20) equation:

$$isSpeechReg = signalEnergy \geq T_E, centroid \leq T_C \quad (20)$$

In the implementation, *isSpeechRegion* is further characterized by *regionStartPos* (indices of frames where a speech-to-silence or silence-to-speech transition occurs), *regionLengths* (length of all-silence or all-speech regions), start and end indices (SI, EI) for each speech region. Once the active speech regions are detected, the intersecting speech segments are merged and fed for the feature extraction mechanism (segments).

The speech signal changes over time, but is stationary on short time scales; thus, their processing is often done in windows of 20–40 ms. For each speech segment, a periodic hamming window [63] with 80% overlap is used and then concatenated into sequences (each vector contains 92 features, each sequence 40 feature vectors). The features used are GTCC, MFCC, pitch, slope, skewness, spread, flux, rolloff, decrease, flatness, kurtosis and entropy. These 12 features are concatenated together and can be combined in numerous ways. A feature can be removed from the sequence or swapped with other if necessary.

With this configuration setup, the next step is feature transferring to tall array T (this provides a way to work with data backed by an audio data store (*audioDataStore*) that can have millions or billions of rows) on the GPU. These feature sequences are re-evaluated (*featureSequences*) and normalized (mean and standard deviation for each coefficient is computed). Such normalized GPU features are ready to be supplied for training Bidirectional Long Short-Term Memory (BiLSTM) deep learning neural network. LSTM can learn long-term dependencies between time steps of sequence data (forward and backward directions). In the training process,

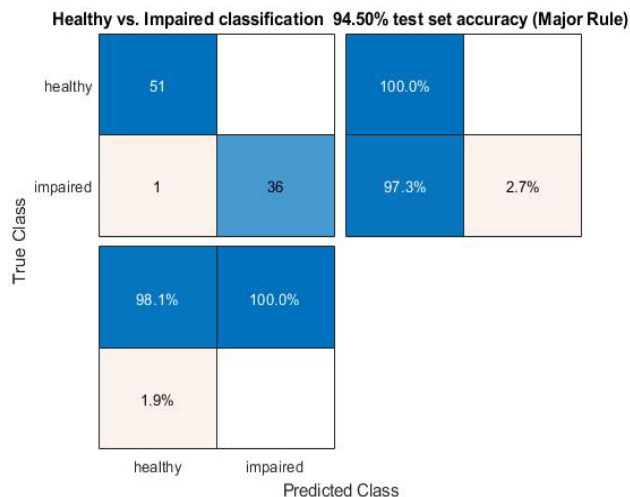


FIGURE 6. BiLSTM accuracy on test set is 94.50%.

MATLAB parallel pool is invoked (Number of Workers used is 4) for processing features in tall array faster. BiLSTM training adapts epoch – based approach. Configuration setup for BiLSTM training: algorithm is RMSProp (root mean square propagation) optimizer, MaxEpochs = 10, MiniBatchSize = 128, shuffle on every epoch, learning rate drop factor = 0.1, learning rate schedule ‘piecewise’. A convergence of BiLSTM network was observed in 320-th iteration (10-th epoch) with 100% of mini-batch accuracy with 0.0035 loss and 0.0001 base learning rate. BiLSTM model build time is 191.68 seconds.

The architecture of proposed BiLSTM neural network has 2 fully connected layers of 100 neurons, followed by a softmax layer and a classification (output) layer.

Figure 6 illustrates healthy vs. impaired classification results on the provided dataset by applying Major Vote method (rule) for tuning classifier performance i.e. overall model accuracy of 94.50%.

C. SUMMARY OF EXPERIMENTS

Both experiments are implemented using MATLAB Audio Toolbox R2019a (Mathworks Inc., USA) software. Audio materials are transferred to MATLAB *audioDataStore* object.

For the WST-SVM experiment, initial preparation steps include the creation of a root folder and two sub folders naming ‘healthy’ and ‘impaired’ correspondingly. The names of the subfolders should match the names of the output target classes. The audio files must be provided as 1411 kbps sample rate .wav audio files at 22050 Hz sample rate.

