Archives of Acoustics Vol. 50, No. 1, pp. 95–105 (2025), https://doi.org/10.24425/aoa.2024.148818

Research Paper

Method for Vocal Fold Paralysis Detection Based on Perceptual and Acoustic Assessment

Rafał HALAMA, Krzysztof SZKLANNY*, Danijel KORŽINEK

Polish-Japanese Academy of Information Technology
Warsaw, Poland

*Corresponding Author e-mail: kszklanny@pjwstk.edu.pl

(*received November 24, 2023; accepted November 30, 2024; published online December 12, 2024*)

This study is aimed to evaluate a method for distinguishing between healthy and pathological voices. The evaluation was carried out using several acoustic parameters including COVAREP (collaborative voice analysis repository for speech technologies), the auditory-perceptual RBH (roughness, breathiness, hoarseness) scale, and AVQI (acoustic voice quality index). Finally, a classifier is trained using machine learning algorithms from the WEKA (Waikato Environment for Knowledge Analysis) platform.

The study group comprised 75 voice recordings of individuals affected by vocal fold paralysis. The control group consisted of 49 voice recordings of healthy individuals. The results indicate that the voice quality of the study group is significantly different than the voice quality of the control group. Acoustic parameters implemented in COVAREP and the RBH scale have proven to be reliable methods assessing voice quality. In addition, data classification achieved over 90 % accuracy for every classifier.

Keywords: voice quality; AVQI; COVAREP; RBH scale; vocal fold paralysis.

1. Introduction

Voice is a key element in everyone's daily life as it is needed to communicate with other people. Three components are required for proper voice production: breathing, phonation, and articulation (Majkowska, 2004). For a human to produce a sound, simultaneous orchestration of several organs is required. The human breathing apparatus consists of lungs, diaphragm, trachea, and bronchi. It generates a driving force in the form of a stream of air exhaled from the lungs, which is needed to produce air turbulence and, therefore, sound. The phonation apparatus consists of the larynx with the vocal folds, vocal muscles, and the laryngeal nerve system. The airflow through the bronchi and trachea into the larynx causes vocal folds to vibrate, which are the sound source for voiced parts of speech. The articulation apparatus consists of the oral cavity, along with the tongue, the pharynx, and the nasal cavity. The oral cavity's role is to amplify and filter the sound produced in the larynx, thus transforming it into an articulated sound that is intelligible as speech. When the uvula of the soft palate is properly positioned the sound wave is emitted through the nasal cavity and nostrils (Tadeusiewicz, 1988).

A healthy voice, also known as euphonic, is characterised by correct and clear articulation, good diction, and the smooth change of intensity and fundamental frequency depending on the content of the utterance. The air pressure of a person with such a voice is perfectly regulated. The close-ups of the vocal fold and the onset of exhalation occur at the same time. The opposite of a euphonic voice is a pathological voice. Voice pathology manifests itself in the form of aphonia and dysphonia. Dysphonia is characterised by hoarseness, abnormal timbre, loudness, and duration of the utterance (Kosztyła-Hojna *et al.*, 2014). Aphonia is defined as the inability to produce a voice. It may be caused by surgery, tumor, or psychological means (Roper, 2014).

Vocal fold paralysis is caused by damage to the laryngeal nerves. The patient can suffer from unilateral or bilateral paralysis, the former of which is more common. We distinguish between central and periph-

eral vocal fold paralysis. Peripheral causes can be divided into traumatic and non-traumatic causes. Traumatic causes are mostly caused by surgery on the thyroid gland, either because of goiter or cancer. Other causes include communication injuries, heart, lungs, neck vessels or tracheal tumor surgeries, and intubation injuries. Non-traumatic causes include respiratory diseases such as tuberculosis, cancer, or enlarged lymph nodes. They also include viral infections such as shingles, influenza, esophageal, tracheal, and bronchial neoplasms, aortic aneurysm, myocardial hypertrophy, and mediastinal diseases. Patients with vocal fold paralysis have impaired defensive function of the larynx, which may cause choking on saliva or food. The voice of such a person is monotonous and dull. The fundamental frequency and timbre of the voice can change rapidly (Chen *et al.*, 2007).

In the medical environment, the assessment of voice pathology is based on multiple different factors, including questionnaires for self-assessment, expert derived perceptual analysis (e.g., using the GRBAS scale (Hirano, 1981)), acoustic analysis (e.g., jitter, shimmer, noise-to-harmonic ratio (Boersma, 2001)), aerodynamic analysis (e.g., maximum phonation time, mean airflow rate (Speyer *et al.*, 2010)), and vocal range analysis (e.g., fundamental frequency and intensity range (Cooper, Sorensen, 1981)). This assessment is repeated at several stages of administrating medication or therapy, thus allowing for a correlation comparison of various methods of medical treatment – see Table 1 (Jeong *et al.*, 2022).

In recent years, especially spurred by the COVID-19 pandemic, much of the diagnosis and pre-screening have been performed in a purely remote setting (Montalbaron *et al.*, 2023). As in this study, a common approach relies on using artificial intelligence (AI) in computer-aided diagnosis (Verikas *et al.*, 2006; Crowson *et al.*, 2020). Early systems relied on simple Mel-frequency cepstrum coefficients and hidden Markov models (Dibazar *et al.*, 2006) – an approach common in speech recognition systems of the era. More modern solutions rely on deep learning and other novel machine learning techniques (Compton *et al.*, 2022; Tirronen *et al.*, 2023; Suvvari, 2023). The results outlined in the cited literature were relatively compared to those discussed in this paper. However, the comparison is difficult as the datasets of other authors are generally not available for a direct comparison.

This study aimed to prove that both acoustic and perceptual analysis are valuable tools for detecting changes in voice quality. Through a series of experiments using several classifiers, the data were successfully classified into voice recordings of people suffering from vocal fold paralysis and voice recordings of healthy individuals.

2. Speech database

The recordings were conducted in 1973–1996 in the Institute of Phoniatrics at the Central Clinical Hospital, 1a Banacha St. in Warsaw. The Nagra IV S series professional tape recorder was used to record the speech in non-acoustically adapted room. (Wow and flutter (9.5 cm/s) ±0.012 %, according to DIN 45507 standard, 0.043 % according to NAB standards. Signal-to-noise ratio (SNR) ASA *A*-weighted, reference 1 mW 125 dBm). The recordings contain 416 recordings of patients with various diseases affecting voice quality, such as after adenoidectomy, tubectomy, cordectomy with vocal fold paralysis, or dysphonia. Each patient underwent a phoniatric examination. A significant number of patients had their voices recorded repeatedly, which may allow us to compare the performance of our system on the same voice before and after rehabilitation.

