Archives of Acoustics Vol. ***, No. ***, https://doi.org/10.24423/archacoust.2026.4371

Online First August 25, 2026

Technical Note

Implementation of a COVAREP-Integrated Application for Voice Analysis Using LLMs

Hubert Daniszewski[(1)], Krzysztof Szklanny[(1)*], Piotr Wrzeciono[(2)]

[(1)] *Multimedia Department, Polish-Japanese Academy of Information Technology* Warsaw, Poland

[(2)] *Institute of Information Technology, Warsaw University of Life Sciences* Warsaw, Poland

*Corresponding Author: kszklanny@pjwstk.edu.pl

Received November 7, 2025; revised April 26, 2026; accepted May 24, 2026; version of record August 25, 2026; published issue XXXX.

This paper presents VocalCheck, a mobile application designed to monitor voice conditions using recordings made with a smartphone. The initial phase involved a literature review and an evaluation of existing voice processing applications, which informed the project's design requirements. Based on this analysis, an application was developed to record and preliminarily analyze voice samples using parameters provided by the Collaborative Voice Analysis Repository (COVAREP) for the Speech Technologies Toolkit. Additionally, an experimental study was conducted to assess the effectiveness of large language models (LLMs), specifically Microsoft Copilot and ChatGPT, in generating MATLAB scripts for computing voice quality parameters. Although these tools proved helpful for general programming tasks, their performance in speech signal processing was unsatisfactory.

Keywords: voice quality, voice diagnostics, mobile app, speech processing, LLM.

1. Introduction

Voice disorders affect individuals across various professions. Teachers, in particular, are highly susceptible to excessive strain on their vocal apparatus. Similar challenges can also affect choristers, singers, actors, public speakers, lecturers, and fitness instructors. Health factors, such as gastroesophageal reflux disease and rare genetic conditions like Pompe disease or Morquio syndrome type A (MPS IVA) (Hendriksz *et al.*, 2014), also affect the quality of the voice. Children may experience voice problems caused by excessive crying or screaming, which can lead to the development of vocal nodules. The project aimed to develop a tool to detect voice abnormalities, enabling the maintenance of vocal hygiene through preventive measures. As part of this goal, the mobile application VocalCheck was created, inspired by a prototype described in (Szklanny, 2019).

2. Literature review and existing solutions

The topic of using smartphones in voice therapy has garnered significant interest, which aided in the decision to choose the type of device for the VocalCheck application. The available hardware solutions included desktop computers, laptops, tablets, smartphones, and smartwatches. According to Google Scholar, approximately 15 800 scientific papers on this topic were published between 2015 and 2025 (Fig. 1).

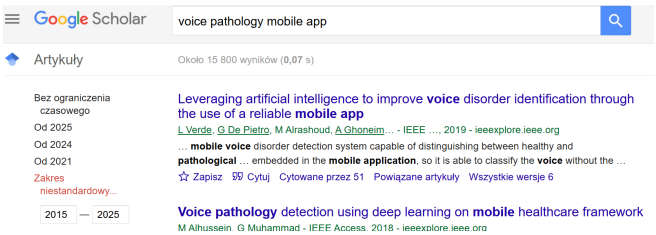

FIG. 1. Search results for 'voice pathology mobile app' on Google Scholar.

Descriptions of research on this subject are drawn from scientific journals such as the *Journal of Voice*, *European Archives of Oto-Rhino-Laryngology*, *IEEE Access*, and *Archives of Acoustics*. Many of the articles presented in these journals address issues related to devices and conditions for voice recording, as well as the support of machine learning tools. In the following, we describe some papers that utilize various devices in human voice diagnostics.

In (MANFREDI *et al.*, 2017), the usefulness of smartphone recordings was compared with results obtained under laboratory conditions. The experiment compared two phone models: HTC One and Wiko CINK SLIM2. Recordings from these devices were compared with recordings from the Sennheiser MD421U microphone, which, according to the mentioned article, is commonly used in acoustic laboratories. Measurements of jitter, shimmer (BOERSMA, 2001), and HNR (YUMOTO *et al.*, 1982) parameters for the phonation of the vowel /: a/ were strongly correlated with each other for all devices. The study demonstrated that even inexpensive smartphones can reliably record voices with pathological changes, comparable to those of professional laboratory microphones.

The authors of the article (ULOZA *et al.*, 2015) used natural phonation to compare values of fundamental frequency (f_0) (COOPER, SORENSEN, 1981), normalized noise energy (NNE) (KASUYA *et al.*, 1986), and signal-to-noise ratio (SNR) (KLINGHOLZ, 1987) obtained for recordings from a Samsung Galaxy Note 3 phone and an AKG Perception 220 microphone. The calculated parameters, except for f_0 and shimmer differed statistically between phone and microphone recordings. The study described in the mentioned article used a classifier based on a random forest model (HO, 1995), which distinguished between healthy and pathological voices with an accuracy of 73.7 % for phone recordings and 79.5 % for microphone recordings.

Similar binary classifiers were described in (VERDE *et al.*, 2019) (accuracy at 84.5 %) and (ASCI *et al.*, 2020). The authors of the latter achieved an 88.5 % accuracy in distinguishing the age of the speaker regardless of gender, using a set of 20 most significant acoustic parameters, including SNR, Mel-frequency cepstrum coefficients (MFCC) (ZHENG *et al.*, 2001), jitter, shimmer, and smoothed cepstral peak prominence (CPPs) (HILLENBRAND, HOUDE, 1996), selected from over 6000 parameters describing the voice. The mentioned article noted that age affects the human voice and that speech analysis is possible using recordings made in home conditions, without specialized equipment or acoustically adapted rooms.

In another study (ALHUSSEIN, MUHAMMAD, 2019), a voice pathology detection system based on parallel convolutional neural networks (CNNs) was presented, achieving 94.9 % accuracy when trained on the Saarbrücken Voice Database (SVD) and was tested in the Massachusetts Eye and Ear Infirmary (MEEI) database. When both the training and test sets were drawn from SVD, the accuracy was 95.5 %. The system was also able to classify three types of pathologies (unilateral paralysis, polyps, and cysts of the vocal folds) with accuracies of 95.1 %, 95.9 %, and 99.7 %, respectively.

In addition to diagnosing changes in the vocal folds, such as polyps and cysts, studies also focus on other voice disorders. SZKLANNY *et al.* (2016) described the process of monitoring the voice condition of individuals with late-onset Pompe disease. This disease, in its advanced form, often causes speech disorders by negatively affecting the work of nerves and muscles in the respiratory system. Electroglottographic (EGG) recordings were analyzed using the Collaborative Voice Analysis Repository (COVAREP) (DEGOTTEX *et al.*, 2014) repository and PRAAT (BOERSMA, 2001). These analyses provided detailed information on the functioning of the vocal tract, consistent with the phoniatric diagnosis. All nine participants with the juvenile form of the disease were diagnosed with voice abnormalities, and 7 out of 10 patients with the mature form of the disease were found to have dysphonia.

Dysphonia is characterized by hoarseness, abnormal timbre, loudness, and duration of the utterance (Kosztyła-Hojna *et al.*, 2014).

The continuation of the study on Pompe disease patients was described in (Szklanny, Tylki-Szymańska, 2018). Fifteen patients with the mature form of the disease underwent re-evaluation three years after the initial study. All patients participated in enzyme replacement therapy. Analysis of acoustic and EGG measurements revealed that the patient's voice condition had deteriorated. An increased degree of vocal fold insufficiency and increased vocal tension were also observed. Three years after the previous study, the average value of the closed quotient (CQ H) (Howard, 1995) parameter decreased from 0.345 to 0.307, which, with a p-value of 0.025, was considered statistically significant. Similarly, the average value of peak slope (PS) among patients decreased from −0.304 to −0.372 (Kane, Gobl, 2011). The article indicates that parameters such as CQ H and PS not only allow for the detection of voice disorders but also for tracking the development of diseases that negatively affect the voice.