For the BiLSTM experiment, the procedure starts by creating a root folder and two subfolders, naming ‘train’ and ‘test’ correspondingly. Initial audio materials must be provided as 64 kbps sample rate files (.mp3 format). Two .csv files (one for training set, another for testing set) are prepared for storing the summarized information about the collected files by using this format: linkage to the stored audio file in disk

TABLE 2. Requirements summary for experimental setup.

Experiment	Audio File Format	Software	Hardware
WST	1411 kbps sample rate .wav audio files	MATLAB Audio Toolbox R2019a (requires DSP System Toolbox, Signal Processing Toolbox)	CUDA-enabled NVIDIA GPUs (3.0 or higher)
	64 kbps sample rate .mp3 format audio files	Microsoft Visual C++ 2015 or newer compiler. Windows, Mac, Linux. operating system:	
BiLSTM			

and sick or impaired indicator (as text string). In addition, .csv file structure can be expanded with transcript of read poem, recording duration.

Table 2 shows the data format, software and hardware requirements for the proposed WST and BiLSTM based PD classification methods.

V. DISCUSSION AND CONCLUSIONS

In this paper, we presented an investigation for detecting speech impairments occurring to the CNDS patients. A dataset of audio files (including early stage CNDS patients and healthy subjects) was collected during a pilot study carried out in Lithuania with the usage of a smart noninvasive interface, i.e. Neural Impairment Test Suite mobile app. For proper task execution, test subject should be acquainted with Lithuanian or English languages (speech dialect is not important).

Three domains of feature extraction methods (and their combinations) from audio signals were considered in this research: cepstrum domain (pitch contours, MFCC, GTCC), auditory spectrograms (slope, skewness, spread, centroid, decrease, kurtosis, flux, rolloff, entropy, flatness) and WST (wavelet time scattering, analytic Gabor). BiLSTM and support vector machine (SVM) with polynomial kernel methods were adapted for classifying target test subjects into healthy and impaired groups. WST-SVM achieved 96.3% accuracy and BiLSTM 94.50% accuracy on test set, thus showing strong expectations for decision support in speech impairment detection in targeting related diseases (e.g., Alzheimer's) in various progression stage. WST-SVM excels over BiLSTM considering the related research findings that collected voice recording from CNDS patients were significantly long, i.e., up to 47 sec (observed from the juvenile HD patient).

The proposed speech detection models can be compared with works of other researchers in competitive study. Tsanas *et al.* targeted identification of PD based on vocal performance (SVM classifier, 90% accuracy) [64]. Caesarendra *et al.* analysed pattern recognition with voice features in PD stage classification (SVM, 79.17% accuracy) [65]. Hauptman *et al.* adopted SVM (77.20 %) for identification of distinctive acoustic and spectral features in PD [35]. Moreover, Extreme learning machine

(ELM, 91.11% accuracy) approach for the classification of PD severity was introduced by Oung *et al.* [44], and Jeancolas *et al.* [48] adapted Bootstrap aggregation classifier (83% accuracy) for sound classification of Parkinsonism.

Speech disorders in HD and PD tend to progress over time, so proposed classification methods could function as a decision support system for monitoring the health state of the CNDS patients and provide insight about disease status. The designed WST-SVM and BiLSTM models are integrated into the NITS mobile app for triggering screening alert to a CNDS patient about his deterioration of speech impairment before such symptoms become much worse. The developed models also can be used as a service in the context of Internet of Health Things (IoHT) [66] ecosystem of services and devices.

ACKNOWLEDGMENT

The authors would like to thank the President of Lithuania Huntington disease association Dr. Z. Navikiene for help carry out the experiments described in this article as well as for her practical support and advices.

HUMAN STUDIES

Research was approved by an Institutional Review Board of the Faculty of Informatics, Kaunas University of Technology.