Following speech signals were recorded: vowels /: *i*/ /: *y*/ /: *a*/ /: *e*/ /: *o*/ /: *u*/ read at equal intervals, simple announcing, and questioning sentences, and scientific text that the patient was not familiar with before the study began.

In addition, 10 sentences of text were recorded. All the recordings, which were conducted using a rarely employed cost-effective speed of 9.5 cm/s, are stored on analogue reel tapes in the Institute of Phoniatrics's archives. The speed does not influence the quality of recorded speech.

The crucial process was to digitise the recordings. It was conducted at the Polish-Japanese Academy of

Table 1. Voice outcome measures as outlined by Jeong *et al.* (2022).

Category of outcome measurement	Definitions and examples
Visuo-perceptual	Subjective rating of laryngeal anatomy function, e.g., videostroboscopy, laryngoscopy, stroboscopy research tool
Auditory-perceptual	Subjective rating of the perceptual vocal quality, e.g., GRBAS (Hirano, 1981), CAPE-V (Nemr *et al.*, 2012)
Acoustic	Computerized measurements of features of the speech sound signal, e.g., jitter, shimmer, noise-to-harmonic ratio, cepstral peak prominence
Aerodynamic	Measures of respiratory components of phonation, e.g., maximum phonation time, S/Z ratio, subglottal pressure
Voice-related quality-of-life measures	Patient rated assessment of the impact of dysphonia, e.g., vocal handicap index (Wilson *et al.*, 2004), V-RQOL (Hogikyan, 2004)

Information Technology using the Studer A812 reel-to-reel tape recorder (Rosłanowski, 2008). The analogue signal from the recorder was sent to the computer via an E-MU 1616 audio interface. The connection was made using a symmetric cable with one end plugged into the CH1 connection of the recorded and the other end plugged into the audio interface's input. The signal was recorded in Sony Sound Forge at a sampling rate of 44.1 kHz and a 16-bit depth.

A database containing the patient's name, date of recording, disease description, ID, and file name recording, keywords, age, gender, and tape number was also created.

Examples of the transcript used are provided in Appendix A, together with its translation in Appendix B. A subset of the recordings, where patients phonated the sustained /a:/ vowel and uttered sentences in Polish: "Ten dzielny żołnierz był z nim razem. Ola lubi bezy", were included in the experiments. The voice recording was excluded if a vowel's phonation was not sustained for at least 1 second.

2.1. Pre-processing of recordings

The acoustic background and reverberation in the room used for recording exceeded appropriate levels, which affected the quality of the voice recordings. All of them had to be subjected to a noise reduction process. Firstly, the SNR was calculated for every voice recording. The SNR is a difference, measured in decibels, between the speech level and the background noise level.

Previous studies reported that recommended levels of SNR are above 42 dB, acceptable: above 30 dB, and unacceptable: below 30 dB (Ingrisano *et al.*, 1998; Deliyski *et al.*, 2005). To eliminate mains hum, we used a FIR high-pass filter to reduce all frequencies below 60 Hz (Fig. 1), which greatly improved the SNR levels of all recordings. Before the process, the SNR ranged from 17.9 dB to 40.9 dB, averaging 26.2 dB.

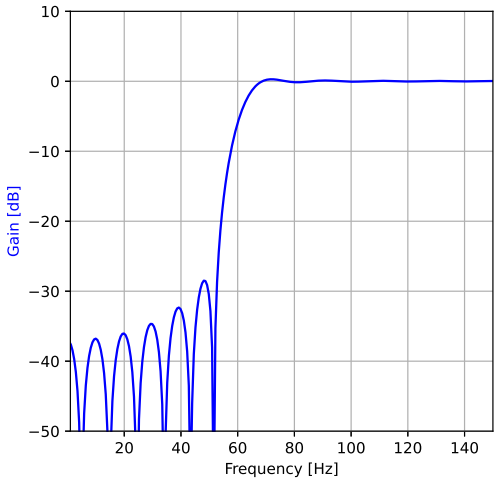

Fig. 1. FIR filter used to process the voice recordings.

After the process, the SNR ranged from 23.5 dB to 48.7 dB, averaging 36.1 dB. Only one recording was deemed unusable and was excluded from the study.

Voice recordings were sorted by the type of vocal disorder. Vocal cord paralysis, which is the goal of this study, was the only one that appeared more than a few times in the database. Only 75 recordings were used in further experiments. Forty-nine recordings which came from 17 healthy individuals were used as a control group.

2.2. Perceptual assessment of voice quality

All voice recordings used in this study were assessed by independent voice specialists using the RBH scale (Nawka *et al.*, 1994). The scale is widely recognised as the easiest method of perceptual voice evaluation by institutions including the Committee on Phoniatrics of the European Laryngological Society (Dejonckere *et al.*, 2001). The RBH scale consists of three features: R – roughness; B – breathiness; H – hoarseness.

Every feature can receive a score from 0 to 3, which describes the severity of a vocal disorder: 0 – normal voice, 1 – a slight change, 2 – medium change, and 3 – high change.

The RBH scale, despite looking uncomplicated, is a reliable method of assessing voice quality, provided it is used by voice specialists such as phoniatrists or speech therapists (Behrbohm *et al.*, 2011).

Perceptual assessment of voice quality was carried out on both occasions by the same two independent voice specialists who had completed an RBH training program and had extensive experience in voice/speech signal assessment. On both occasions, the experts were blindfolded for the assessment duration. The two experts underwent an audiometry test, and the test results for both indicated normal hearing.

2.3. Control group recordings

Because the original dataset contained only voices with voice quality disorders, an additional set of recordings was created to capture the vocal properties of healthy individuals for control purposes. The recordings were made in the recording studio of the Polish-Japanese Academy of Information Technology. The microphone used in the recordings was a Rode NT-1A and it has the following parameters: frequency range 2 Hz–20 kHz, sensitivity 25 mV/Pa, equivalent noise level 5 dBA, maximum SPL – 137 dB SPL, polar pattern – cardioid. The signal was registered with a 48 kHz sampling rate and a 16-bit resolution (standard WAV PCM).

