Biomedical Signal Processing and Control 78 (2022) 103878

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Videokymogram Analyzer Tool: Human–computer comparison

Aleš Zita [a,*], Adam Novozámský [a], Barbara Zitová [a], Michal Šorel [a], Christian T. Herbst [b,c], Jitka Vydrová [d], Jan G. Švec [b,d]

[a] The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod vodárenskou věží 4, Prague 8, 182 08, Czech Republic
[b] Palacký University, Faculty of Sciences, Department of Experimental Physics, Voice Research Lab, 17. listopadu 12, Olomouc, 771 46, Czech Republic
[c] Mozarteum University, Mirabellplatz 1, Salzburg, 5020, Austria
[d] Voice Centre Prague, Medical Healthcom, Ltd., Národní 11, Prague, 110 00, Czech Republic

ARTICLE INFO

Keywords:
Image analysis
Videokymography
Vocal fold vibrations

ABSTRACT

Videokymography (VKG) is a modern video recording technique used in laryngology and phoniatrics to examine vocal fold vibrations. To obtain quantitative information on the vocal fold vibration, VKG image analysis is needed but no software has yet been validated for this purpose. Here, we introduce a validated software tool that aids clinicians to evaluate diagnostically important vibration characteristics in VKG and other types of kymographic recordings. State-of-the-art methods for automated image evaluation were implemented and tested on a set of videokymograms with a wide range of vibratory characteristics, including healthy and pathologic voices. The automated image segmentation results were compared to manual segmentation results of six evaluators revealing average differences smaller than one pixel. Furthermore, the automatically categorized vibratory parameters precisely agreed with the average visual assessment in 84 and 91 percent of the cases for pathological and healthy patients, respectively. Based on these results, the newly developed software was found to be a valid, reliable automated tool for the quantification of vocal fold vibrations from VKG images, offering a number of novel features relevant for clinical practice.

1. Introduction

The vibration characteristics of the laryngeal tissues – particularly those of the vocal folds – are critical for the evaluation of voice disorders by laryngologists and phoniatricians [1,2]. The vocal folds are a pair of elastic tissues in larynx (see Fig. 1) and their vibrations produce phonation. The vibrating folds gradually close and open the space between them (*rima glottidis*, or glottis) with the frequency range of about 60–1000 Hz [3]. To visualize the vocal folds, laryngeal endoscopy is routinely used in clinical practice (Fig. 1, left). There are three standard techniques to display and evaluate the vibration of the vocal folds using laryngeal endoscopy: *videostroboscopy*, *high-speed laryngeal videoendoscopy*, and *videokymography* [4,5].

Videostroboscopy displays an illusory slowed-down motion of the vocal folds in real time by temporarily synchronizing the images, captured by a standard endoscopic video camera, with vocal fold vibratory cycles [6,7]. While this method is most frequently used in clinical practice, it is not suitable for documenting and quantifying irregular vibrations typical for disordered voices.

High-speed videoendoscopy (HSV) captures laryngeal images with a high-speed camera at frame rates well above the fundamental frequency of phonation, typically exceeding 1000 frames per second [4,8,9]. This method accurately documents each oscillatory cycle of the vocal fold oscillations and produces large amounts of data which are beneficial for research purposes and can be used for various types of analyses, such as glottal segmentation and extraction of glottal area waveforms, digital kymography, phonovibrography, laryngotopography, etc. [10–14]. However, the HSV method has not yet been widely implemented in clinical practice, mainly because it does not provide real time visual feedback and is time-wise demanding, due to the large volume of acquired data [15]. Addressing this, a viable strategy to diminish the vast amount of data generated by HSV is to reduce the two spatial image dimensions to a single one. This is achieved via the videokymographic imaging, which is the main subject of the present study.

In *videokymography* (VKG) special cameras are utilized to capture images of the vibrating vocal folds at a single line perpendicular to glottal axis with the rate of 7200 line images per second and allow simultaneous observation in standard and videokymographic modes (see Fig. 1). This allows the clinician to flexibly orient the camera in order to record the desired line of interest [16–18]. VKG method offers apparent benefits by combining real time imaging feedback

[*] Corresponding author.
E-mail address: zita@utia.cas.cz (A. Zita).

https://doi.org/10.1016/j.bspc.2022.103878
Received 30 December 2021; Received in revised form 8 May 2022; Accepted 4 June 2022
Available online 8 July 2022

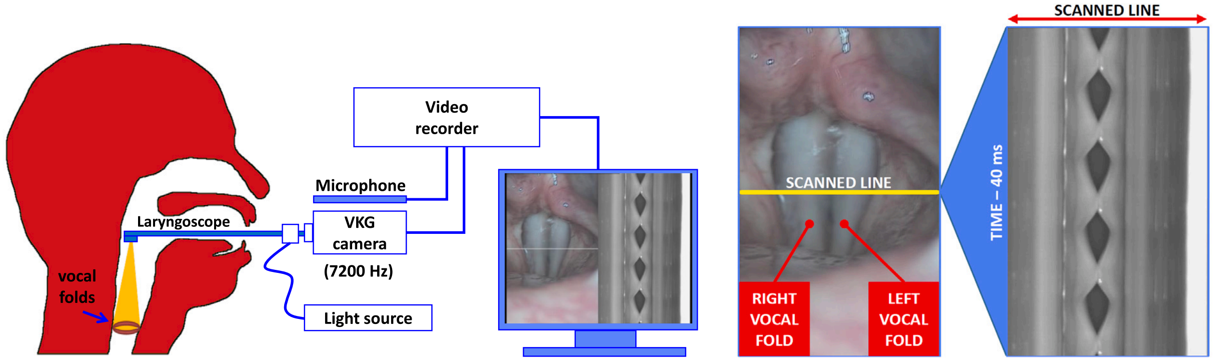


Fig. 1. Videokymography: On the left there is the examination of vocal fold vibrations by laryngeal endoscopy using a videokymographic (VKG) camera. On the right there are two parallel imaging modes of the videokymographic (VKG) camera: standard (left) and videokymographic (right). The videokymographic image is composed of successively acquired scanned lines at the location indicated in the standard mode. The time is mapped onto the vertical axis within the VKG image, going from top to bottom. The standard duration is 40 ms (resulting from the standard rate of 25 frames/s [16,17]). This also applies to the other figures with VKG images in this article.