The same signal analysis methods were applied in (Szklanny *et al.*, 2018). This rare genetic disease negatively affects the larynx and vocal tract. Due to the accumulation of glycosaminoglycans (GAG), MPS IV A can cause acute obstruction of the upper airways and deformities of the laryngeal vestibule, vocal folds, and posterior laryngeal mucosa. In individuals with MPS IV A, an increased value of the PS parameter was observed compared to the control group, indicating a breathy voice, a decreased value of the CPP parameter, characteristic of dysphonia, and a low value of HNR, which is common in a hoarse voice. Shimmer and jitter were elevated compared to the control group, indicating instability in vocal fold vibration.

Szklanny *et al.* (2019) described a similar process of analyzing EGG recordings in the diagnosis of vocal nodules in children. Vocal fold nodules are typically caused by voice overuse or vocal hyperfunction, and their symptoms include persistent hoarseness, a disturbance in vocal fold vibrations that results in turbulent airflow in the glottis, manifesting itself as a raspy, rough voice. The study involved 57 patients aged 5 to 14 years who had been diagnosed with vocal nodules. CQ H, PS, normalized amplitude quotient (NAQ), and CPP were used to determine voice quality. An elevated PS value of −0.19, compared with to –0.28 in the control group consisting of 37 healthy children aged 7.5 to 16 years, indicated a breathy voice. Tension in the voice was detected in 14 patients, signaled by an elevated NAQ value of 0.14. For the control group, NAQ was 0.12. CPP allowed for the detection of dysphonia in 30 subjects, with an average value of 11.64 among the sick and 11.66 in the control group. The average CQ H was 0.31 in sick children and 0.41 in the control group, indicating vocal fold insufficiency in patients with vocal nodules.

The study of vocal nodules in children was expanded with the RBH scale in (Szklanny, Wrzeciono, 2019b). It described the Voice Pathology Analysis Database (VPADB), which was created as part of this study. It contains recordings of voices with vocal nodules, data obtained using an electroglottograph, and perceptual assessments of roughness, breathiness, and hoarseness (RBH) scale (Nawka *et al.*, 1994). Data from healthy voices from the SVD database were used for the control group. The assessments were made based on the opinions of two independent experts in voice analysis. To determine the RBH scale values for a given voice sample, the parameters PS, NAQ, and CQ H were used. Each of the three RBH features was analyzed for three verdicts: 0 for healthy individuals (denoted as R0*, B0*, and H0*), 0 for individuals with vocal nodules (R0, B0, H0), and 1 for individuals with vocal nodules (R1, B1, H1). Using ANOVA, it was found that CQ H and NAQ allow distinguishing among R0*, R0, and R1, as well as B0*, B0, and B1. For the distinction between H0*, H0, and H1 values, the PS parameter was also needed, although it was considered significantly less effective. The results of the voice quality tests agreed with the assessments of both experts. A further supplement to the study of vocal nodules was described in (Szklanny, Wrzeciono, 2019a).

As part of this manuscript, a classifier based on a genetic algorithm (Holland, 1975) was prepared, which diagnosed the degree of vocal fold insufficiency based on voice samples and electroglottograph recordings. The final form of the classifier used the following parameters: CQ H, PS, NAQ, and RDS. The accuracy of the classifier was evaluated at 92 % for the training set and 94.73 % for the validation set. The high effectiveness of the obtained classifier, combined with the possibility of its representation using arithmetic functions understandable to humans, enabled its implementation in the LibreOffice package using Visual Basic for Applications (VBA), making it available, for example, in spreadsheets.

The literature also describes the impact of voice pitch on subglottal pressure values and the NAQ parameter. In (Björkner *et al.*, 2005), professional opera artists were asked to perform a diminuendo, gradually softening their voice by reducing the sound level while singing the syllable /pae:/. The recording was repeated three times, each performed at 25 % and 75 % of the vocal range of the subjects, expressed in the number of semitones. Airflow calculations were performed using the Rothenberg mask. The subglottal pressure was measured by connecting a plastic tube, held in the corner of the singers' mouths, to a subglottal pressure sensor. It was observed that while subglottal pressure varies among singers, it is strongly correlated with their fundamental frequency component, f_0. In the case of high tones, it was about twice as high as in recordings where the same singers used lower tones. After normalizing the subglottal pressure relative to the main component, it was observed that as the normalized measurement increased, the NAQ value decreased. This parameter decreases significantly in the case of low pressure. It was concluded that this illustrates the manner of phonation, which in the case of professional opera artists is independent of the sound level and pitch of the sung tones.

The aforementioned acoustic parameters, part of the COVAREP toolkit, are reliable and enable a detailed analysis of voice quality. The COVAREP toolkit is currently not being developed; its maintenance is important for conducting further experiments, which is part of our article.

Many experiments are conducted in acoustic cabins to minimize the influence of the acoustic background. The acceptable level of ambient noise to obtain reliable measurements was described in (Lebacq *et al.*, 2017). In this experiment, the maximum acceptable level of ambient noise that does not affect the values of jitter [%] and NHR was determined. Voice samples were generated by a synthesizer described in (Manfredi *et al.*, 2017). Artificial recordings differed in terms of phonation type, f_0 frequency, and the level of jitter and noise to simulate the speaker's voice condition. The ambient noise used in the experiment was described as 'shopping small ambient.' In the conclusions, the researchers stated that using smartphones to measure NHR and f_0 perturbations is possible, provided that the maximum level of ambient sounds does not exceed 50 dBA and that voice samples are limited to the phonation of the vowel /a/. In relation to the article, it should be noted that parameters such as PS, NAQ, and H1–H2, H1–A3 are weakly sensitive to low SNR levels. According to the mentioned publication, the required SNR level is 10 dB. This corresponds to a situation where the ambient noise level is 50 dBA, and the phonation level is 60 dBA.

In (Uloza *et al.*, 2023), the influence of ambient noise and the device used on the values read by PRAAT was presented. The level of noise added to the recordings averaged 29.61 dB, and its SNR, compared to voice recordings, was 38.11 dB. To compare the acoustic voice quality index (AVQI) calculated for recordings made with an AKG Perception 220 microphone and an iPhone 6s smartphone, a Bland–Altman plot was used. This is a scatter plot of two variables, where the horizontal axis represents the arithmetic mean of both measurements, and the vertical axis represents the difference between these measurements. It aims to illustrate the difference between the two measurements relative to the size of the measured values (Kaur, Stoltzfus, 2017). Despite a p-value of 0.07, the study's authors considered the differences in recordings from the microphone and the phone with added noise to be not statistically significant. Additionally, they emphasized that the obtained Bland–Altman plots for all three versions of the recordings did not demonstrate a systematic difference in the calculated AVQI. Therefore, according to the study's authors, smartphones can be an acceptable tool for recording voice for therapeutic purposes.

In the literature on voice diagnostics using smartphones, there are also studies examining the usefulness of such applications. Angadi *et al.* (2023) described the MyVocalHealth application, which was only available during the study described in that paper. The application aimed to motivate users to participate in voice therapy exercises on a regular basis. At the end of the study, data showed that application users performed about 50 % more than the control group, which did not use the application. In addition, the study's authors placed special emphasis on usability tests, i.e., tests in which users had to perform a given task using the application independently. Seventeen participants took part in them, of whom over 95 % performed them correctly. The most common mistake was skipping or failing to pause the instructional video attached to the exercise. Thirteen of the test participants stated that the application was easy to use, and seven noted that the instructional videos were helpful. The most significant reservations regarding usability were problems with the button leading to the next

task and the video. 12 respondents answered a survey enabling the calculation of the application's score on the system usability score (SUS) scale. The application scored 78.13 points, exceeding the average rating of 68 points for web systems considered usable. After conducting the survey, six testers participated in a group discussion, during which the following suggestions for improving the quality of the application were formulated:

1. The application should contain a clear indication showing that the exercise has been completed.
2. A modifiable system of reminders and snooze functions would help in regularly participating in exercises.
3. The best result in the exercises should be displayed at the top of the screen. The application only displayed a chart of chronologically performed exercises over the last 7 and 30 days.