During the recording, the healthy individuals phonated the vowel /a:/ three times with a sound pressure level of 60 dBA–80 dBA, 1 meter from the microphone, for a sustained period of at least 4 seconds.

Following that, the recorded individual was made to briefly strain his/her voice by reading out a few sentences, and then again to phonate the vowel /:a/ four times.

The last four phonations of the vowel /:a/ were used to calculate the acoustic parameters. All the participants phonated neutrally. Phonations with higher or lower values of the fundamental frequency of a speech signal, often denoted by $F0$, were not considered in the analyses.

A lot of consideration was taken to match the conditions of the original dataset while preparing the control samples. It is obviously impossible to recreate the conditions perfectly, but the chosen signal analysis methods were not affected by the differences in the acquisition and storage of the signal data. Given the overall low levels of background noise and good levels of SNR, both sets of recordings showed negligible levels of change in parameter values.

3. Acoustic voice evaluation

Acoustic methods for voice quality assessment are growing in popularity amongst clinicians focusing on voice research, because these methods benefit from being non-invasive and give the opportunity of utilising automation (Maryn *et al.*, 2009). They are an easy and reliable way of comparing voice dysphonia levels before surgeries and after them (Maryn *et al.*, 2009). Traditionally, sustained vowel phonation is used for testing instead of continuous speech (Askenfelt, Hammarberg, 1986). In the case of vowels, features such as talking speed, pauses, the context of a sentence, accent, or type of language spoken are not relevant. On the other hand, this approach can sometimes be worse than continuous speech because sustained vowel phonation is not representative of everyday use of speech in a normal spontaneous setting (Parsa, Jamieson, 2001). That is why the best results are obtained while using both methods.

One example of using acoustic analysis is an acoustic parameter, which evaluates the voice quality based on the parametrized sound signal. Collaborative voice analysis repository for speech technologies (COVAREP) is a free toolkit with many implementations of acoustic parameters (Degottex *et al.*, 2014) and it is available as an open source public repository online written in MATLAB. The following acoustic parameters which were used for our study were implemented in COVAREP: peak slope – PS (Kane, Gobl, 2011), normalised amplitude quotient – NAQ (Alku *et al.*, 2002), parabolic spectral parameter – PSP (Alku *et al.*, 1997), quasi-open quotient – QOQ (Hacki, 1989), cepstral peak prominence – CPP (Hillenbrand, Houde, 1996), H1H2 (Hanson, 1997), harmonic richness factor – HRF (Childers, Lee, 1991), and maxima dispersion quotient – MQD (Kane, Gobl, 2013). Voice recordings included only the sustained phonation of the /:a/ vowel, which meant they could not be used in the experiments in which continuous speech was also needed.

3.1. Peak slope

The PS is calculated by observing the wavelet decomposition given the following formula for the mother wavelet:

$$g(t) = -\cos\left(2\pi f_n t\right) \cdot \exp\left(\frac{t^2}{2\tau^2}\right), \tag{1}$$

where $f_n = \frac{f_s}{2}$, for f_s being the sampling frequency of 16 kHz and $\tau = \frac{1}{2f_n}$. This decomposition results in an octave band filter bank with centre frequencies at 8 kHz, 4 kHz, 2 kHz, 1 kHz, 500 Hz, and 250 Hz. From this filterbank, a local maximum is located for each band and a regression line is computed based on the amplitudes of the observed maxima (see Fig. 1 in (Kane, Gobl, 2011)).

This acoustic parameter differentiates between a modal, tense, or breathy voice. According to previous studies, the PS parameter has a certain advantage compared to other parameters (Kane, Gobl, 2011). It is completely independent, meaning that no other algorithm is used to compute its value. It is especially useful when the voice recording has an ambient noise that may disturb other algorithms and, consequently, affect the obtained values.

3.2. Normalised amplitude quotient

The NAQ is a time-based acoustic parameter used for speech signal analysis. Studies have suggested that the parameter effectively differentiates types of phonations and demonstrates resistance to the presence of noise in the speech signal (Alku *et al.*, 2002).

It is computed for each glottal flow period using the following formula (Alku *et al.*, 2002):

$$\frac{A_{\mathrm{ac}}}{T_{\mathrm{av}} \cdot d_{\mathrm{min}}} = \frac{A_{\mathrm{max}} - A_{\mathrm{min}}}{T_{\mathrm{av}} \cdot d_{\mathrm{min}}}, \tag{2}$$

where A_{max} is the amplitude for each period of the signal, A_{min} is the lowest amplitude for each period of the signal, T_{av} is the average fundamental period length, d_{min} is the minimum derivative glottal flow, and A_{ac} is the maximal flow of amplitude.

3.3. Parabolic spectral parameter

The PSP is an acoustic parameter based on fitting a parabolic function to the low-frequency part of the calculated glottal flow spectrum. The parameter is a single numerical value that describes how the spectral decay of the resulting glottal flow behaves with respect to the theoretical limit corresponding to the

maximum decay. The PSP is commonly compared with other time-based acoustic parameters (Alku *et al.*, 1997).

3.4. Maxima dispersion quotient

The MDQ is an acoustic parameter used to differentiate between modal, breathy, or tense voice. Previous studies show that the parameter is effective in assessing voice type based on sustained vowel phonation and continuous speech, which achieves better results than the NAQ parameter (Kane, Gobl, 2013).

For a tense voice, the maxima tend to appear around glottal closure instants (GCIs), which mark the moments of greatest excitation of vocal folds in the glottal airflow. Otherwise, if the voice is breathier, it has been observed that the maxima are scattered. The MDQ parameter recognises the scale of maxima scattering and thus effectively indicates the type of voice, and it obtains particularly good results during the analysis of continuous speech (Kane, Gobl, 2013).

3.5. Quasi-open quotient

The QOQ is an acoustic time domain parameter. It is calculated by measuring the distance between two points around and closest to the maximum of the glottal flow pulse, which are exactly 50 % of the maximum's amplitude value. This duration is also normalised with respect to the pitch period T_0 (Fig. 2).

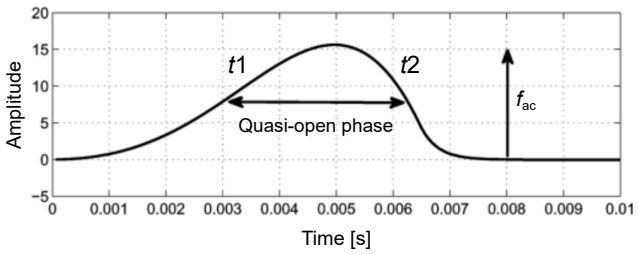

Fig. 2. Amplitude of a glottal flow impulse.