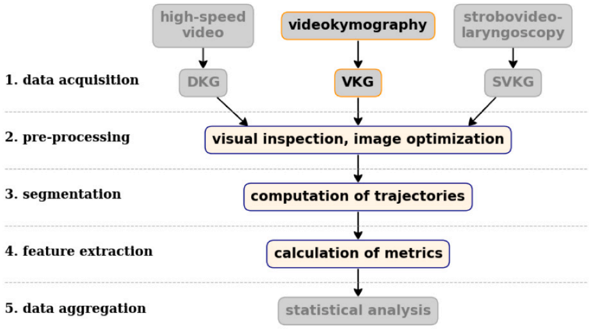


Fig. 2. Proposed data processing pipeline for kymographic documentation and analysis of vocal fold vibration. This study particularly focuses on layers 2 through 4 (i.e., pre-processing, segmentation, and feature extraction). The algorithms and software presented in this manuscript have been validated with videokymographic footage.

found in videostroboscopy with the advantages of high-speed imaging, i.e., sufficient frame rates to truthfully document each oscillatory cycle of the vocal folds [19].

In a VKG system, data acquisition and kymographic image generation is facilitated simultaneously on the hardware layer. In contrast, two further strategies offer the possibility to generate surrogate kymographic images from previously recorded endoscopic laryngeal footage [5]: (a) digital kymography (DKG) [14,20], operating on HSV data; and (b) strobovideokymography (SVKG) [21–23], operating on videostroboscopic data.

Here we propose a system for data analysis in which either of these three types of kymographic image (VKG, DKG, or SVGK) is processed in a pipeline that is schematically illustrated in Fig. 2. In particular, the available kymographic images are visually inspected, optimized and selected for further treatment (see **layer 2: pre-processing** in Fig. 2). In analogy to HSV data, the vibrating glottal edge is segmented, thus computing the time-varying medio-lateral deflections of the vocal folds (**layer 3: segmentation**). However, while in HSV the segmentation operates on two spatial dimensions, VGK images have a reduced dimensionality, thus requiring a fundamentally different segmentation approach. The resulting data is then subjected to extraction of dedicated metrics that allow for quantitative assessment (**layer 4: feature extraction**) and can finally be used for statistical analysis or intra-subjective comparison if so required (layer 5: data aggregation).

While exclusively manual treatment within this analysis pipeline has been partially pursued for scientific exploration [24–26], this is rather time-consuming and thus not feasible in clinical practice. Ideally, layers 2 through 4 should be automated and completed by computer-aided (semi)automatic software algorithms and a supporting graphical user interface (GUI).

Addressing this, some systems have been developed previously. For instance, the commercially marketed Kay Elemetrics Image Processing System (KIPS) offers features to produce DKGs from HSV, segmentation of these DKGs, as well as limited analysis of the segmented DKG contours in the form of metrics addressing glottal opening (glottal width), mean fundamental frequency (f_o), the vibratory amplitudes of left and right vocal folds, as well as the percentage of time when the glottis is closed with respect to the duration of the analysis. Another system, proposed by Manfredi et al. [27] operates directly on VKG images and offers the image segmentation and extraction of basic quantitative parameters, such as the left-to-right amplitude and period ratios, open-to-closed phase ratios, and phase symmetry index [28,29].

Despite the existence of these previously established systems, which constitute commendable groundbreaking work in their own right, a comprehensive coverage of layers 2–4 in the proposed analysis pipeline (Fig. 2) is still missing. Neither of the developed software tools has been available for analyzing the sets of existing clinical VKG recordings: the tool of Manfredi et al. [27] has not been released for external use and its exploration has been limited to preliminary or case studies [29,30], whereas the KIPS software allows analyzing only DKG and not VKG recordings. Furthermore, performance of neither of these software tools has been validated against visual assessment.

The goal of this project was to fundamentally address these issues, targeting two particular objectives:

Objective I was to develop and test a user-friendly software tool for automated analysis of clinical videokymographic recordings that can be used in a clinical setting. This software predominantly targets layers 2–4 in the proposed analysis pipeline (see Fig. 2). In this context, we present (a) a GUI for acquiring kymographic videos or images and their pre-processing; (b) a segmentation algorithm that supports user-defined image adaptations; and (c) implementation of a number of clinically relevant metrics.

Objective II was constituted by a rigorous validation of the implemented segmentation and feature extraction algorithms. This is achieved by comparing the proposed toolset with manual visual assessments, in order to verify the accuracy and applicability of the proposed solution. For this, we took advantage of a previously developed protocol for visual analysis of videokymograms using pictograms [24,31], transferring the visual pictogram features into quantitative vibratory parameters. This was achieved for clinically relevant features, such as: the relative duration of glottal closure, left–right differences in vibratory amplitudes, frequencies and phases, left–right axis shifts, opening versus closing durations, and cycle-to-cycle variability [1,32].