The respondents stated that the application was most often used at home, in the car, in the parking lot, and in restaurants. Most respondents performed the exercises according to the recommendations, in the morning and evening, but some also performed them during their lunch break. The MyVocalHealth application did not calculate any voice-related parameters; it only measured the phonation duration. However, because for the study of the application's usability, it serves as a good reference point when designing applications that support voice diagnostics.

In summary, numerous studies have been conducted on the use of recordings in voice analysis. Many of them allow for a comparison of the effectiveness of using different types of recording devices, ranging from specialized ones such as electroglottographs to professional microphones and commonly available mobile phones. In analyzing this topic, a variety of measures are used to assess voice condition. Some of the parameters used, such as CQ H, require the use of an electroglottograph; however, there are also many others, including PS, NAQ, NHR, CPP, AVQI, jitter, and shimmer, which can be calculated based on audio recordings. They enable detection of vocal fold insufficiency and allow observations of changes in the voice caused by various factors, such as vocal strain or aging. Thanks to the PRAAT program and the COVAREP toolkit (Degottex *et al.*, 2014), there are already ready-made implementations of algorithms calculating the values of useful parameters. The conditions under which the recording is made are also important, as well as what is being recorded. Some studies utilize soundproof rooms or soundproof booths, while other recordings to be made in standard rooms or public settings, such as restaurants or hospitals. Attention should be paid to the distance between the device and the speaker, the angle of its inclination, and the level of ambient noise. It is important that the phone is far enough from the speaker's mouth and that the level of background noise does not exceed 50 dB. Study participants were often asked to phonate the vowel /a/ or read a sentence that allowed the analysis of intonation changes in the voice. Phonation, due to its simplicity, is understandable to many user groups and is effectively processed by existing algorithms analyzing audio signals. Many of the cited articles affirm the use of mobile phones for diagnostic purposes and present real applications supporting voice therapy. Both machine learning and traditional statistical methods achieve high accuracy in distinguishing between healthy and pathological voices. Artificial intelligence methods, such as machine learning, have already been used in voice diagnosis in studies by Verikas *et al.* (2006), Crowson *et al.* (2020), and Suvvari (2023). Furthermore, Compton *et al.* (2022) employed a deep learning approach, and Tirronen *et al.* (2023) utilized transfer learning.

The literature review in this area also provided important information on how to design such tools to be user-friendly and useful for the target audience. Attention was paid, among other things, to the clarity of instructions and the visualization of changes in voice condition assessment over time, which allows tracking the therapy progress.

In summary, numerous studies confirm the usefulness of audio recordings as a tool for analyzing and assessing voice conditions. The literature compares the effectiveness of various types of recording devices – from specialized devices such as electroglottographs to high-quality laboratory microphones and commonly available smartphones. The use of these devices enables the determination of a wide range of acoustic parameters, such as NAQ, PS, NHR, CPP, AVQI, which provide significant information on the quality of phonation and enable the detection of abnormalities such as vocal fold insufficiency. Some parameters, such as closed quotient, require an electroglottograph. Still, many of them can be calculated solely based on acoustic recordings using available tools such as the COVAREP (Degottex *et al.*, 2014) computational package or the PRAAT program.

However, the effectiveness of the analysis depends not only on the selection of parameters and algorithms but also on the conditions under which the recordings are made. Key factors include the microphone's distance from the speaker's mouth, its angle, and the level of background noise – it is recommended that background noise not exceed 50 dB. Depending on the research design, recordings are made in both controlled conditions (e.g., soundproof booths) and natural environments, such as living rooms, restaurants, or medical facilities. Such a variety of conditions enables the assessment of the resistance of analytical algorithms to interference, bringing them closer to clinical applications and daily therapeutic practice.

In studies, tasks that involve prolonged phonation of the vowel /:a/ or reading a short sentence are most often used. Prolonged phonation is a straightforward and intuitive method for obtaining voice data, that is understandable and feasible for a wide range of users, while simultaneously providing a signal of sufficient length and stability for a reliable acoustic analysis.

The results of numerous scientific papers suggest that both classical statistical methods and modern machine learning techniques achieve high accuracy in distinguishing healthy from pathological voices, based on recordings made using a mobile phone. Importantly, the literature emphasizes the need to design applications that support diagnosis and therapy, including a clear user interface, comprehensible instructions, and visualization of the patient's progress. These elements are crucial to the acceptability and usability of such tools in clinical practice and daily monitoring of voice conditions.

3. Existing applications

The process of preparing for the project involved testing selected applications thematically related to voice processing. The choice of applications was limited to those that are functional and have support. This section describes five applications, including the functionalities they provide, along with a list of identified advantages and disadvantages for each solution. The first of the tested applications is the Sound Meter (Decibel Meter)[1]. The central part of the application's screen displays the sound level in dB. Below it, depending on the settings, there is a detailed graph of the signal level readings over time or a list of sound level thresholds with examples of corresponding sources. At the bottom of the screen, control buttons are available to perform various functions, including calibration, resetting and pausing the measurement, and changing the color scheme. The main advantages of this tool is its simple and intuitive interface; however, it does not provide functions sufficient for assessing the condition of the voice.

The Vocal Pitch Monitor[2], offers a more advanced method of measurement. This program is designed for individuals involved in music, so some of its features may be incomprehensible to the average user. The main panel of the tool displays the received pitch and the voice pitch in real time. The application enables recording, playback, and sharing of recordings in the WAV format. However, this functionality does not always work correctly, as the application cuts off fragments of recordings and does not allow for their free scrolling. The application offers a wide range of options, allowing users to adjust, among other things, the graph zoom, select the format of note names (letters and solmization), display frequencies, and change the colors of interface elements on the RGB scale.

The Voice Tools application[3], is designed for transgender individuals, and its main purpose is to help them practice speech to affirm their gender identity. However, the exercises and tools available in it can allow for a deeper analysis of audio signals than the previous solutions. It consists of five modules: pitch, which allows to see a graph of the frequency reading of a sample sentence; tone, containing a list of buttons playing synthesized sounds in different tones; a panel similar to pitch, volume, with a graph of the sound level; analyse, which shows a longer fragment of text to be read aloud, and then displays statistical data regarding frequency, signal level, and ambient noise; and the 'more' button, which provides access to statistics from the entire time of using the tool, application settings, and a spectrogram generated in real time.

[1] Available on Google Play at: https://play.google.com/store/apps/details?id=com.gamebasic.decibel.

[2] Available on Google Play at: https://play.google.com/store/apps/details?id=com.tadaoyamaoka.vocalpitchmonitor&hl=en_IN&gl=US.

[3] Available on Google Play at: https://play.google.com/store/apps/details?id=com.DevExtras.VoiceTools&hl=en_US.

The aforementioned options – in addition to visual customization – include settings that help maintain vocal health, such as a reminder to drink water before vocal exertion and support for pronunciation practice. The user can choose the difficulty level of the sentences that will be suggested by the application and add his or her own sentences. An advantage of the program is also the presence of a tutorial in the form of dialog boxes with clear instructions. The biggest limitation of Voice Tools is the lack of support for analyzing pre-recorded audio, it supports only live recording, with a maximum duration of 120 s.

Loud and Clear for Parkinson's[4] is a solution designed for practicing diction and is specifically aimed at individuals with Parkinson's disease. Its creator is a speech therapist and graduate of the University of Texas at Dallas. He also obtained a certificate from the American Speech-Language-Hearing Association (ASHA). The therapy involves doing warm-up exercises, daily exercises, and homework assignments. The tasks in the application are varied, ranging from simple phonations of individual sounds to reading simple sentences, and include language puzzles and answering questions that simulate conversation (e.g., about a favorite ice cream flavor). Before the exercises, a short video is displayed showing how to perform them. The exercise itself is accompanied by a meter that indicates whether the user should speak louder, softer, or maintain the current sound level. The application also contains educational resources on Parkinson's disease, suggested therapists, information on the length of the daily exercise series with achievements, and sequences of commands with a voice recording function, which allows the user to monitor progress in diction training. The tool uses gamification elements, including a series of daily exercises and achievements, to encourage regular therapy. The gamification process involves introducing elements that make a given experience similar to a game. Such an approach to application design is explored, among others, in education. Among the negative aspects of the experience with this program, one can mention the lengthy module load times compared to other solutions and the random requests to access sensitive data that the application does not require for proper operation, such as multimedia files stored on the device or location data. It also does not provide any precise data, for example, regarding the voice level in dB.