As confirmed in studies (Kane, Gobl, 2013), the QOQ parameter achieves weaker results than the MDQ and NAQ parameters. Only in the case of SNR ranging between 0 dB and 10 dB, this parameter works better.

3.6. Cepstral peak prominence

In 2018 CPP was recommended by the American Speech-Language-Hearing Association (ASHA) as a tool that allows to measure the degree of noise and other unwanted sounds in the voice signal as well as to detect the degree of dysphonia (Patel *et al.*, 2018). CPP is defined by the distance between the top of the cepstrum and its regression line. As shown in the research of Hillenbrand and Houde (1996), the cepstral maxima are more visible in the cepstrum of a breathy voice than in the cepstrum of a modal voice which makes it possible to distinguish between these types of phonations using this parameter.

3.7. H1–H2

This acoustic parameter helps to distinguish between breathy and tense voices, which was confirmed in the studies by Hanson (1997), Airas and Alku (2007). It is calculated by the difference between the amplitude of the first two vocal harmonies in the spectrum of the voice source. It is described in decibels [dB]. The H1–H2 parameter is less accurate than the MDQ and NAQ parameters (Kane, Gobl, 2013). Only when the SNR of the recording oscillates between 0 dB and 10 dB, this parameter achieves better results than its counterparts.

3.8. Harmonic richness factor

The HRF is described as the ratio of the sum of the harmonic amplitudes in the glottal flow to the component amplitude at the fundamental frequency. In previous studies (Childers, Lee, 1991), the HRF parameter's scores were higher by 6.8 dB for a modal voice compared to a breathy voice, which effectively allows to distinguish between these types of phonations.

4. Acoustic voice quality index

The AVQI is a tool developed to measure overall voice quality using acoustic markers for clinical purposes. For the voice quality evaluation to be accurate and representative, the AVQI needs continuous speech and sustained vowel phonation, which lasts for a few seconds (Maryn, Roy, 2012).

The AVQI ranges between 0 and 10 and has a cut-off score between a healthy and pathological voice, which differs depending on the language, but generally, it is around 3 (Fig. 3). The more an AVQI score exceeds the cut-off threshold the higher the severity of voice dysphonia. The threshold for English and Australian equals 3.46 (Reynolds *et al.*, 2012), 2.70 for German (Barsties, Maryn, 2012), 3.07 for French (Maryn *et al.*, 2014), 2.95 for Dutch (Maryn *et al.*, 2010), 2.97 for Lithuanian (Uloza *et al.*, 2017), 3.15 for Japanese (Hosokawa *et al.*, 2017), 2.02 for Korean (Maryn, Weenink, 2015), and 3.09 for Finnish language (Kankare *et al.*, 2020). Measurement errors must be considered while using the AVQI. The difference in results between the two recordings should be at least 0.54 (Barsties, Maryn, 2012) to mark that the voice quality has changed. To our knowledge, there is no data about AVQI parameters for the Polish language.

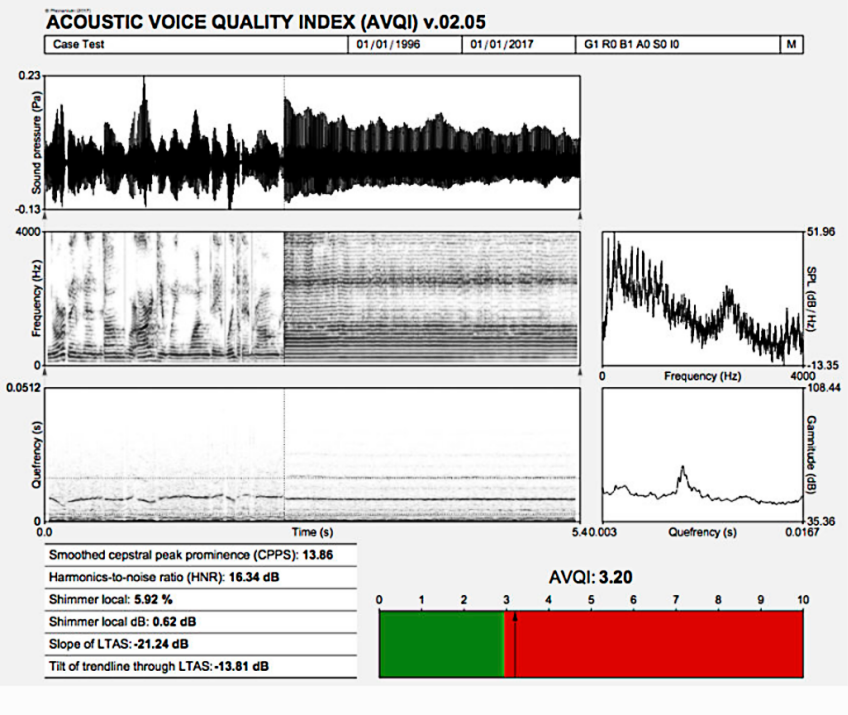

Fig. 3. Example of AVQI results.

5. Inter-rater reliability

Two independent experts who used the RBH scale to assess the voice quality of recordings in the database of non-healthy individuals were tested for inter-rater reliability because one of them had a sensitive hearing, which could heavily affect the results of experiments. Tests were conducted using MedCalc software and Real Statistics Resource Pack addon for Excel. To check the expert's agreement a single measure of the intraclass correlation coefficient (ICC) was used, which was previously used in other studies (Maryn *et al.*, 2014). The suggested limit between a good and a weak or an average agreement is 0.75 (Portney, Watkins, 2009). The obtained results for the R, B, and H parameters were as follows 0.56, 0.5, 0.46, which gave us an average of 0.51. In our experiment, we noticed a shift in the annotation of the recordings between the voice specialists. The scores recorded by expert 2 proved to be more sensitive to changes in voice quality than those recorded by expert 1. The two experts underwent an audiometry test, and the test results for both indicated normal hearing. For further discussion on the inter-rater reliability of experts can be found in previous studies; Table 2 (Szklanny, Wrzeciono, 2019).