2. Materials and data

2.1. Datasets

All the data used in our study were acquired in cooperation with the Voice and Hearing Centre Prague, a medical institution specialized

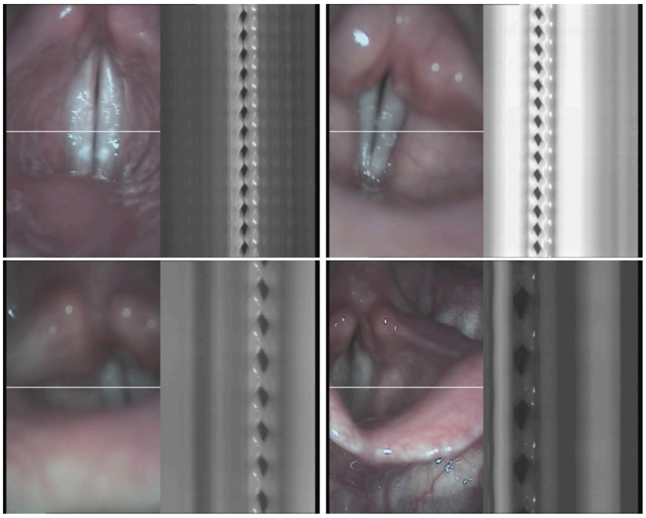


Fig. 3. Examples of video frames extracted from routine clinical VKG recordings demonstrate the variability of laryngeal settings, vibratory patterns, and image quality. The vibrations may show lack of glottal closure or can be asymmetrical. The glottal opening can contain specular reflections, be over-saturated, blurred, or even off-center. Such variability had to be taken into account for developing the VKG Analyzer software.

in voice diagnostics. Routinely acquired videokymographic recordings of patients with and without voice disorders were used; no special recordings were made for this study. No exclusion criteria were applied here on the subjects; our primary goal was to obtain and analyze recordings showing the largest possible variability of findings and image quality, regardless of the clinical diagnosis, gender or age. Healthy as well as disordered patients were included to ensure the robustness of the proposed methods. The most common diagnoses included were: laryngitis chronica, hyperfunctional dysphonia, oedema laryngis, hemorrhage, vocal fold atrophy, paresis mm. interni and voice fatigue. However, since clinical diagnoses have traditionally been based mostly on structural rather than vibrational features, our goal was not to obtain clinical diagnoses but rather to accurately capture the vibratory features of the vocal folds which provide crucial additional information on the functionality of the vocal folds [32]. For the evaluations, we have therefore selected images containing large variation of vibratory patterns and having various levels of image quality in order to test the robustness of image segmentation and vibration analysis. Examples of the clinical VKG images subjected to our analysis are shown in Fig. 3.

Three datasets were used to create and validate the software. The first dataset consisted of 500 randomly chosen images from clinical VKG examinations of healthy as well as voice-disordered subjects. It was used for a heuristic adaptation of the algorithms and fine-tuning of the parameters. We named this dataset the *“Training Dataset”*. This dataset was used in the **layer 2: pre-processing and layer 3: segmentation** of the processing pipeline as depicted in Fig. 2.

A second dataset, the *“Segmentation Validation Dataset”*, was created to test the performance of the segmentation algorithm (**layer 3: segmentation** in the processing pipeline depicted in Fig. 2). The dataset consisted of manual annotations of 834 key points, i.e., the opening, closing, lateral and medial extrema for left and right vocal fold movement contour, performed by 6 raters, yielding the total of 5004 annotations. Details on this dataset are provided in Section 3.6 devoted to the validation studies.

A third dataset, the *“Attributes Validation Dataset”*, was used to evaluate of the accuracy of the extracted vibration attributes (**layer 4: feature extraction** in diagram displayed in Fig. 2), testing the overall analyzer performance. This dataset contained the total of 13500 visually-based manual evaluations of 9 vibratory features obtained from ten evaluators. These evaluations were performed on 50 VKG images from 50 patients with various voice disorders showing the largest possible range of pathological vibratory patterns and 200 VKG images from 40 healthy patients. Further details on this dataset are also provided in Section 3.6 devoted to the validation studies.

2.2. Videokymographic data acquisition/voice recording

In order to acquire the videokymographic images, we used a commercially available 2nd generation videokymography camera (Cymo, Netherlands) connected to a 90° rigid laryngoscope (type 130310529, Xion, Germany) with a bright light source (300-W xenon, type FX 300 A, Fentex, Germany) (Fig. 1). Examples of the videokymographic images from different patients can be seen in Fig. 3. Audio signal has also been captured together with the videokymographic data using an electret microphone (Xion) for perceptual monitoring of the recorded voices.

2.3. Software tool implementation

Initial development has been realized in *Image Processing Toolbox for Matlab* [33]. The final application is programmed in C++, complemented by the *openCV* library [34] for image and video handling and by the *Qt* library [35] for the graphical user interface. The *SQLite* database system [36] was used for data storage.