FUNDING

This research received no external funding.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] P. H. H. Nguyen and A. M. Cenci, Eds., *Behavioral Neurobiology of Huntington's Disease and Parkinson's Disease* (Current Topics in Behavioral Neurosciences), vol. 22. Springer, 2015, pp. 3–29.
- [2] P. Podoll, P. Caspary, H. W. Lange, and J. Noth, "Language functions in Huntington's disease," *Brain*, vol. 111, no. 6, pp. 1475–1503, 1988.
- [3] T. Thies, D. Muecke, A. Lowit, E. Kalbe, J. Steffen, and M. Barbe, "Cognitive skills and prominence production: Highlighting prominent elements in the speech of patients with Parkinson's disease," in *Proc. Hanyang Int. Symp. Phonetics Cogn. Sci. Lang.*, 2019, pp. 1–3.
- [4] C. L. Ludlow, N. P. Connor, and C. J. Bassich, "Speech timing in Parkinson's and Huntington's disease," *Brain Lang.*, vol. 32, no. 2, pp. 195–214, 1987, doi: [10.1016/0093-934x\(87\)90124-6](https://doi.org/10.1016/0093-934x(87)90124-6).
- [5] S. Polychronis, G. Dervenoulas, T. Yousof, F. Niccolini, G. Pagano, and M. Politi, "Swallowing and chewing difficulties are associated with presynaptic dopaminergic dysfunction and greater non-motor symptom burden in early drug-naïve Parkinson's patients," *BioRxiv*, Mar. 2019, Art. no. 577148, doi: [10.1101/577148](https://doi.org/10.1101/577148).
- [6] K. M. Heilman, "Apraxic and action-intentional disorders associated with vascular and degenerative dementing diseases," in *Vascular Disease, Alzheimer's Disease, and Mild Cognitive Impairment: Advancing an Integrated Approach*. London, U.K.: Oxford Univ. Press, 2020, p. 146.
- [7] O. Sinanović, "Psychiatric disorders in neurological diseases," *Mind and Brain*. Cham, Switzerland: Springer, 2020, pp. 65–79.
- [8] C. R. Pereira, D. R. Pereira, S. A. T. Weber, C. Hook, V. H. C. de Albuquerque, and J. P. Papa, "A survey on computer-assisted Parkinson's disease diagnosis," *Artif. Intell. Med.*, vol. 95, pp. 48–63, Apr. 2019, doi: [10.1016/j.artmed.2018.08.007](https://doi.org/10.1016/j.artmed.2018.08.007).
- [9] S. A. S. Lee, "Virtual speech-language therapy for individuals with communication disorders: Current evidence, limitations, and benefits," *Current Develop. Disorders Rep.*, vol. 6, no. 3, pp. 119–125, Sep. 2019.