Table 2. Perceptual score distribution among experts. Scores of 0, 1, 2, and 3 are used for all parameters on the RBH scale, with reference to the different degrees of vocal disorder: 0 – a normal voice; 1 – a slight degree; 2 – a medium degree; 3 – a high degree.

Score	Expert	0	1	2	3
R	Expert 1	172	149	70	26
	Expert 2	38	172	148	59
B	Expert 1	100	195	87	35
	Expert 2	41	129	179	68
H	Expert 1	49	243	92	33
	Expert 2	1	110	149	157

6. Acoustic analysis results

The AVQI score was tested to correlate with RBH scores for the same voice recordings. We used Spearman's rank-order correlation coefficient and RBH scores of experts were averaged. AVQI and the R feature had a weak correlation, while AVQI and two other features noted a higher-than-average level of correlation. Table 3 presents the results.

Table 3. Results of Spearman's Rank-Order Correlation coefficient for AVQI and RBH.

R and AVQI	B and AVQI	H and AVQI
0.371	0.655	0.594

Acoustic voice parameters obtained through COVAREP were tested for a correlation with the AVQI score for the same voice recording. With the use of Spearman's Rank-Order Correlation coefficient it was noted that the PS parameter from COVAREP had a significant correlation with the AVQI score amounting to 0.62 for a vowel, and 0.69 for a continuous speech. The parameter CPP, which is used for cal-

Table 4. Results of Spearman's Rank-Order Correlation coefficient of AVQI with various acoustic parameters of non-healthy individuals.

Phonation type	NAQ	QOQ	H1H2	HRF	PSP	PS	MDQ	CPP
Vowel	−0.11	−0.4	0.03	−0.08	0.18	0.62	0.37	−0.84
Continuous speech	−0.35	−0.53	−0.16	−0.05	0.07	0.69	0.36	−0.77

Table 5. Results of Mann–Whitney U-test and Student t-test.

Parameter	Mean ±SD for non-healthy individuals	Mean ±SD for healthy individuals	Test results
CPP	12.41 ±0.66	11.47 ±0.47	p ¡ 0.0001 $U = 481$
H1H2	12.97 ±7.69	5.36 ±3.82	p ¡ 0.0001 $t = 6.76$
HRF	19.21 ±6.88	23.37 ±8.5	$p = 0.0014$ $U = 1214$
NAQ	0.174 ±0.05	0.11 ±0.02	p ¡ 0.0001 $t = 8.647$
PSP	0.27 ±0.08	0.16 ±0.06	p ¡ 0.0001 $t = 8.06$
QOQ	0.5 ±0.08	0.38 ±0.07	p ¡ 0.0001 $U = 435$
MDQ	0.11 ±0.02	0.1 ±0.02	$p = 0.0001$ $t = 4.068$
PS	−0.42 ±0.05	−0.31 ±0.04	p ¡ 0.0001 $U = 202$

culations in AVQI, had a significant negative correlation with the AVQI score amounting to −0.84 for a vowel, and −0.77 for a continuous speech. Similar results were observed in the study on the Finnish language, where the correlation between the CPP parameter and the AVQI score was equal to −0.35 (Laukkanen, Rantala, 2022). Table 4 shows the tests results.

The Shapiro–Wilk test was used to check whether acoustic parameters had normal distribution or not. An F-test was run to check if the variance was equal. Two variants of the Student t-test were used for acoustic parameters with normal distribution: the Student t-test for an equal variance or the Student-test for unequal variance. As the distribution for other parameters was not normal, the Mann–Whitney U-test was used in their case. Table 5 shows that all acoustic voice parameters calculated for recordings of healthy individuals were statistically different from their counterparts calculated for individuals suffering from vocal fold paralysis.

The Shapiro–Wilk test was used to check whether RBH scores had a normal distribution or not. As their distribution was not normal, the Mann–Whitney U-test was used. Results showed that RBH scores for healthy individuals were statistically different from RBH scores for non-healthy individuals.

7. Classification

During the final experiment, we tried to differentiate a healthy voice from a voice affected by vocal cord paralysis using the classification based on acoustic voice parameters. All calculations were done in the WEKA software.

For the experiment, we used five classifiers, which were proven to be effective in previous studies on voice disorders (Verde *et al.*, 2018): Naïve Bayes, support vector machine (SVM), decision tree, logistic model tree, instance-based learning algorithm k-NN.

Naïve Bayes is a classifier based on Bayes' theorem and the probability theory. The features of such a classifier are independent, so neither of them affects the other (Friedman *et al.*, 1997).

The SVM is a classifier defined by a hyperplane, that separates data belonging to different classes with the widest possible margin. This technique distinguishes between a healthy and pathological voice because it natively splits the data into up to two classes. The classification accuracy can be increased by changing the parameters and a function of the kernel (Godino-Llorente *et al.*, 2005). This study used the polynomial function, which is one of the two most popular kernel functions used in the SVM (Alpaydin, 2004).

The decision tree is a technique used for classifying categorised data based on the training method represented by a decision tree. Decision trees are easy to interpret and can handle both continuous and categorical data. In this work, we used J48, based on the C4.5 algorithm (Quinlan, 1999), which is the most popular tree-based classifier.

The logistic model tree is a technique that combines logistic regression, a probability-based machine learning algorithm, with a decision tree. In the WEKA software, it is implemented by the SimpleLogistic class (Landwehr *et al.*, 2005).

Instance-based learning algorithms are algorithms that use specific instances to obtain the results of a classifier. In this study, we used the k-NN algorithm (Aha *et al.*, 1991), which bases its results on the k-number of nearest neighbours in a new instance.

The dataset containing 75 voice recordings of non-healthy individuals and 49 voice recordings of healthy individuals with their acoustic parameters calculated

Table 6. Classification results.

Classifier	Parameters	Accuracy [%]	Sensitivity [%]	Specificity [%]	MAE
Naïve Bayes	NAQ, QOQ, H1H2, CPP, PSP, PS	95.16	98.59	90.57	0.059
SVM	NAQ, QOQ, H1H2, CPP, PSP, PS, HRF, MDQ	94.35	98.57	88.89	0.057
Decision tree	NAQ, QOQ, H1H2, CPP, PSP, PS	91.94	95.77	86.79	0.09
Logistic tree	NAQ, QOQ, H1H2, CPP, PSP, PS	94.35	97.22	90.38	0.12
k-NN	NAQ, QOQ, H1H2, CPP, PSP, PS, HRF	98.39	100	96.08	0.024

was prepared. Then, it was imported to the software WEKA and then underwent a classification process with the use of 10-fold cross-validation. We have calculated every classifier's accuracy, sensitivity, specificity, and mean absolute error (MAE). Accuracy describes the percentage of correctly classified data. Sensitivity describes the effectiveness of classifying positive cases. Specificity describes the effectiveness of the classification of negative cases. The MAE is a measure that determines how much on average the forecast period deviates from the real value.