3. Method

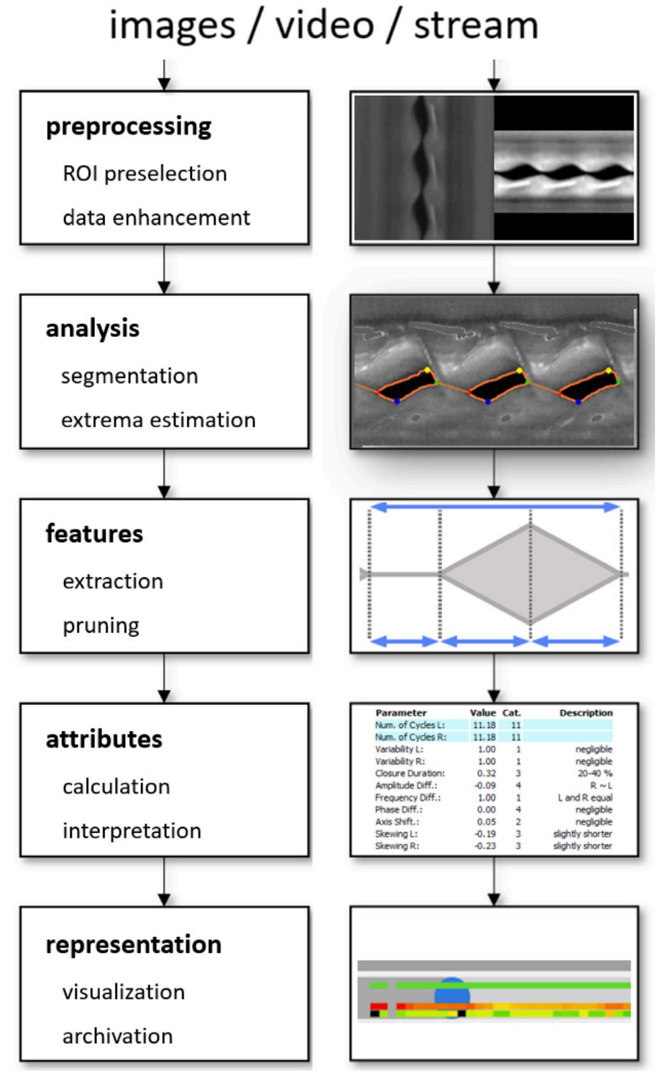
The proposed software solution, addressing the **objective I**, consists of five main building blocks, as shown in Fig. 4. The input data can be in the form of single kymographic images (from DKG, VKG or SVKG modality), a VKG video file, or a live video stream of the VKG examination session. Following the initial information rich frames detection and preprocessing, the software localizes the fundamental vibration structures for every vibration cycle — the contours of glottal openings, the lateral movement extrema, and the opening/closing points. These basic features are used for the derivation of advanced features, and ultimately for computation of the final vibration attributes. Finally, the software visualizes the results in the graphical user interface. In the following paragraphs, the individual pipeline blocs from Fig. 4 are described.

3.1. Information-rich-frames detection

A typical VKG video data acquisition process produces many frames containing irrelevant data (the cases where the patient moved, did not phonate, or the resulting images are off-center or low quality). As a part of the preprocessing, the image content richness of every frame is estimated. The content richness detection is based on searching for the vocal fold oscillations’ amplitude using the column frequency analysis. The absolute values of the first 32 coefficients of each column’s Fourier transformation determine the vibration amplitudes for the relevant frequencies (see Fig. 5). The maximum of calculated amplitudes therefore signifies the level of vibrations in the VKG image. Frames achieving a higher maximum value than the empirically defined threshold are marked. The process of preselecting the content-rich frames helps the physician to focus on the relevant parts of the VKG video, where the vocal folds are visible and vibrating. An interactive visualization tool (the representation part in Fig. 4) helps the user to select the information-rich-frames of interest.

3.2. Preprocessing of VKG images

The acquired data contain various degradations (examples are depicted in Fig. 3). Primary goal of Image preprocessing (**layer 2** Fig. 2)



Parameter	Value	Cat.	Description
Num. of Cycles L:	11.18	11	
Num. of Cycles R:	11.18	11	
Variability L:	1.00	1	negligible
Variability R:	1.00	1	negligible
Closure Duration:	0.32	3	20-40 %
Amplitude Diff.:	-0.09	4	R ~ L
Frequency Diff.:	1.00	1	L and R equal
Phase Diff.:	0.00	4	negligible
Axis Shift.:	0.05	2	negligible
Skewing L:	-0.19	3	slightly shorter
Skewing R:	-0.23	3	slightly shorter

Fig. 4. Streamline processing pipeline schema. The input sequence is processed frame by frame. The first stage focuses on image preprocessing (layer 2 in Fig. 2). Next, the glottal openings are segmented (layer 3). The segmentation determines the lateral extrema and opening/closing points. Then the derived vibration features and final attributes are calculated (layer 4). Lastly, the software visualizes the results in the graphical user interface.

is to reduce unwanted artifacts caused by the data acquisition process and normalize the input for further processing.

After loading an image containing the VKG data, the system performs adaptive histogram equalization [37] to normalize the image. In this phase, the operator can further adjust the image contrast and brightness using the controls in the program interface.

When the user initiates the automatic extraction of attributes, the algorithm first removes any specular light reflections caused by the mucosal secretions. Here, the pixels having values higher than the preset threshold (set to 200) are replaced by the mean value of all pixels in the same column. This rudimentary impainting approach is sufficient for the segmentation process. Next, the algorithm cuts out the image borders, which have no informational value. By default, the cutout is set to 1/4th of the image width from both sides. For the rest of the pipeline, only the middle part of the image containing the relevant vibration structures is kept.

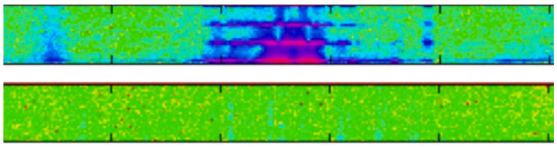


Fig. 5. Spectral analysis for the selection of content-rich images. The example shows the spectral analysis of each column of the VKG image with (top) and without (bottom) pronounced vibrations. The x-axis of the graph shows the VKG image spatial domain; the y-axis shows the first 32 Fourier coefficients' absolute values.