- [10] A. Lauraitis, R. Maskeliunas, R. Damaševičius, D. Polap, and M. Wozniak, "A smartphone application for automated decision support in cognitive task based evaluation of central nervous system motor disorders," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 1865–1876, Sep. 2019, doi: [10.1109/JBHI.2019.2891729](https://doi.org/10.1109/JBHI.2019.2891729).
- [11] F. Marxreiter, U. Buttler, H. Gassner, F. Gandor, T. Gladow, B. Eskofier, and J. Klucken, "The use of digital technology and media in German Parkinson's disease patients," *J. Parkinson's Disease*, vol. 10, no. 2, pp. 717–727, 2020, doi: [10.3233/JPD-191698](https://doi.org/10.3233/JPD-191698).
- [12] D. Polap, M. Woźniak, R. Damaševičius, and R. Maskeliūnas, "Bio-inspired voice evaluation mechanism," *Appl. Soft Comput.*, vol. 80, pp. 342–357, Jul. 2019, doi: [10.1016/j.asoc.2019.04.006](https://doi.org/10.1016/j.asoc.2019.04.006).
- [13] D. Polap and M. Wozniak, "Voice recognition by neuro-heuristic method," *Tsinghua Sci. Technol.*, vol. 24, no. 1, pp. 9–17, Feb. 2019, doi: [10.26599/TST.2018.9010066](https://doi.org/10.26599/TST.2018.9010066).
- [14] V. Illner, P. Sovka, and J. Ruzs, "Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease," *Biomed. Signal Process. Control*, vol. 58, Apr. 2020, Art. no. 101831, doi: [10.1016/j.bspc.2019.101831](https://doi.org/10.1016/j.bspc.2019.101831).
- [15] J. S. Almeida, P. P. R. Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, and V. H. C. de Albuquerque, "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognit. Lett.*, vol. 125, pp. 55–62, Jul. 2019, doi: [10.1016/j.patrec.2019.04.005](https://doi.org/10.1016/j.patrec.2019.04.005).
- [16] H. Gunduz, "Deep learning-based Parkinson's disease classification using vocal feature sets," *IEEE Access*, vol. 7, pp. 115540–115551, 2019.
- [17] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.
- [18] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, Apr. 2019.
- [19] A. Sabzi Shahrebabaki, A. S. Imran, N. Olfati, and T. Svendsen, "A comparative study of deep learning techniques on frame-level speech data classification," *Circuits, Syst., Signal Process.*, vol. 38, no. 8, pp. 3501–3520, Aug. 2019.
- [20] N. P. Narendra and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Commun.*, vol. 110, pp. 47–55, Jul. 2019.
- [21] Z. Soumaya, B. D. Taoufiq, B. Nsiri, and A. Abdelkrim, "Diagnosis of parkinson disease using the wavelet transform and MFCC and SVM classifier," in *Proc. 4th World Conf. Complex Syst. (WCCS)*, Apr. 2019.
- [22] M. Wozniak and D. Polap, "The use of wavelet transformation in conjunction with a heuristic algorithm as a tool for feature extraction from signals," *Inf. Technol. Control*, vol. 46, no. 3, pp. 372–381, Sep. 2017, doi: [10.5755/j01.itc.46.3.17582](https://doi.org/10.5755/j01.itc.46.3.17582).
- [23] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
- [24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [25] T. Sing Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959–971, Oct. 1996, doi: [10.1109/34.541406](https://doi.org/10.1109/34.541406).
- [26] D. W. Scharre, S.-I. Chang, R. A. Murden, J. Lamb, D. Q. Beversdorf, M. Katakai, H. N. Nagaraja, R. A. Bornstein, "Self-administered gerocognitive examination (SAGE) administration and scoring instructions," *Alzheimer Disease Associated Disorders*, vol. 4, no. 1, pp. 64–71, 2010.
- [27] P. Gillivan-Murphy, N. Miller, and P. Carding, "Voice tremor in Parkinson's disease: An acoustic study," *J. Voice*, vol. 33, no. 4, pp. 526–535, Jul. 2019.
- [28] A. Gaballah, V. Parsa, M. Andreetta, and S. Adams, "Objective and subjective speech quality assessment of amplification devices for patients with Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1226–1235, Jun. 2019.
- [29] L. R. Rabiner and R. W. Schafer, *Theory and Application of Digital Signal Processing*. London, U.K.: Pearson, 2010.
- [30] X. Valero and F. Alias, "Gammamaton cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012.
- [31] P. K. Ajmera, D. V. Jadhav, and R. S. Holambe, "Text-independent speaker identification using radon and discrete cosine transforms based features from speech spectrogram," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2749–2759, Oct. 2011, doi: [10.1016/j.patcog.2011.04.009](https://doi.org/10.1016/j.patcog.2011.04.009).
- [32] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010. [Online]. Available: <http://ieeexplore.ieee.org/document/5547575>
- [33] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomed. Signal Process. Control*, vol. 8, no. 3, pp. 311–314, May 2013.
- [34] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity nonintrusive speech quality assessment," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.
- [35] Y. Hauptman, R. Aloni-Lavi, I. Lapidot, T. Gurevich, Y. Manor, S. Naor, N. Diamant, and I. Opher, "Identifying distinctive acoustic and spectral features in Parkinson's disease," in *Proc. Interspeech*, Sep. 2019, pp. 2498–2502.
- [36] N. S. Nehe and R. S. Holambe, "DWT and LPC based feature extraction methods for isolated word recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2012, no. 1, p. 7, Jan. 2012.
- [37] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995, doi: [10.1109/89.365379](https://doi.org/10.1109/89.365379).
- [38] P. A. Varotsos, N. V. Sarlis, E. S. Skordas, and M. S. Lazaridou, "Entropy in the natural time domain," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 1, Jul. 2004, Art. no. 011106.
- [39] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 2004, p. I-193.
- [40] K. Wu, D. Zhang, G. Lu, and Z. Guo, "Learning acoustic features to detect Parkinson's disease," *Neurocomputing*, vol. 318, pp. 102–108, Nov. 2018.
- [41] M. Perez, W. Jin, D. Le, N. Carozzi, P. Dayalu, A. Roberts, and E. M. Provost, "Classification of huntington disease using acoustic and lexical features," in *Proc. Interspeech*, Sep. 2018, pp. 1898–1902.
- [42] D. Le, K. Licata, C. Persad, and E. M. Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2187–2199, Nov. 2016.
- [43] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul, and H. Apaydin, "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, Jan. 2019.
- [44] Q. W. Oung, H. Muthusamy, S. N. Basah, H. Lee, and V. Vijejan, "Empirical wavelet transform based features for classification of Parkinson's disease severity," *J. Med. Syst.*, vol. 42, no. 2, p. 29, Feb. 2018.
- [45] H. Ali, S. M. Adnan, S. Aziz, W. Ahmad, and M. Obaidullah, "Sound classification of Parkinsonism for telediagnosis," *Tech. J.*, vol. 24, no. 1, pp. 90–97, 2019.
- [46] H. Kuresan, D. Samiappan, and S. Masunda, "Fusion of WPT and MFCC feature extraction in Parkinson's disease diagnosis," *Technol. Health Care*, vol. 27, no. 4, pp. 363–372, Jul. 2019, doi: [10.3233/THC-181306](https://doi.org/10.3233/THC-181306).
- [47] B. R. Burk and C. R. Watts, "The effect of parkinson disease tremor phenotype on cepstral peak prominence and transglottal airflow in vowels and speech," *J. Voice*, vol. 33, no. 4, pp. 580.e11–580.e19, Jul. 2019.
- [48] L. Jeancolas, G. Mangone, J.-C. Corvol, M. Vidailhet, S. Lehéry, B.-E. Benkelfat, H. Benali, and D. Petrovska-Delacrétaz, "Comparison of telephone recordings and professional microphone recordings for early detection of Parkinson's disease, using mel-frequency cepstral coefficients with Gaussian mixture models," in *Proc. Interspeech*, Sep. 2019, pp. 3033–3037.
- [49] A. Lauraitis. (2018). *Neural Impairment Test Suite*. [Online]. Available: https://play.google.com/store/apps/details?id=com.alauraitis.test_suite
- [50] *Android 10 Compatibility Definition. Audio Codecs Detail*. Accessed: Apr. 26, 2020. [Online]. Available: https://source.android.com/compatibility/android-cdd#5_multimedia_compatibility
- [51] I. Shoulson and S. Fahn, "Huntington disease: Clinical care and evaluation," *Neurology*, vol. 29, pp. 1–3, Jan. 1979.
- [52] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, no. 6B, pp. 1687–1697, 1972.
- [53] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. 19th Eur. Signal Process. Conf.*, Aug./Sep. 2011, pp. 451–455.
- [54] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, Feb. 1967.