Table 6 presents that the best results were received while using the k-NN classifier with a group of acoustic parameters (NAQ, QOQ, H1H2, CPP, PSP, PS, HRF). The decision tree (NAQ, QOQ, H1H2, CPP, PSP, PS) achieved the lowest accuracy: 91.94 %. The biggest MAE was received using the logistic model tree (NAQ, QOQ, H1H2, CPP, PSP, PS).

8. Discussion

Conducted experiments have shown that both the perceptual evaluation and the acoustic evaluation have the potential to distinguish a healthy voice from a pathological voice affected by vocal fold paralysis.

The biggest difficulty was encountered while processing the database of non-healthy individuals. This database contained voice recordings from 40–50 years ago, which were recorded on old analogue tapes. In addition, the standard of research has changed drastically over the last decades, so a significant part of the recordings could not be used for this study. Noise reduction due to unwanted background noise also turned out to be very time-consuming and the process should have been automated.

The perceptual assessment of experts who graded voice recordings of non-healthy individuals using the RBH scale was a significant problem. One expert's sensitive hearing led him to grade voice recordings differently from the other expert. Undoubtedly, this fact has influenced the results of some experiments.

An interesting finding was the negative correlation of the CPP parameter, which is one of the components needed to calculate the AVQI score. A similar correlation was found in the studies on the Finnish language (Laukkanen, Rantala, 2022).

Every used classifier, whose accuracy was confirmed in the previous studies (Verde *et al.*, 2018), achieved over 90 % accuracy, which is a very high result for data classification. Such scores are reported in the literature to be on par with the level of human experts (Suvvari, 2023).

A similar study was carried out in (Szklanny, 2019), which investigated the differences in the values of acoustic parameters between choral singers and individuals with a healthy voice. The values of acoustic parameters were compared with a group of men with a healthy voice. Significant differences were only observed for parameters H1H2 and HRF.

Other studies utilise deep learning approaches (Compton *et al.*, 2022) and transfer learning (Tirronen *et al.*, 2023) providing a similarly high score at the cost of reduced interpretability of results.

9. Conclusion

The study shows that acoustic and perceptual analyses are valuable tools for detecting differences in voice quality. Using several classifiers, several experiments classified the data successfully into voice recordings of people suffering from vocal fold paralysis and voice recordings of healthy individuals.

Statistical tests have shown a medium-high correlation of the AVQI parameter with B and H features from the RBH perceptual scale. The acoustic parameter PS has shown a strong correlation with AVQI, while the CPP parameter has shown a strong, negative correlation with AVQI.

For further research, it would be advisable to expand the database with additional recordings of patients with vocal fold paralysis as well as healthy subjects, considering prolonged phonation of the vowel /: a/.

Appendix A. Examples of recorded sentences in Polish

Ten dzielny żołnierz był z nim razem. Ola lubi bezy. Czy Ola lubi bezy? Idziemy do domu.

Czy idziemy do domu? Dzień dobry. Do widzenia. Warszawa miasto pokoju. Warszawa stolica Polski. Do widzenia Pani. Do zobaczenia Panu. Dziś jest ładna pogoda. Czy dziś jest ładna pogoda?

Przeszło sto lat minęło od pojawienia się na ulicach Warszawy pierwszego konnego tramwaju, łączącego

dworce na Pradze z Dworcem Wiedeńskim przy ulicy Marszałkowskiej. Jeszcze dwa razy Stolica przeżywała podobnie uroczyste momenty – w 1908 roku i 15 września 1945 roku. Wtedy w zniszczonej stolicy na lewym brzegu Wisły rozpoczął kursowanie pierwszy powojenny tramwaj. Odbudowa Stolicy i rozbudowa linii tramwajowych następowały równie szybko. Rejon otaczający Dworzec Centralny stanowi obecnie również wielki plac budowy, chociaż prowadzi się tu dopiero różne roboty przygotowawcze. Załogi wielu przedsiębiorstw inżynieryjnych przekładają urządzenia podziemne. Coraz bliżej jest termin zakończenia budowy objazdów tramwajowych w Alejach Jerozolimskich oraz w ulicach Marchlewskiego i Chałubińskiego. Na usunięcie czekają jeszcze słupy oświetleniowe, stojące na linii zastępczego torowiska. Długość objazdowych torów wynosi ponad dwa kilometry. Będą się one przecinały przy ulicy Chałubińskiego w miejscu gdzie rozebrano narożny budynek.

Appendix B. Examples of recorded sentences translated to English

The brave soldier was with him. Ola likes meringue. Does Ola like meringue? We are going home. Are we going home? Good morning. Goodbye. Warsaw, the city of peace. Warsaw, the capital of Poland. Goodbye Mrs. Goodbye Mr. Today is nice weather. Is it nice weather today?