The last application tested as part of preparing to create our tool is AI SoundLab[5]. It requires registration, during which the user can provide several medical data, including information regarding voice and respiratory system diseases. AI SoundLab consists of submodules called 'studies,' access to which requires entering codes provided by authorized entities. During the tests, the publicly available COVID-19 study, as well as the Health AI Demo and Respiratory AI Demo studies, were reviewed. The first of these is a survey during which, in addition to answering questions about health status, the user performs several voice tasks, including recording their cough, laughter, phonations of sounds, and reading text. Health AI Demo involves making a 30 s to 60 s voice recording of any content, after which the application evaluates the detected sentiment in it. The result is presented as a point on a graph, where the horizontal axis is labeled 'positivity' and the vertical axis is labeled 'activation.' The Respiratory AI Demo is a study that evaluates voice health. After recording coughs, reading a text fragment, and a five-second phonation of the vowel /i/, the user receives an assessment based on four metrics:

- respiratory health: detection of respiratory system disease symptoms, such as wet and dry cough and sore throat, calculated by a deep neural network,
- vocal strength: voice strength measured by the energy of high frequencies of the signal,
- speech rate: articulation speed, measured by the average distance between speech spectra adjacent in time,
- voicing control: the speaker's control over the vocal cords, measured by the average length of speech fragments.

In summary, AI SoundLab is the most complex of all the tools tested as part of this project, allowing for the acquisition of valuable information regarding the design of applications focused on voice analysis. The application highlights the importance of clear instructions for performing tests and serves as a reference point for the user interface's graphic design. The quality of solutions in AI SoundLab largely depends on the person designing the study. Still, the application also contains noteworthy universal elements, such as the ability to record and analyze audio samples.

[4]Available on Google Play at: https://play.google.com/store/apps/details?id=com.ateamo.loud_and_clear&hl=en_US.
[5]The application is available online at: https://aisoundlab.audeering.com/welcome.

All the tested applications contributed to determining the key functions of the target project and to preparing the user interface based on the expectations of the target group.

4. VocalCheck application

After testing applications on similar topics and establishing the project's main assumptions, a mobile application called VocalCheck was created. The application's backend was written in Java with the Spring Boot framework, designed for building such applications. The frontend of the project was implemented using HTML pages, Cascading Style Sheets (CSS), and JavaScript. When attempting to access the application without logging in, it displays a panel with a button that allows login using a Google account (Fig. 2). The primary advantage of using the Google platform is that the application administrator does not need to provide a registration panel or manage user data independently, which may be considered sensitive under certain countries' legal regulations. For example, Article 4(1) of the General Data Protection Regulation (GDPR), adopted on April 27, 2016, defines personal data as information that allows direct or indirect identification of a natural person. Data related to voice may be classified as a special category of personal data under Article 9 of the regulation, as it may reveal the ethnic origin of a person or the diseases they are struggling with. Thanks to the Google Cloud API, all data regarding the voice test history and personal user data, such as gender, is stored directly in the application's files on the logged-in user's Google Drive. This avoids complications related to storing sensitive information on one's servers.

Fig. 2. VocalCheck application login panel.

The screen for recording a voice sample (Fig. 3), which will be evaluated, allows for two ways of providing the recording:

- recording a five-second phonation of the vowel /a/ in real time,
- providing a wave file with a voice recording from the device.

Rather than regular speech, sustained vowel phonation has been chosen since it is the traditional form of voice testing (Askenfelt, Hammarberg, 1986). The drawback of this approach is that it does not represent the natural form of speech in a standard spontaneous setting (Parsa, Jamieson, 2001).

Pressing the microphone icon starts the recording. Before the signal is saved, the user must grant the web browser access to the microphone; otherwise, an error message will be displayed. During recording, the sound wave is shown in the center of the screen as a waveform. Below the graph, buttons are available to cancel the recording (red cross) or end it before the allotted time expires (green checkmark) (Fig. 4). The application checks whether the recording is clipped and whether it has the appropriate signal strength. Clipping occurs when the input signal voltage exceeds the range of values that a given device can process. This results in the excess signal value being 'cut off,' introducing additional harmonic components that are audible as a characteristic distortion in the sound. Using the Web Audio API in JavaScript, the application detects the time-domain moment at which

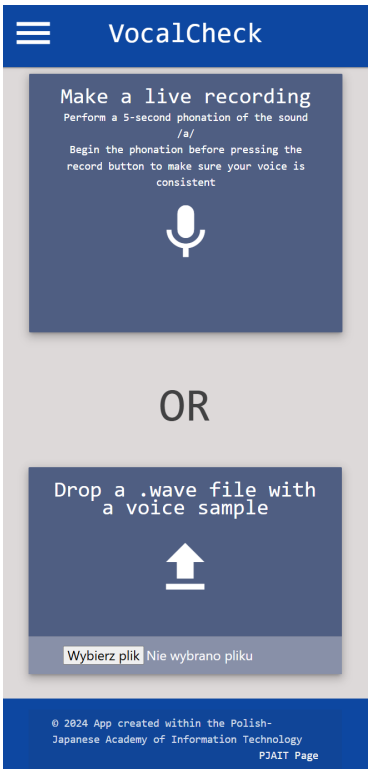

FIG. 3. Voice sample collection panel in the VocalCheck application.

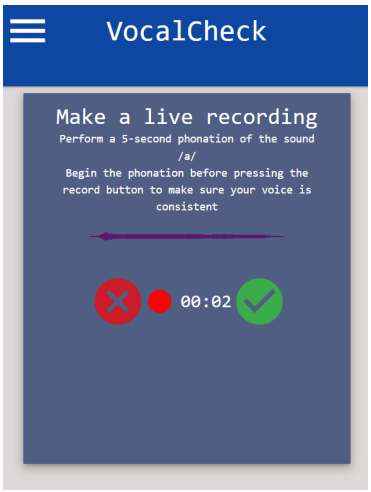

FIG. 4. Recording user voice in the VocalCheck application.

the signal amplitude exceeds its maximum value and displays an error message prompting the user to check their microphone settings.

To additionally detect whether the signal strength corresponds to the sufficient strength needed for calculations, the power of each sampled signal is also counted. This was done using Eq. (1). The total signal level is compared with the −20 dB threshold. This is the minimum signal level required to obtain correct calculation results. This level was determined empirically during application tests:

$$P = \frac{1}{N}\sum_{n=0}^{N-1}|f(n)|^2, \tag{1}$$

$$L = 10\log_{10} P. \tag{2}$$

After submitting the voice sample in the form of a live recording or a wave file, functions from the COVAREP (Degottex *et al.*, 2014) repository are launched, calculating the following parameters:

- PS: calculated by a function from the file *glottalsource/peakslope.m*,
- CPP: calculated by a function from the file *glottalsource/cpp.m*,
- NAQ: calculated by a function from the file *glottalsource/get_vq_params.m*.

The PS is defined as the average decrease in amplitude (in dB) between successive spectral peaks within a specified frequency range of the magnitude spectrum of a speech frame. It reflects the spectral sharpness and correlates with voice quality features such as breathiness and tenseness. It allows for the detection of breathy and strained voice. The formula for calculating PS was taken from (Kane, Gobl, 2017). The PS is calculated by observing the wavelet decomposition, given the following formula for the mother wavelet:

$$g(t) = -\cos(2\pi f_n t) \cdot \exp\left(\frac{t^2}{2\tau^2}\right), \tag{3}$$

where f_n is $\frac{f_s}{2}$ for sampling frequency $f_s = 16\,\text{kHz}$, τ is $\frac{1}{2f_n}$.

The CPP is a measure calculated based on the cepstrum, which is the spectrum obtained by calculating the inverse Fourier transform for the logarithm of the spectrum of the audio signal. The cepstrum in the real number domain was described as follows:

$$C_r(q) = \mathcal{F}\left\{\log|S(f)|^2\right\}, \tag{4}$$

where $\mathcal{F}$ is the Fourier transform, $S^2(f)$ is the power spectrum of the signal $s(t) : S^2(f) = \mathcal{F}\{E[s(t) \cdot s(t-\tau)]\}$.