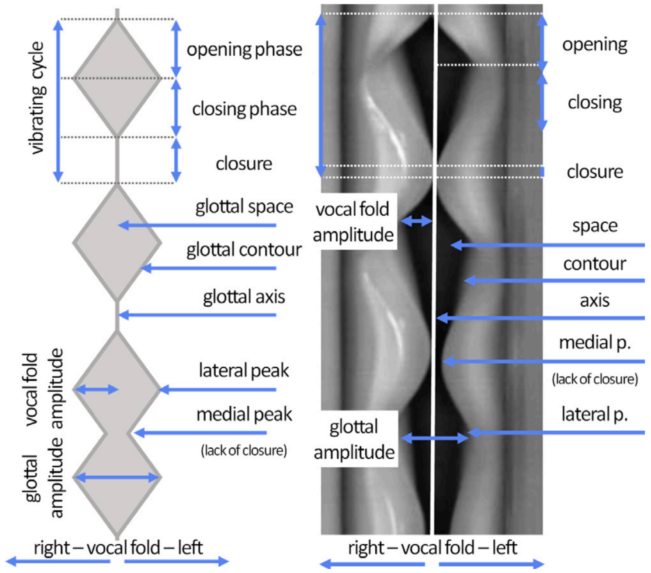


Fig. 6. Vibration features in videokymograms. (a) Schema and (b) real case.

3.3. Segmentation and extrema estimation

Vibration characteristics of the vocal folds in the acquired VKG images are computed from the glottal openings (see Fig. 6). The segmentation of the openings outlines the border between the laryngeal tissue and the open glottis. This part of the processing pipeline covers the **layer 3: segmentation** in the schematic overview (recall Fig. 2).

First, the algorithm determines the active part of the VKG image using the intensity variations in every column. It finds the left-most block of 5 columns of image pixels, all having the standard deviation higher than 0.5. Then it finds the right-most columns of pixels with the same property. The found columns define the part of the VKG image with pronounced vibrations containing the glottal openings for segmentation. In the next phase of the processing, the algorithm executes a segmentation of the glottal area using pixel intensity thresholding. In order to find the global threshold for the image segmentation, the algorithm first estimates the middle line of the glottal opening. Next, the global threshold is estimated using the sorted middle line in the next phase. (see the Algorithm 1). The procedure performs the final segmentation by selecting pixels having an intensity lower than the calculated threshold. All mentioned parameters in both steps of the algorithm were established and fine-tuned empirically on the randomly selected data of 500 VKG frames from healthy and unhealthy patients (the *Training Dataset* defined in Section 2.1).

The global segmentation method can produce unwanted artefacts such as false 'holes' in dark areas of vocal folds. To remove these incorrectly segmented areas, the algorithm performs a morphological opening [38] using a rectangular morphological element of size 3 × 3.

The achieved segmentation determines the contour pixels of the glottal openings (refer to Fig. 7(b)). The extremal points of the contours in the temporal domain (up or down on image) denote the *glottal*

Algorithm 1: Find the Global Threshold

```
$sorted\_image$ := SortColumns($image$)
$new\_height$ := Height($image$) * 0.55
$sorted\_image$ := $sorted\_image$[1:$new\_height$, :]
$column\_sums$ := Sum($sorted\_image$, 1)
$middle\_idx$ := ArgMin($column\_sums$)
$sorted\_column$ = Sort($image$[:, $midddle\_idx$])
$sorted\_column$ := Filter1D($sorted\_column$, GAUSS)
for $i$ in 1:Length($sorted\_column$) do
    if $sorted\_column[i] \leq 0.1$ then
        $min\_idx := i$
    end
    if $sorted\_column[i] \leq 0.22$ then
        $max\_idx := i$
    end
end
$sorted\_column$ := $sorted\_column$[$min\_idx$:$max\_idx$]
$gradient\_vector$ := $sorted\_column$[1:end-1] - $sorted\_column$[2:end]
for $i$ in 1:Length($sorted\_column$) do
    if $gradient\_vector[i] > 0.03$ then
        $global\_threshold$ := $sorted\_column[i]$
        return $global\_threshold$
    end
end
```

opening (top) and *closing* (bottom) *points*. Between the glottal cycles (defined by the segmented glottal opening(s)), the glottis can be closed (Fig. 7(b)). In such a case, the line connecting the closing of one cycle to the opening of the following cycle is used to approximate the position of the boundary between the left and the right vocal fold during glottal closure. The final tracing contours of the movements of the vocal folds are formed as curves running from the first glottal opening, each on one side, including the connecting lines when the vocal folds are closed, repeatedly for every glottal cycle, until the end of the last glottal opening on the analyzed image frame. These tracing lines are then used for finding the lateral and medial peaks (see Figs. 6 and 7(b)).

Each of the established tracing lines (left and right) can be viewed as a continuous curve. Therefore, we can use the first derivative test to find the lateral peaks (the violet points in Fig. 7(b)) as well as the medial peaks when glottal closure is missing (Fig. 6). Places where the first-order derivative is zero signify the places of either extreme or saddle point. A second-order derivative is used to distinguish between an extreme and the saddle point. The lateral peaks are used for finding the vibration amplitudes. The medial peaks are important to localize for the cases, where the vocal folds do not close completely.

3.4. Features

The main calculation pipeline (**layer 4**, Fig. 2) of the proposed software starts with extracting the basic and advanced (derived) features. The basic vibration features targeted here are: the frequency and regularity of vocal fold vibration, the relative duration of glottal closure, opening versus closing duration, and the left–right vibratory asymmetry. These features are calculated directly from the detected extrema points — namely the frequency, lateral amplitude (vocal fold vibration amplitude), and the maximum opening point (lateral peak). From the combination of the left and right extrema points, we can also determine the phases where the glottis is closed and open. The lateral peaks and the closed phase (or the medial peak when there is no closed phase) are used to detect the opening and closing points. (See Table 1 for reference; the upper index R or L denotes correspondence to the right or left vocal fold; the lower index $i \in \{1, \dots, n\}$ denotes the number of the corresponding vibration cycle, where n is the number of cycles in the videokymogram.).