References

1. Aha D.W., Kibler D., Albert M.K. (1991), Instance-based learning algorithms, *Machine learning*, **6**: 37–66, https://doi.org/10.1007/bf00153759.
2. Airas M., Alku P. (2007), Comparison of multiple voice source parameters in different phonation types, [in:] *Eighth Annual Conference of the International Speech Communication Association*, https://doi.org/10.21437/interspeech.2007-28.
3. Alku P., Bäckström T., Vilkman E. (2002), Normalized amplitude quotient for parametrization of the glottal flow, *The Journal of the Acoustical Society of America*, **112**(2): 701–710, https://doi.org/10.1121/1.1490365.
4. Alku P., Strik H., Vilkman E. (1997), Parabolic spectral parameter – A new method for quantification of the glottal flow, *Speech Communication*, **22**(1): 67–79, https://doi.org/10.1016/s0167-6393(97)00020-4.
5. Alpaydin E. (2004), *Introduction to Machine Learning*, MIT Press.
6. Askenfelt A.G., Hammarberg B. (1986), Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures, *Journal of Speech, Language, and Hearing Research*, **29**(1): 50–64, https://doi.org/10.1044/jshr.2901.50.
7. Barsties B., Maryn Y. (2012), Der acoustic voice quality index [in German: Ein Messverfahren zur allgemeinen Stimmqualität], *HNO*, **60**(8): 715–720, https://doi.org/10.1007/s00106-012-2499-9.
8. Behrbohm H., Kaschke O., Nawka T., Swift A.C. (2011), *Ear, Nose and Throat Diseases with Head and Neck Surgery* [in Polish: *Choroby ucha, nosa i gardła z chirurgią głowy i szyi*], 2nd ed., Edra Urban & Partner.
9. Boersma P. (2001), Praat, a system for doing phonetics by computer, *Glot International*, **5**(9/10): 341–345.
10. Chen H.-C., Jen Y.-M., Wang C.-H., Lee J.-C., Lin Y.-S. (2007), Etiology of vocal cord paralysis, *ORL*, **69**(3): 167–171, https://doi.org/10.1159/000099226.
11. Childers D.G., Lee C.K. (1991), Vocal quality factors: Analysis, synthesis, and perception, *The Journal of the Acoustical Society of America*, **90**(5): 2394–2410, https://doi.org/10.1121/1.402044.
12. Compton E.C. *et al.* (2022), Developing an Artificial Intelligence tool to predict vocal cord pathology in primary care settings, *The Laryngoscope*, **133**(8): 1531–4995, https://doi.org/10.1002/lary.30432.
13. Cooper W.E., Sorensen J.M. (1981), *Fundamental Frequency in Sentence Production*, Springer Science & Business Media.
14. Crowson M.G. *et al.* (2020), A contemporary review of machine learning in otolaryngology–head and neck surgery, *The Laryngoscope*, **130**(1): 45–51, https://doi.org/10.1002/lary.27850.
15. Degottex G., Kane J., Drugman T., Raitio T., Scherer S. (2014), COVAREP – A collaborative voice analysis repository for speech technologies, [in:] *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 960–964, https://doi.org/10.1109/icassp.2014.6853739.
16. Dejonckere P.H. *et al.* (2001), A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques, *European Archives of Oto-rhino-laryngology*, **258**: 77–82, https://doi.org/10.1007/s004050000299.
17. Deliyski D.D., Shaw H.S., Evans M.K. (2005), Adverse effects of environmental noise on acoustic voice quality measurements, *Journal of Voice*, **19**(1): 15–28, https://doi.org/10.1016/j.jvoice.2004.07.003.
18. Dibazar A.A., Berger T.W., Narayanan S.S. (2006), Pathological voice assessment, [in:] *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, **2006**: 1669–1673, https://doi.org/10.1109/IEMBS.2006.259835.
19. Friedman N., Geiger D., Goldszmidt M. (1997), Bayesian network classifiers, *Machine Learning*, **29**: 131–163, https://doi.org/10.1023/A:1007465528199.
20. Godino-Llorente J.I., Gómez-Vilda P., Sáenz-Lechón N., Blanco-Velasco M., Cruz-Roldán F., Ferrer-Ballester M.A. (2005), Support vector machines applied to the detection of voice disorders,

[in:] *Nonlinear Analyses and Algorithms for Speech Processing. NOLISP 2005. Lecture Notes in Computer Science*, Faundez-Zanuy M., Janer L., Esposito A., Satue-Villar A., Roure J., Espinosa-Duro V. [Eds.], pp. 219–230, https://doi.org/10.1007/11613107_19.

21. Hacki T. (1989), Classification of glottal dysfunctions on the basis of electroglottography [in German: Klassifizierung von glottiscysfunktionen mit hilfe der elektroglottographie], *Folia phoniatrica*, **41**(1): 43–48, https://doi.org/10.1159/000265931.
22. Hanson H.M. (1997), Glottal characteristics of female speakers: Acoustic correlates, *The Journal of the Acoustical Society of America*, **101**(1): 466–481, https://doi.org/10.1121/1.417991.
23. Hillenbrand J., Houde R.A. (1996), Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech, *Journal of Speech, Language, and Hearing Research*, **39**(2): 311–321, https://doi.org/10.1044/jshr.3902.311.
24. Hirano M. (1981), *Clinical Examination of Voice*, Springer-Verlag, New York.
25. Hogikyan N.D. (2004), The voice-related quality of life (V-RQOL) measure: History and ongoing utility of a validated voice outcomes instrument, *Perspectives on Voice and Voice Disorders*, **14**(1): 3–5, https://doi.org/10.1044/vvd14.1.3.
26. Hosokawa K. *et al.* (2017), Validation of the acoustic voice quality index in the Japanese language, *Journal of Voice*, **31**(2): 260.e1–260.e9, https://doi.org/10.1016/j.jvoice.2016.05.010.
27. Ingrisano D.R., Perry C.K., Jepson K.R. (1998), Environmental noise: A threat to automatic voice analysis, *American Journal of Speech-Language Pathology*, **7**(1): 91–96, doi: https://doi.org/10.1044/1058-0360.0701.91.
28. Jeong G.-E. *et al.* (2022), Treatment efficacy of voice therapy following injection laryngoplasty for unilateral vocal fold paralysis, *Journal of Voice*, **36**(2): 242–248, https://doi.org/10.1016/j.jvoice.2020.05.014.
29. Kane J., Gobl C. (2011), Identifying regions of non-modal phonation using features of the wavelet transform, [in:] *Twelfth Annual Conference of the International Speech Communication Association*, pp. 177–180, https://doi.org/10.21437/interspeech.2011-76.
30. Kane J., Gobl C. (2013), Wavelet maxima dispersion for breathy to tense voice discrimination, [in:] *IEEE Transactions on Audio, Speech, and Language Processing*, **21**(6): 1170–1179, https://doi.org/10.1109/tasl.2013.2245653.
31. Kankare E. *et al.* (2020), The acoustic voice quality index version 02.02 in the Finnish-speaking population, *Logopedics Phoniatrics Vocology*, **45**(2): 49–56, https://doi.org/10.1080/14015439.2018.1556332.
32. Kosztyła-Hojna B., Moskal D., Kuryliszyn-Moskal A., Rutkowski R. (2014), Visual assessment of voice disorders in patients with occupational dysphonia, *Annals of Agricultural and Environmental Medicine*, **21**(4): 898–902, https://doi.org/10.5604/12321966.1129955.
33. Landwehr N., Hall M., Frank E. (2005), Logistic model trees, *Machine Learning*, **59**: 161–205, https://doi.org/10.1007/s10994-005-0466-3.
34. Laukkanen A.-M., Rantala L. (2022), Does the acoustic voice quality index (AVQI) correlate with perceived creak and strain in normophonic young adult Finnish females?, *Folia Phoniatrica et Logopaedica*, **74**(1): 62–69, https://doi.org/10.1159/000514796.
35. Majkowska M. (2004), Basic issues of voice emission and hygiene [in Polish: Podstawowe zagadnienia emisji i higieny głosu], [in:] *Prace Naukowe Akademii im. Jana Długosza w Częstochowie*, **5**: 93–101.
36. Maryn Y., Corthals P., Van Cauwenberge P., Roy N., De Bodt M. (2010), Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels, [in:] *Journal of Voice*, **24**(5): 540–555, https://doi.org/10.1016/j.jvoice.2008.12.014.
37. Maryn Y., De Bodt M., Barsties B., Roy N. (2014), The value of the acoustic voice quality index as a measure of dysphonia severity in subjects speaking different languages, *European Archives of Oto-Rhino-Laryngology*, **271**: 1609–1619, https://doi.org/10.1007/s00405-013-2730-7.
38. Maryn Y., Roy N. (2012), Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity, *Jornal da Sociedade Brasileira de Fonoaudiologia*, **24**: 107–112, https://doi.org/10.1590/s2179-64912012000200003.
39. Maryn Y., Roy N., De Bodt M., Van Cauwenberge P., Corthals P. (2009), Acoustic measurement of overall voice quality: A meta-analysis, *The Journal of the Acoustical Society of America*, **126**(5): 2619–2634, https://doi.org/10.1121/1.3224706.
40. Maryn Y., Weenink D. (2015), Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index, *Journal of Voice*, **29**(1): 35–43, https://doi.org/10.1016/j.jvoice.2014.06.015.
41. Montalbaron M.B. *et al.* (2023), Presumptive diagnosis in tele-health laryngology: A multi-center observational study, *The Annals of Otology, Rhinology, and Laryngology*, **132**(12): 1511–1519, https://doi.org/10.1177/00034894231165811.
42. Nawka, T., Anders, L., Wendler, J. (1994), The auditory assessment of hoarse voices according to the RBH system [in German], *Sprache, Stimme, Gehör*, **18**: 130–133.
43. Nemr K. *et al.* (2012), GRBAS and Cape-V scales: High reliability and consensus when applied at different times, *Journal of Voice*, **26**(6): 812.e17–218.e22, https://doi.org/10.1016/j.jvoice.2012.03.005.
44. Parsa V., Jamieson D.G. (2001), Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech, *Journal of Speech, Language, and Hearing Research*, **44**(2): 327–339, https://doi.org/10.1044/1092-4388(2001/027).