- [55] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, Jan. 1988.
- [56] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Aug. 2011, pp. 1973–1976.
- [57] J. Anden and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.
- [58] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," IRCAM, Paris, France, Tech. Rep. 54, 2004.
- [59] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Apr. 1997, pp. 1221–1224.
- [60] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Sel. Areas Commun.*, vol. SAC-6, no. 2, pp. 314–323, 2nd Quart., 1988.
- [61] B. Farell, S. Li, and S. P. McKee, "Coarse scales, fine scales, and their interactions in stereo vision," *J. Vis.*, vol. 4, no. 6, p. 8, Jun. 2004.
- [62] G. I. Gaudry and R. E. Edwards, *Littlewood-Paley and Multiplier Theory*, vol. 90. Springer, 2012.
- [63] O. M. Essenwanger, *Elements of Statistical Analysis*. Amsterdam, The Netherlands: Elsevier, 1986.
- [64] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 1, pp. 181–190, Jan. 2014.
- [65] W. Caesarendra, F. T. Putri, M. Ariyanto, and J. D. Setiawan, "Pattern recognition methods for multi stage classification of Parkinson's disease utilizing voice features," in *Proc. IEEE Int. Conf. Adv. Intell. Mechatronics (AIM)*, Jul. 2015, pp. 802–807.
- [66] J. J. P. C. Rodrigues, D. B. De Rezende Segundo, H. A. Junqueira, M. H. Sabino, R. M. Prince, J. Al-Muhtadi, and V. H. C. De Albuquerque, "Enabling technologies for the Internet of health things," *IEEE Access*, vol. 6, pp. 13129–13141, 2018, doi: [10.1109/access.2017.2789329](https://doi.org/10.1109/access.2017.2789329).