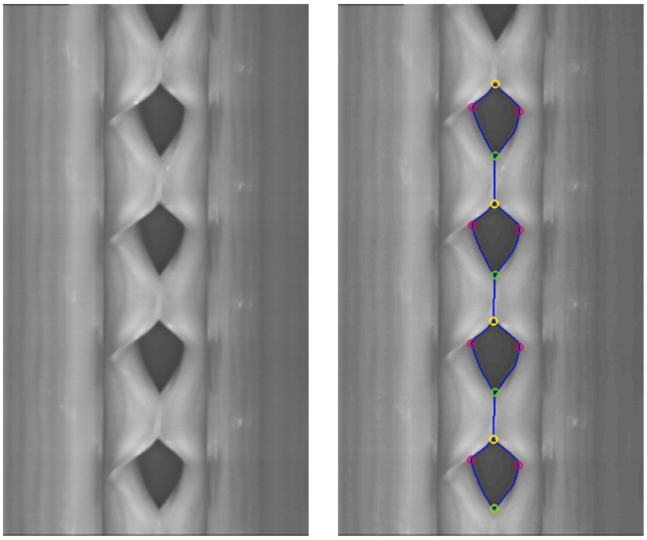

(a) (b)

Fig. 7. Glottal space contouring and subsequent main features detection. (a) Original VKG image; (b) Detected glottal opening border with the main feature points: Opening Points, Closing Points, and Lateral Peaks.

Table 1
Basic glottal features in videokymograms (see Fig. 6); upper indices R and L denote the right and left vocal folds and lower index i denotes the number of the vibration cycle in the videokymogram.

Basic feature	Notation
Opening points	O_i
Closing points	C_i
Lateral peaks	A_i^R, A_i^L
Medial peaks	M_i^R, M_i^L

Table 2
Derived glottal features in videokymograms (see Fig. 6); upper index $j \in \{R, L\}$ denotes the right and left vocal folds and lower index i denotes the number of the vibration cycle in the videokymogram.

Advanced features	Notation and definition
Generalized opening points	$\tilde{O}_i^j = \{O_i, M_i^j\}$
Generalized closing points	$\tilde{C}_i^j = \{C_i, M_i^j\}$
Opening phase duration	$t_i^{oj} = A_i^j(y) - \tilde{O}_i^j(y)$
Closing phase duration	$t_i^{cj} = \tilde{C}_i^j(y) - A_i^j(y)$
Open phase duration	$T_i^{oj} = t_i^{oj} + t_i^{cj} = \tilde{C}_i^j(y) - \tilde{O}_i^j(y)$
Closed phase duration	$T_i^{cj} = \tilde{O}_{i+1}^j(y) - \tilde{C}_i^j(y)$
Vibration cycle duration	$T_i^j = T_i^{oj} + T_i^{cj} = t_i^{oj} + t_i^{cj} + T_i^{cj} = \tilde{O}_{i+1}^j(y) - \tilde{O}_i^j(y)$
Vocal fold amplitudes	$a_i^j = \text{mean}(\lvert A_i^j(x) - \tilde{O}_i^j(x)\rvert, \lvert A_i^j(x) - \tilde{C}_i^j(x)\rvert)$
Glottal amplitudes	$a_i = A_i^L(x) - A_i^R(x)$

The derived glottal features are computed from the basic features using the definitions in Table 2. The generalized opening points are defined as the union of opening points and medial peaks, and similarly, generalized closing points are defined as the union of closing points and medial peaks. The generalized opening and closing points enclose open phases, while the generalized opening points separate vibration cycles.[1]

3.5. Attributes

The set of vocal fold vibration attributes used by clinicians was previously implemented by the authors into a visually-perceptual VKG

[1] Depending on the definition, vibration cycles can be separated by the generalized opening points, by the lateral peaks, or by the generalized closing points.

Table 3
Cycle-to-cycle amplitude variability: correspondence between the numerical values of the Amplitude Periodicity Index (API) and categories of the parameter in the VKG visual evaluation form [24,31].

Category	Description	VariabilityR, VariabilityL
1	Negligible	(0.85, 1]
2	Small	(0.61, 0.85]
3	Medium	(0.5, 0.61]
4	Large	[0, 0.5]

evaluation sheet [24,31]. In our software, we aimed at evaluating these visual attributes automatically using a set of parameters derived from the detected glottal features. The intervals for the parameters' discretization (see Tables 3–9) were obtained by manually measuring the pictograms depicting the typical idealized VKG waveforms — those served as visual anchors for the previous visual VKG evaluation studies [24,31]. The discretization of the calculated values is mandatory for a backward reference to manual annotations performed using the VKG visual evaluation tool. Additionally, the human-to-computer comparative study utilizes the discretized values for more straightforward performance evaluation.

The meaning of the parameters and their relation to the VKG features was defined as follows:

Number of cycles

(1) $\text{NumberOfCyclesR} = \frac{y_{max}}{\overline{T}^{R}}$

(2) $\text{NumberOfCyclesL} = \frac{y_{max}}{\overline{T}^{L}}$

The "*Number of cycles*" parameter is defined by the duration of the recorded videokymogram y_{max} and the average length of the vibration cycle $\overline{T}^{j} = \frac{1}{n_j} \sum_{i=1}^{n^j} T_i^j$, where $j = R, L$ and n^R and n^L denote the number of full cycles of the right and left vocal fold in the videokymogram determined from the total number of detected opening points O^R and O^L, respectively.

This parameter is directly related to the fundamental frequency of oscillations of the vocal folds and consequently to the produced fundamental frequency of voice.