CPP is the difference between the maximum value on the cepstrum graph and the corresponding point of linear regression for the entire cepstrum. It is defined as the spectrum calculated from the spectrum, whose low value is usually correlated with a greater intensity of dysphonia (Patel *et al.*, 2018).

The NAQ is defined as the ratio between the maximum amplitude of the differentiated glottal flow and the product of the maximum negative amplitude of the glottal flow derivative and the fundamental period. It reflects how abrupt or smooth the glottal closing phase is during vocal fold vibration. Its ability to express the phonation type has been confirmed in the study by Airas and Alku (2007). NAQ was described by the following formula:

$$\text{NAQ} = \frac{\text{AQ}}{T}, \tag{5}$$

where AQ is the amplitude quotient, T is the fundamental period of the flow.

The transmitted audio signal is temporarily saved on the server's disk. Then, using the open-source environment Octave, which allows running scripts in the MATLAB language, the code is executed to compute the values of selected voice parameters. The PS and NAQ parameters allow for the classification of the processed voice into one of three categories:

1. Modal voice – neutral, healthy voice.
2. Strained voice (tense voice or pressed voice) – a voice condition resulting from increased tension in the vocal tract. It is characterized by complete and sharp closure of the vocal folds, accompanied by more intense articulation.
3. Breathy voice – the opposite of a strained voice. It results from low tension in all elements of the vocal tract and is characterized by incomplete closure of the vocal folds, accompanied by an audible inhalation sound.

Other parameters that could be used to distinguish between those categories are maxima dispersion quotient (MDQ) (Kane, Gobl, 2013), parabolic spectral parameter (PSP) (Alku *et al.*, 1997), harmonic richness factor (HRF) (Childers, Lee, 1991), H1H2 (Hanson, 1997), quasi-open quotient (QOQ) (Hacki, 1989).

In (Kane, Gobl, 2011) box plots were presented regarding the statistical values of PS and NAQ parameters for each of the three voice types. Based on the mentioned article, it was assumed that for modal voice, PS values should be in the range PS $\in [-0.19; -0.5]$. PS below -0.35 indicates a strained voice, while values above -0.19

indicate a breathy voice. The same article also contains a range of NAQ values. However, in (Alku *et al.*, 2002), healthy ranges of values for prolonged phonation of the vowel /a/ are additionally distinguished depending on the gender of the speaker. Based on data on the average NAQ value for each voice along with their standard deviations, the range of modal voice for women was set to NAQ $\in [0.14; 0.20]$, and for men to NAQ $\in [0.12; 0.25]$. Exceeding the upper limit of the range indicates a breathy voice, and exceeding the lower limit indicates a strained voice. The CPP parameter, as described in (Szklanny *et al.*, 2019), enables the detection of dysphonia at an early stage of development. In patients without dysphonia, the CPP value was recorded as 11.64 dB, and for patients exhibiting dysphonia, as 11.66 dB. Based on this, a binary assessment was adopted, according to which dysphonia can occur for CPP < 11.6 dB, and otherwise, it does not happen for this study group. To simplify the interpretation of results and make them understandable to a user unfamiliar with voice analysis, six different possible test results were adopted:

– modal voice (marked in green in app):

$$\text{PS} \in [-0.19; -0.35] \wedge \text{NAQ} \in [\text{NAQ\_MIN}; \text{NAQ\_MAX}],$$

– moderate breathy voice (marked in light blue):

$$(\text{PS} > -0.19 \wedge \text{NAQ} \in [\text{NAQ\_MIN}; \text{NAQ\_MAX}]) \vee (\text{PS} \in [-0.19; -0.35] \wedge \text{NAQ} > \text{NAQ\_MAX}),$$

– severe breathy voice (marked in dark blue):

$$\text{PS} > -0.19 \wedge \text{NAQ} > \text{NAQ\_MAX},$$

– moderate strained voice (marked in light red):

$$(\text{PS} < -0.35 \wedge \text{NAQ} \in [\text{NAQ\_MIN}; \text{NAQ\_MAX}]) \vee (\text{PS} \in [-0.19; -0.35] \wedge \text{NAQ} < \text{NAQ\_MIN}),$$

– severe strained voice (marked in dark red):

$$\text{PS} < -0.35 \wedge \text{NAQ} < \text{NAQ\_MIN},$$

– uncertain result (marked in gray):

$$(\text{PS} > -0.19 \wedge \text{NAQ} < \text{NAQ\_MIN}) \vee (\text{PS} < -0.35 \wedge \text{NAQ} > \text{NAQ\_MAX}),$$

where NAQ_MIN is 0.14 for women and 0.12 for men, and NAQ_MAX is 0.20 for women and 0.25 for men. In cases where only one parameter indicates a given voice type, the voice type is classified as a moderate variant. When PS and NAQ indicate it simultaneously, it is treated as a severe form. The result is uncertain if one parameter indicates a strained voice and the other a breathy voice.

All six possible classifications were presented on a 2D board in the voice test result panel in the application (Fig. 5). The shadow cast by the ball representing the phonation indicates its type. The verdict regarding dysphonia depends on CPP and is visualized through the color of the ball and its height relative to the board. The lower the ball is, the higher the CPP value. When the ball is positioned below the board and is red, it indicates

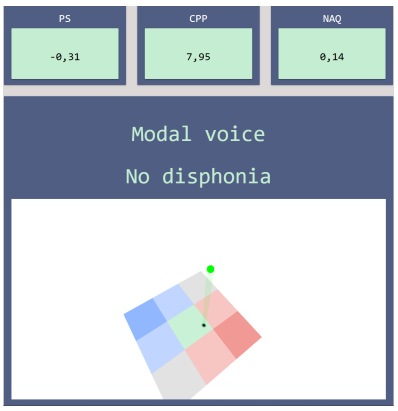

Fig. 5. Voice classification in the VocalCheck application.

that the application has detected the risk of dysphonia in the user's voice. The exact values of all parameters can be read by hovering the cursor over the ball. The backgrounds of the data blocks correspond to the voice conditions indicated by the displayed value.

In addition to the login panel, voice sample panel, and the results panel, VocalCheck includes a panel for viewing the history of voice tests. The data is stored in the form of a .csv file in the application's configuration files on the user's Google Drive and is updated each time the user conducts a test. After entering the Results History tab, the user is redirected to a page displaying a table showing PS, CPP, and NAQ values and the date of the tests (Fig. 6). The colors of the individual cells in the table correspond to the verdicts regarding the voice condition described in Sec. 3. Pressing a row in the table will open a visualization of the result.

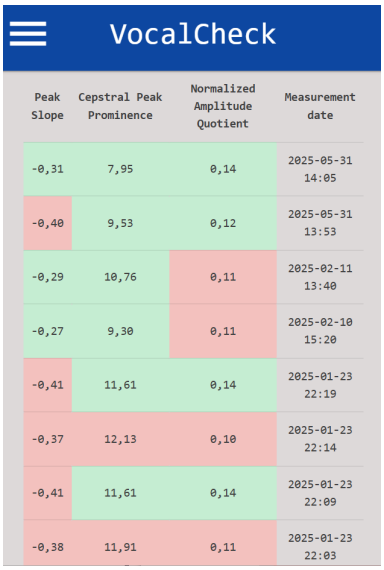

FIG. 6. VocalCheck voice sample test history.

The last tab, gender (Fig. 7), allows setting data regarding the user's gender. This is required because different NAQ ranges for the modal voice vary depending on gender. The gender panel is always open when the application cannot read this data from the configuration files on the user's Google Drive.