ANDRIUS LAURAITIS received the Ph.D. degree in informatics from the Kaunas University of Technology, Lithuania, in 2020. He is currently a Lecturer with the Department of Multimedia Engineering, Kaunas University of Technology, Kaunas, Lithuania. His research interests are computer aided diagnostics, medical decision support, and Huntington disease.



RYTIS MASKELIŪNAS (Member, IEEE) received the Ph.D. degree in computer science, in 2009. He is currently a Professor with the Department of Applied Informatics, Vytautas Magnus University, Kaunas, Lithuania. He is author or coauthor of more than 80 refereed scientific articles and serves as a Reviewer/Committee Member for various refereed journals. His main areas of scientific research are multimodal signal processing, modeling, development and analysis of associative, multimodal interfaces, mainly targeted at elderly, and people with major disabilities. He has won various awards/honors, including the Best Young Scientist Award of 2012, the National Science Academy Award for Young Scholars of Lithuania, in 2010, and others.



ROBERTAS DAMAŠEVIČIUS (Member, IEEE) received the Ph.D. degree in informatics engineering from the Kaunas University of Technology, Lithuania, in 2005. He is currently a Professor with the Department of Applied Informatics, Vytautas Magnus University, Lithuania, and an Adjunct Professor with the Faculty of Applied Mathematics, Silesian University of Technology, Poland. He also lectures software maintenance, human–computer interface, and robot programming courses. His research interests include sustainable software engineering, human–computer interfaces, assisted living, data mining, and machine learning. He is the author of more than 270 articles as well as a monograph published by Springer. He is also the Editor-in-Chief of the *Information Technology and Control Journal*. He has been the Guest Editor of several invited issues of international journals, such as *BioMed Research International*, *Computational Intelligence and Neuroscience*, the *Journal of Healthcare Engineering*, *IEEE ACCESS*, and *Electronics*.



TOMAS KRILAVIČIUS received the Ph.D. degree in computer science from the University of Twente, The Netherlands, in 2006. He is currently a Chief Scientist with the Baltic Institute of Advanced Technology, the Head of the Applied Informatics Department, a Professor with Vytautas Magnus University, and a stakeholder and co-founder of TokenMill. His research interests are modeling of complex systems, machine learning, language technologies, scientific infrastructures, education, and management of research projects.

...