45. Patel R.R. *et al.* (2018), Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function, *American Journal of Speech-Language Pathology*, **27**(3): 887–905, https://doi.org/10.1044/2018_ajslp-17-0009.

46. Portney L.G., Watkins M.P. (2009), *Foundations of Clinical Research: Applications to Practice*, 3rd ed., Pearson/Prentice Hall Upper Saddle River, NJ.

47. Quinlan J.R. (1999), *C4.5: Programs for Machine Learning*, Morgan Kaufman.

48. Reynolds V. *et al.* (2012), Objective assessment of pediatric voice disorders with the acoustic voice quality index, *Journal of Voice*, **26**(5): 672.e1–372.e7, https://doi.org/10.1016/j.jvoice.2012.02.002.

49. Roper T.A. (2014), Clinical Skills, 2nd ed., Oxford University Press.

50. Rosłanowski A. (2008), *Phoniatric database* [in Polish: *Baza nagrań foniatrycznych*], B.Eng., Polish-Japanese Academy of Information Technology.

51. Speyer R. *et al.* (2010), Maximum phonation time: Variability and reliability, *Journal of Voice*, **24**(3): 281–284, https://doi.org/10.1016/j.jvoice.2008.10.004.

52. Suvvari T.K. (2023), The role of Artificial Intelligence in diagnosis and management of laryngeal disorders, *Ear, Nose & Throat Journal*, https://doi.org/10.1177/01455613231175053.

53. Szklanny K. (2019), Acoustic parameters in the evaluation of voice quality of choral singers. prototype of mobile application for voice quality evaluation, *Archives of Acoustics*, **44**(3): 439–446, https://doi.org/10.24425/aoa.2019.129257.

54. Szklanny K., Wrzeciono P. (2019), Relation of RBH auditory-perceptual scale to acoustic and electroglottographic voice analysis in children with vocal nodules, *IEEE Access*, **7**: 41647–41658, https://doi.org/10.1109/ACCESS.2019.2907397.

55. Tadeusiewicz R. (1988), *Speech Signal* [in Polish: *Sygnał mowy*], Wydawnictwa Komunikacji i Łączności, Warszawa.

56. Tirronen S., Javanmardi F., Kodali M., Reddy Kadiri S., Alku P. (2023), Utilizing Wav2Vec in database-independent voice disorder detection, [in:] *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 1–5, https://doi.org/10.1109/ICASSP49357.2023.10094798.

57. Uloza V., Petrauskas T., Padervinskis E., Ulozaité N., Barsties B., Maryn Y. (2017), Validation of the acoustic voice quality index in the Lithuanian language, *Journal of Voice*, **31**(2): 257.e1–257.e11, https://doi.org/10.1016/j.jvoice.2016.06.002.

58. Verde L., De Pietro G., Sannino G. (2018), Voice disorder identification by using machine learning techniques, *IEEE access*, **6**: 16246–16255, https://doi.org/10.1109/access.2018.2816338.

59. Verikas A., Gelzinis A., Bacauskiene M., Uloza V. (2006), Towards a computer-aided diagnosis system for vocal cord diseases, *Artificial Intelligence in Medicine*, **36**(1): 71–84, https://doi.org/10.1016/j.artmed.2004.11.001.

60. Wilson J., Webb A., Carding P., Steen I., MacKenzie K., Deary I. (2004), The voice symptom scale (VoiSS) and the vocal handicap index (VHI): A comparison of structure and content, *Clinical Otolaryngology & Allied Sciences*, **29**(2): 169–174, https://doi.org/10.1111/j.0307-7772.2004.00775.x.