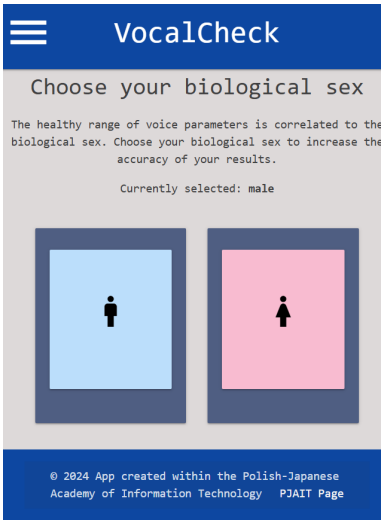

FIG. 7. VocalCheck gender selection panel.

Graphical elements, such as buttons, tables, and cards containing individual elements for providing voice samples and displaying results, were generated in the MaterializeCSS framework[6]. Its basis is the language and

[6]Data from the repository: https://github.com/dogfalo/materialize.

style of Material Design, an application design language created by Google. The goal of Material Design is to introduce a set of guidelines for designing mobile and web applications, including adding the depth to objects and using animations to create a responsive user interface. The use of MaterializeCSS enabled the preparation of a user interface that adapts to screen dimensions, allowing the application to be displayed on both desktop and mobile screens. The colors in VocalCheck were also selected in accordance with Material Design guidelines[7]. The final color palette was chosen based on usability tests, selecting color with the hexadecimal code #0d47a1, provided by MaterializeCSS under the name 'blue darken-4,' as the primary color.

The wavesurfer.js library, along with the wavesurfer.js/dist/plugins/record.esm.js plugin, was used to display the sound wave visualization, and to draw it during recording. This plugin requires manual modification to avoid automatic lowering of the volume. Prolonged phonation was, by default, detected by the getUserMedia() function from the MediaDevices web API class in JavaScript as noise. An attempt was made to use generative models, such as Microsoft Copilot and ChatGPT in version gpt-4o, to verify whether they can generate code that allows for the calculation of PS, CPP, and NAQ parameters.

VocalCheck has been tested by two experts in human-computer interaction, who also specialize in voice analysis. The test has been conducted by students of PJAIT and members of the author's personal network. In total, 15 individuals took part in usability tests, during which they were tasked to perform three actions:

1. Record a live voice sample and interpret the results.
2. Send a prerecorded sample and interpret the results.
3. Check the evaluation history to find the results from a specific date.

Test participants had to complete the tasks with minimal guidance, and each step they performed was recorded, timed, and further analyzed. After performing the tests, the participants were interviewed and asked for feedback. The tests lead to the following conclusion:

1. Recording a live voice sample took, on average, slightly longer than sending a prerecorded sample: 27.68 s versus 20.71 s:
   - during the interviews, 7 participants preferred recording their voices live, 3 preferred sending a prerecorded sample, and 5 had no strong opinion one way or the other,
   - only 2 participants used the drag-and-drop method to upload files, while the others used the system browser,
   - thirteen participants (those who were not familiar with the voice quality parameters) did not understand what each of the three parameters represents, and 7 asked about the meaning of terms such as dysphonia and modal voice.
2. Checking the evaluation history took an average of 16.26 s:
   - one participant suggested adding the ability to filter the result list by voice classification.

ChatGPT is a large language model developed by OpenAI, trained on a vast corpus of texts to provide accurate information on a wide range of topics. The GPT-4o version, where 'o' stands for 'omni,' is the latest version of this model available at the time of writing. It was released in May 2024 and introduces the ability to transmit combinations of text, audio, images, and recordings as input data to a single model. The Microsoft Copilot tool is powered by three AI components, including an older version of the ChatGPT model – GPT-4. In addition, Copilot utilizes the Prometheus module, a proprietary Microsoft set of techniques for working with machine learning models from OpenAI, designed to provide accurate, up-to-date, and targeted results while enhancing security, as well as the DALL-E 3 model for generating images from text. The ChatGPT-4o, ChatGPT-4, and DALL-E 3 models are based on the transformer architecture described in the paper 'Attention is all you need.'

As part of the experiment, both models were asked to generate scripts in Octave that would allow for the calculation of PS, CPP, and NAQ parameters using queries in the following format: How to calculate X when analyzing a voice recording in Octave? Where X is the full name of one of the three parameters, each query was

[7]The tool used for this purpose available at: https://material-foundation.github.io/material-theme-builder/.

performed as the first in a separate English chat. In case of errors during script execution, subsequent messages describing the problems were sent to the model until a functional code was generated. Additionally, after obtaining scripts returning values for the given parameters, information was provided that the result was different from the results provided by COVAREP, with a request to analyze the scripts from this repository, explain the cause of the differences in results, and modify the generated code so that the parameters were calculated according to the actual implementation. These queries were written according to the following format:

```
The output of the script is different from the COVAREP's implementation of the X calculation
algorithm. Please take a look at the script at: <LINK TO THE SCRIPT CALCULATING THE PARAMETER
IN COVAREP> and analyze what may be the cause of the output differences. If possible, adjust
your code so that it matches COVAREP's calculations.
```

All errors encountered during script execution, including those related to COVAREP, were also addressed in subsequent queries to the model. Changes were made to the generated code regarding the names of input files, functions that display results in the console, etc., to enhance the readability of the results. The values calculated by the scripts before and after considering the COVAREP repository were collected, compared with those calculated by the COVAREP scripts, and compiled together in Table 1. There are situations where the Microsoft tool can generate working code that gives the desired results. This occurs when the machine learning model has encountered the given problem multiple times during the training data processing.

Table 1. Comparison of voice parameter values depending on the source of implementation of counting functions.

	COVAREP	Microsoft Copilot	Microsoft Copilot after taking into account COVAREP	ChatGPT	ChatGPT after taking into account COVAREP
PS	–0.395743	0.0903	0.0062304	0.0002711	0
CPP	7.11601	0.78335	6.3649	0.045087	0.8329
NAQ	0.0946473	–0.077603	–0.30935	0.066899	0.27388

The operation of Microsoft Copilot was tested to generate a Java Swing fragment of code. The obtained code runs a graphical application with a text field and the text 'Hello World!'. As part of the query, specific guidelines were provided regarding the displayed window, including requirements to use a JTextArea component, place the text in an H1 header, bold it, use a red font color, and wrap it. The functional code was obtained after two queries were made. Attempting to place HTML tags in a component of the JTextArea class caused Microsoft Copilot to use the non-existent setEditorKit() method in this class. This method does exist, however, in the JEditorPane class, which the Microsoft system uses after receiving the compilation error. Running the final version of the code displays the window configured as expected (Fig. 8).

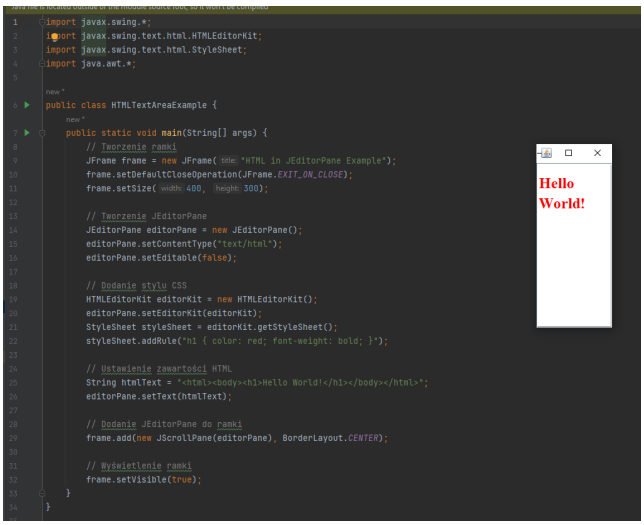

Fig. 8. Output of Java Swing code generated in Microsoft Copilot.

Repeating the experiment with the ChatGPT-4o model gave a different result. In response to the first query, the model correctly identified the inability to render HTML data and printed only the appropriate tags, using the text-wrapping function (Fig. 9). At the same time, the response concluded by explaining why using JTextArea in this case is impossible and suggested replacing it with JEditorPane or JTextPane. The second query, requesting the creation of a version of the program that renders the HTML code, resulted in the functional code using JEditorPane (Fig. 10). Potentially, the greater insightfulness of ChatGPT-4o may stem from it being a newer, more developed model than the ChatGPT-4 used in Microsoft Copilot.