Cycle-to-cycle variabilities

(3) $\text{VariabilityR} = \underset{i=1,\ldots,n-1}{\text{median}}\ API(i, R)$

(4) $\text{VariabilityL} = \underset{i=1,\ldots,n-1}{\text{median}}\ API(i, L)$

The cycle-to-cycle amplitude variability indicates how much the vocal fold vibration amplitudes deviate from ideal periodic vibrations. This feature is related to the degree of voice roughness [39]. The "*Cycle-to-cycle amplitude variability*" parameter is defined by the Amplitude Periodicity Index (API) [40] $API(i,j) = \frac{\min\{a_i^j, a_{i+1}^j\}}{\max\{a_i^j, a_{i+1}^j\}}$, where $i = 1, \ldots, n-1, \quad j = R, L$. Analogously, the "*Cycle-to-cycle period variability*" can be defined through the Time Periodicity Index (TPI) [40] $TPI(i,j) = \frac{\min\{T_i^j, T_{i+1}^j\}}{\max\{T_i^j, T_{i+1}^j\}}$, where $i = 1, \ldots, n-1, \quad j = R, L$.

Duration of closure

(5) $\text{ClosureDuration} = \underset{i=1,\ldots,n}{\text{median}}\, CQ(i)$

The relative duration of glottal closure is a classic feature that indicates how well the vocal folds close during phonation [32,41]. The relative duration of the closure is defined by the Closed Quotient (CQ) [40] as $CQ(i) = \frac{T_i^c}{T_i}$, $i = 1, \ldots, n$.

Amplitude differences

(6) $\text{AmplitudeDifferences} = \underset{i=1,\ldots,n}{\text{median}}\ ASI(i)$

Table 4
Duration of closure: correspondence between the numerical values of the Closed Quotient (CQ) and categories of the parameter in the VKG visual evaluation form [24,31].

Category	Description	Closure duration
1	No closure	[0, 0.01]
2	1–20	(0.01, 0.2]
3	20–40	(0.2, 0.4]
4	40–60	(0.4, 0.6]
5	>60	(0.6, 1]

Table 5
Amplitude differences: correspondence between the numerical values of the Amplitude Symmetry Index (ASI) and categories of the parameter in the VKG evaluation form [24,31].

Category	Description	Amplitude difference
1	R much larger	[−1,−0.6)
2	R larger	[−0.6,−0.31)
3	R slightly larger	[−0.31,−0.1)
4	R ∼ L	[−0.1, 0.1]
5	L slightly larger	(0.1, 0.31]
6	L larger	(0.31, 0.6]
7	L much larger	(0.6, 1]

Table 6
Frequency differences: correspondence between numerical values and categories of the parameter in the VKG evaluation sheet [24,31].

Category	Description	Frequency difference
1	R faster than L	(0, 0.91)
2	L and R equal	[0.91, 1.1)
3	L faster than R	[1.1,1)

The difference in vibration amplitude of the left and right vocal folds shows the vocal fold asymmetry and can help clinicians discover unilateral pathologies hindering the vibratory ability of the vocal folds [32,41]. The "*Amplitude difference*" parameter is defined by the Amplitude Symmetry Index (ASI) [40] $ASI(i) = \frac{a_i^L - a_i^R}{a_i^L + a_i^R}$, $i = 1, \ldots, n$.

Frequency differences

(7) $\text{FrequencyDifferences} = \frac{\text{NumberOfCyclesL}}{\text{NumberOfCyclesR}}$

This parameter allows discovering differences in the fundamental frequencies of the left and right vocal folds. In normal phonation, the left and right vocal folds are expected to vibrate at the same fundamental frequencies. In the case of left–right frequency differences, the voice may become biphonic or diplophonic [32,41,42]. The "*Frequency difference*" parameter is defined as a ratio between the number of left and right cycles (see parameters (1)–(2)).

Phase differences

(8) $\text{PhaseDifferences} = \underset{i=1,\ldots,n}{\text{median}}\ PSI(i)$

The "*Phase difference*" parameter is defined by the Phase Symmetry Index (PSI) as [40] $PSI(i) = \frac{A_i^L(y) - A_i^R(y)}{T_i}$, $i = 1, \ldots, n$. This parameter provides information on the possible asymmetry between the tension of the left and right vocal folds.

Axis shifts

(9) $\text{AxisShift} = \underset{i=1,\ldots,n}{\text{median}}\ AS(i)$

The "*Axis shift*" parameter is the third parameter revealing the left–right asymmetry of the vocal fold vibration [32]. In contrast to the phase differences, which are mainly visible during the open phase of the glottal vibratory cycle, the axis shift allows discovering the left–right asymmetries during the closed phase of the glottal vibratory cycle [32,41]. The "*Axis shift*" parameter (AS) is defined as [43] $AS(i) = \frac{O_{i+1}(x) - C_i(x)}{a_i}$, $i = 1, \ldots, n$.

Table 7
Phase differences: correspondence between the numerical values of the Phase Symmetry Index (PSI) and categories of the parameter in the VKG visual evaluation form [24,31].

Category	Description	Phase differences
1	R ahead of L: large	(0.3, 1]
2	R ahead of L: medium	(0.15, 0.3]
3	R ahead of L: small	(0.05, 0.15]
4	Negligible	[−0.05, 0.05]
5	L ahead of R: small	[−0.15,−0.05)
6	L ahead of R: medium	[−0.3,−0.15)
7	L ahead of R: large	[−1,−0.3)
14	lambada: large	Yet to be quantified

Table 8
Axis shift: correspondence between numerical values and categories of the parameter in the VKG evaluation form; the evaluation sheet denotes the "R→L" category by 2, the "negligible" category by 1, and the "complex" category by 4.