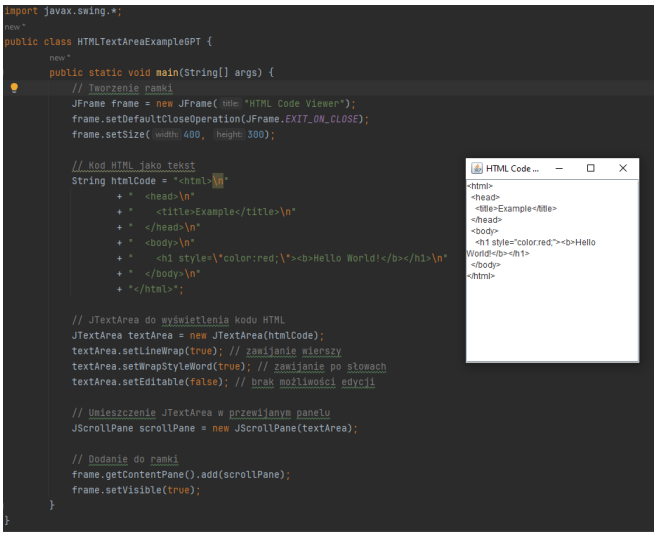

FIG. 9. Code generated by ChatGPT after the first query.

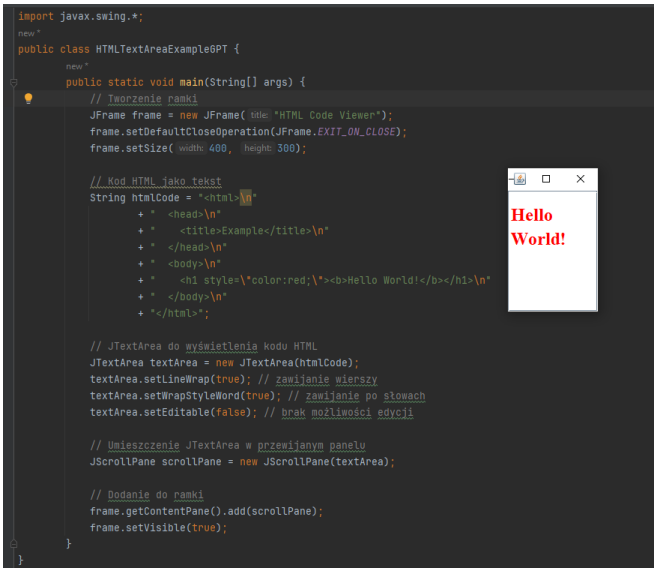

FIG. 10. Code generated by ChatGPT after requesting the HTML rendering version.

5. Discussion

5.1. Using the mobile diagnostics application

The application presented in this study, VocalCheck, was developed as a tool for the preliminary assessment of vocal health. It is intended to support early-stage voice diagnostics by providing accessible, smartphone-based analysis of vocal parameters. Contemporary approaches in computational voice assessment shape its design and aim to serve as a practical aid for singers, vocal coaches, and researchers in the field of voice science.

The underlying approach to automatic diagnostics in VocalCheck draws on several scientific studies that utilize quantitative voice features derived using the COVAREP library. These works provide the empirical basis for identifying indicators of vocal strain or dysfunction and have been instrumental in shaping the diagnostic logic embedded in the application.

A key component of the research involved validating the results produced by VocalCheck through comparative analysis with previously published experimental studies, notably that of (Szklanny, 2019). The outcomes generated by VocalCheck closely mirrored those reported in that study, suggesting that the application's analytical mechanisms provide a comparable level of diagnostic resolution. This fact reinforces VocalCheck's credibility as a practical alternative or supplement to more established laboratory-grade voice assessment tools.

The application underwent empirical testing among members of the PJAIT choir. This group was chosen because of their regular vocal activity, which makes them particularly sensitive subjects for voice monitoring. Notably, the application identified signs of vocal strain among male singers – an observation that led to improved vocal hygiene practices within the ensemble. In contrast, no significant vocal anomalies were detected in the female participants, indicating a generally healthy vocal condition within this subgroup. The tests were conducted using the choir members' personal smartphones, encompassing a variety of device models with varying microphone specifications and audio processing capabilities. Despite these hardware variations, VocalCheck demonstrated a high degree of usability. Its consistent performance across different devices suggests that the application is not only technically sound but also adaptable to real-world conditions, making it a promising tool for non-invasive widely accessible voice screening in diverse user populations.

5.2. Software development using large language models

As part of the project, a fully functional application for assessing voice condition from recordings was created. The latest articles suggest that, under appropriate conditions, mobile phones can be used effectively to obtain recordings from which sufficiently precise calculations of various voice parameters are possible. The application's interface facilitates the interpretation of parameter values, and the recording of test history allows tracking changes in the user's voice over time. VocalCheck takes into account widely used standards for graphic interface design, which positively influences users' feelings when using the application compared to competing solutions.

Ultimately, the scripts generated by the language models do not correctly calculate the PS, CPP, and NAQ parameters. Considering COVAREP in the queries did not have a significant impact on the quality of the results. The reasons for this may be as follows: a high level of complexity in the algorithms calculating the parameters – GPT models are trained comprehensively. Even with the use of fine-tuning methods, they are still susceptible to 'hallucinations,' i.e., situations where the system fabricates information, thereby risking misleading the user. Insufficient query quality – despite using good practices, such as writing specific commands, expanding abbreviations, and guiding the model through a 'chain of thought,' we are uncertain that the generated responses will be accurate. Errors may result, among other things, from the model's inability to verify the transmitted information, overfitting, or bias in the training data.

The second example of using large language model (LLM) in code production shows the potential and development of generative models in this area. Reproducing methods from the COVAREP repository may have been too niche a task for these systems; however, solving more popular problems, such as creating simple graphical interfaces, is susceptible to automation. To achieve greater precision in voice condition evaluation, VocalCheck can be expanded with additional parameters, such as those available in plugins for the PRAAT tool. Another direction for the development of research is the use of LLM for processing functions in the MATLAB language. Generative machine learning models demonstrate promising results in performing simple programming tasks. In addition to ChatGPT and Microsoft Copilot, a potential candidate for further testing is the DeepSeek model from China. Version R1 of this model is publicly available on the GitHub server. Other competitive LLM models include, for example, the Grok AI model, integrated with the X platform (formerly Twitter), and the Claude AI model, created by former OpenAI employees. With more precise construction of queries, there is a chance of successfully generating functions that calculate voice parameters in MATLAB, as is the case with COVAREP.

An attempt can then be made to rewrite them in other programming languages, such as Python or Java. The last COVAREP release dates back to 2016, and since then, this project has not been updated. The high complexity of the functions implemented there makes it difficult to support this tool, and machine learning techniques offer a chance to reproduce them in a more accessible programming environment.

The naming convention for scripts generated by the LLMs is as follows: name of the LLM and the name of the calculated parameter, separated by an underscore. Versions of the scripts generated after the prompt that points to the COVAREP implementations are marked by the text covarep_mentioned at the end of the filename.

The tests conducted using artificial intelligence tools demonstrate that software development can be effectively supported by such technologies, highlighting their potential to automate and accelerate the implementation of computational solutions, particularly for general programming tasks. However, for highly specialized systems with a single established implementation, such as COVAREP, the results remain suboptimal, as AI-generated code often falls short of the precision and performance required for such domain-specific frameworks. It can nevertheless be assumed that, when sufficiently detailed prompts are provided, the generated code could function correctly, as evidenced by operational examples presented in this study. In the case of extensive and intricate projects like COVAREP, the challenge of regenerating computational code that reproduces identical results constitutes a distinct research problem within computer science, encompassing issues such as reproducibility, algorithmic fidelity, and the validation of AI-assisted implementations.

6. Conclusion

The application VocalCheck, presented in this study, has demonstrated its utility as a tool for preliminary voice diagnostics. Designed with accessibility and practical deployment in mind, it offers an effective means of assessing vocal health using widely available mobile devices.

Usability tests conducted as part of the study confirmed the application's ease of use, as evidenced by positive UX (user experience) evaluations. These findings indicate that VocalCheck can be easily used by users without specialized training, making it a practical option for both individual and group voice monitoring contexts.

In addition, exploratory experiments were conducted to evaluate the potential of LLMs for generating diagnostic software code. These tests provided valuable insights into how contemporary AI tools can be used to support the development of applications aimed at human voice assessment.

Unlike many existing mobile applications for vocal analysis, VocalCheck incorporates state-of-the-art methodologies from modern voice diagnostics. Its analytical framework is aligned with current scientific standards, ensuring precision and reliability that distinguish it from simpler or more narrowly focused tools.

Data availability statement

The following supporting information can be downloaded at: https://github.com/s19110/VocalCheck. The materials contain code generated by chatGPT and MS Copilot.

The link to the VocalCheck application: https://vocalcheck.pjwstk.edu.pl.

The application supporting the findings of this study is available from the corresponding author upon reasonable request.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contribution

Hubert Daniszewski – software development, validation, formal analysis, investigation, resources, writing – original draft preparation, visualization. Krzysztof Szklanny – conceptualization, methodology, resources, writing – review and editing, supervision. Piotr Wrzeciono – conceptualization, methodology, resources, writing – review and editing, supervision. All authors reviewed and approved the final manuscript.

References

1. Alhussein M., Muhammad G. (2019), Automatic voice pathology monitoring using parallel deep models for smart healthcare, *IEEE Access*, **7**: 46474–46479, https://doi.org/10.1109/ACCESS.2019.2905597.
2. 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.
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. Airas M., Alku P. (2007), Comparison of multiple voice source parameters in different phonation types, [in:] *Proceedings Interspeech 2007*, https://doi.org/10.21437/interspeech.2007-28.
5. Angadi V., Chih M.-Y., Stemple J. (2023), Developing and testing a smartphone application to enhance adherence to voice therapy: a pilot study, *International Journal of Environmental Research and Public Health*, **20**(3): 2436, https://doi.org/10.3390/ijerph20032436.
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. Asci F. *et al.* (2020), Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender, *Sensors*, **20**(18): 5022, https://doi.org/10.3390/s20185022.
8. Björkner E., Sundberg J., Alku P. (2005), Subglottal pressure and NAQ variation in voice production of classically trained baritone singers, [in:] *Proceedings Interspeech 2005*, pp. 1057–1060, https://doi.org/10.21437/interspeech.2005-424.
9. Boersma P. (2001), Praat, a system for doing phonetics by computer, *Glot International*, **5**(9/10): 341–345.
10. 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.
11. Compton E.C. *et al.* (2022), Developing an Artificial Intelligence tool to predict vocal cord pathology in primary care settings, *The Laryngoscope*, **133**(8): 1952–1960, https://doi.org/10.1002/lary.30432.
12. Cooper W.E., Sorensen J.M. (1981), *Fundamental Frequency in Sentence Production*, Springer Science & Business Media.
13. 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.
14. 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.
15. 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.
16. 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.
17. Hendriksz C.J. *et al.* (2014), Burden of disease in patients with Morquio A syndrome: results from an international patient-reported outcomes survey, *Orphanet Journal of Rare Diseases*, **9**(1): 32, https://doi.org/10.1186/1750-1172-9-32.

18. 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.
19. Ho T.K. (1995), Random decision forests, [in:] *Proceedings of 3rd International Conference on Document Analysis and Recognition*, **1**: 278–282, https://doi.org/10.1109/ICDAR.1995.598994.
20. Holland J.H. (1975), *Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, The MIT Press.
21. Howard D.M. (1995), Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers, *Journal of Voice*, **9**(2): 163–172, https://doi.org/10.1016/s0892-1997(05)80250-4.
22. Kane J., Gobl C. (2011), Identifying regions of nonmodal phonation using features of the wavelet transform, [in:] *Proceedings Interspeech 2011*, pp. 177–180, https://doi.org/10.21437/interspeech.2011-76.
23. 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.
24. Kasuya H., Shigeki O., Mashima K., Ebihara S. (1986) Normalized noise energy as an acoustic measure to evaluate pathologic voice, *The Journal of the Acoustical Society of America*, **80**(5): 1329–1334, https://doi.org/10.1121/1.394384.
25. Kaur P., Stoltzfus, J.C. (2017), Bland–Altman plot: a brief overview, *International Journal of Academic Medicine*, **3**(1): 110–111, https://doi.org/10.4103/IJAM.IJAM_54_17.
26. Klingholz R. (1987), The measurement of the signal-to-noise ratio (SNR) in continuous speech, *Speech Communication*, **6**(1): 15–26, https://doi.org/10.1016/0167-6393(87)90066-5.
27. 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.
28. Lebacq J., Schoentgen J., Cantarella G., Bruss F.T., Manfredi C., DeJonckere P. (2017), Maximal ambient noise levels and type of voice material required for valid use of smartphones in clinical voice research, *Journal of Voice*, **31**(5): 550–556, https://doi.org/10.1016/j.jvoice.2017.02.017.
29. Manfredi C. *et al.* (2017), Smartphones offer new opportunities in clinical voice research, *Journal of Voice*, **31**(1): 111, https://doi.org/10.1016/j.jvoice.2015.12.020.
30. 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.
31. 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).
32. 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.
33. Suvvari T.K. (2023), The role of Artificial Intelligence in diagnosis and management of laryngeal disorders, *Ear, Nose & Throat Journal*, **104**(11), https://doi.org/10.1177/01455613231175053.
34. 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.
35. Szklanny K., Gubrynowicz R., Iwanicka-Pronicka K., Tylki-Szymańska A. (2016), Analysis of voice quality in patients with late-onset Pompe disease, *Orphanet Journal of Rare Diseases*, **11**(1): 99, https://doi.org/10.1186/s13023-016-0480-5.
36. Szklanny K., Gubrynowicz R., Ratyńska J., Chojnacka-Wądołowska D. (2019), Electroglottographic and acoustic analysis of voice in children with vocal nodules, *International Journal of Pediatric Otorhinolaryngology*, **122**: 82–88, https://doi.org/10.1016/j.ijporl.2019.03.030.
37. Szklanny K., Gubrynowicz R., Tylki-Szymańska A. (2018), Voice alterations in patients with Morquio A syndrome, *Journal of Applied Genetics*, **59**(1): 73–80, https://doi.org/10.1007/s13353-017-0421-6.

38. Szklanny K., Tylki-Szymańska A. (2018), Follow-up analysis of voice quality in patients with late-onset Pompe disease, *Orphanet Journal of Rare Diseases*, **13**(1): 189, https://doi.org/10.1186/s13023-018-0932-1.

39. Szklanny K., Wrzeciono P. (2019a), 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.

40. Szklanny K., Wrzeciono P. (2019b), The application of a genetic algorithm in the noninvasive assessment of vocal nodules in children, *IEEE Access*, **7**: 44966–44976, https://doi.org/10.1109/ACCESS.2019.2908313.

41. 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.

42. Uloza V. *et al.* (2015), Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening, *European Archives of Oto-rhino-laryngology*, **272**(11): 3391–3399, https://doi.org/10.1007/s00405-015-3708-4.

43. Uloza V. *et al.* (2023), Accuracy of acoustic voice quality index captured with a smartphone – measurements with added ambient noise, *Journal of Voice*, **37**(3): 465, https://doi.org/10.1016/j.jvoice.2021.01.025.

44. Verde L., De Petro G., Alrashoud M., Ghoneim A., Al-Mutib K.N., Sannino G. (2019), Leveraging Artificial Intelligence to improve voice disorder identification through the use of a reliable mobile app, *IEEE Access*, **7**: 124048–124054, https://doi.org/10.1109/ACCESS.2019.2938265.

45. 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.

46. Yumoto E., Gould W.J., Baer T. (1982), Harmonics-to-noise ratio as an index of the degree of hoarseness, *The Journal of the Acoustical Society of America*, **71**(6): 1544, https://doi.org/10.1121/1.387808.

47. Zheng F., Zhang G., Song Z. (2001), Comparison of different implementations of MFCC, *Journal of Computer science and Technology*, **16**(6): 582–589, https://doi.org/10.1007/BF02943243.