Category	Description	Axis shift
1	R → L	(0.1,1)
2	Negligible	[−0.1, 0.1]
3	L → R	(−1,−0.1)
6	Complex	Yet to be quantified

Table 9
Opening versus closing duration: correspondence between the numerical values of the Speed Index (SI) and categories of the parameter in the VKG evaluation form.

Category	Description	SkewingR, SkewingL
1	Much shorter	[−1,−0.75)
2	Shorter	[−0.75,−0.35)
3	Slightly shorter	[−0.35,−0.05)
4	Equal	[−0.05, 0.05]
5	Slightly longer	(0.05, 0.35]
6	Longer	(0.35, 0.75]
7	Much longer	(0.75, 1]

Opening versus closing durations, cycle skewing

$$(10)\quad \text{SkewingR} = \underset{i=1,\dots,n}{\text{median}}\, SI(i,R)$$

$$(11)\quad \text{SkewingL} = \underset{i=1,\dots,n}{\text{median}}\, SI(i,L)$$

The opening and closing phases of the vibration cycle of the vocal folds can have different duration. These differences appear as a skewing of the vocal fold vibratory pattern and provides clinically interesting information [32,41,44]. The skewing can differ for the left and right vocal fold and reveals the vocal fold vibration's detailed dynamics.

The "*Opening versus closing duration*" / "*Skewing*" parameter can be quantified by the Speed Index (SI) [45] $SI(i,j) = \frac{t_i^{oj} - t_i^{cj}}{T_i^o} = \frac{t_i^{oj} - t_i^{cj}}{t_i^o + t_i^c} = \frac{SQ(i,j)-1}{SQ(i,j)+1}$, $i = 1,\dots,n, j = R, L$, which is derived from the Speed Quotient (SQ) [46,47] $SQ(i,j) = \frac{t_i^{oj}}{t_i^{cj}}$, $i = 1,\dots,n, j = R, L$.

3.6. Verification studies

To address the **objective II**, two studies were done to compare the performance of the proposed image analysis tool with the clinician visual assessments and verify the usability of the proposed algorithms. The first study evaluated the accuracy of the segmentation process, which is the critical tool for further feature extraction. The second study focused on the estimated vibration attributes and their comparison to the clinician visual assessments.

The first verification study addressed the segmentation accuracy of our algorithm (**layer 3** Fig. 2) using the *Segmentation Validation Dataset* described in Section 2.1. It consisted of annotated key extrema points of the vibration waveforms, i.e., the opening and closing points, and the left and right lateral and medial peaks (recall Figs. 6 and 7(b)). The validation procedure was analogous to the one used by Lohscheller et al. [48]: using an auxiliary manual annotation tool, six expert examiners denoted 834 key points on the set of clinical VKG images yielding the total of 5004 annotations. The images were selected to represent different degradation levels (e.g., noise, blur, or presence of specular reflections) and various types of healthy and pathologic vocal fold vibrations that could influence the segmentation accuracy, regardless of particular clinical diagnoses. The annotated key point positions were compared to the mean of the other annotators' key points to verify the robustness of the annotations. The ground truth for the 834 key points was then established as the mean of the manually detected coordinates. This procedure ensured the quality of the annotators' performance. Examples of the annotated points, together with the locations of the points detected automatically by the VKG Analyzer are shown in Fig. 8. The segmentation accuracy of the tool was assessed as the distance errors between the automatically detected key points and those obtained by the manual procedure. We analyzed the errors in both the spatial and temporal domains.

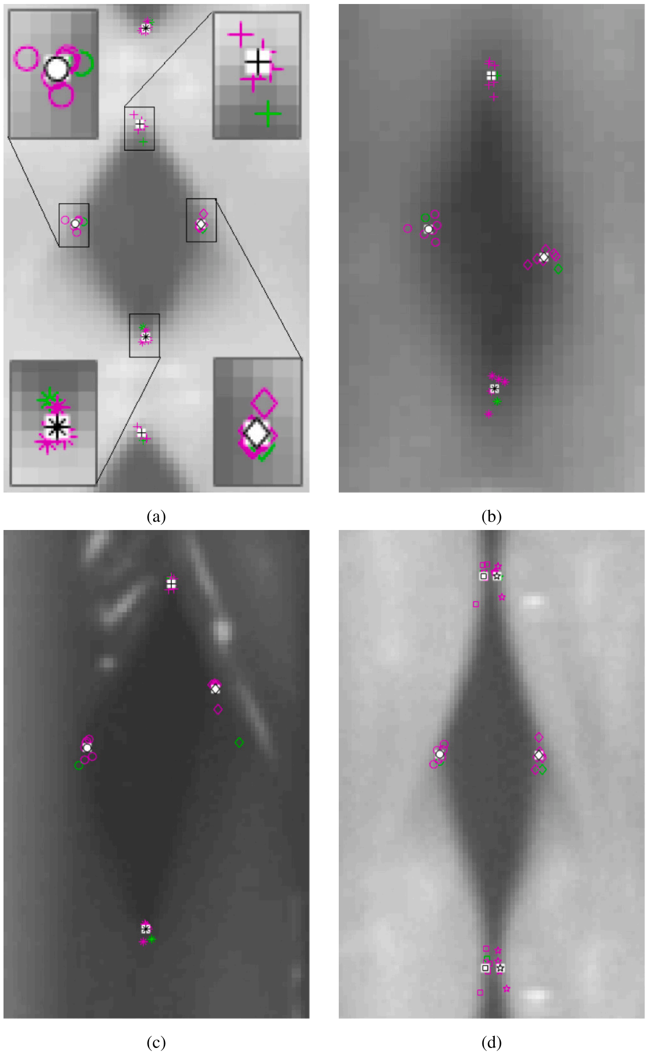
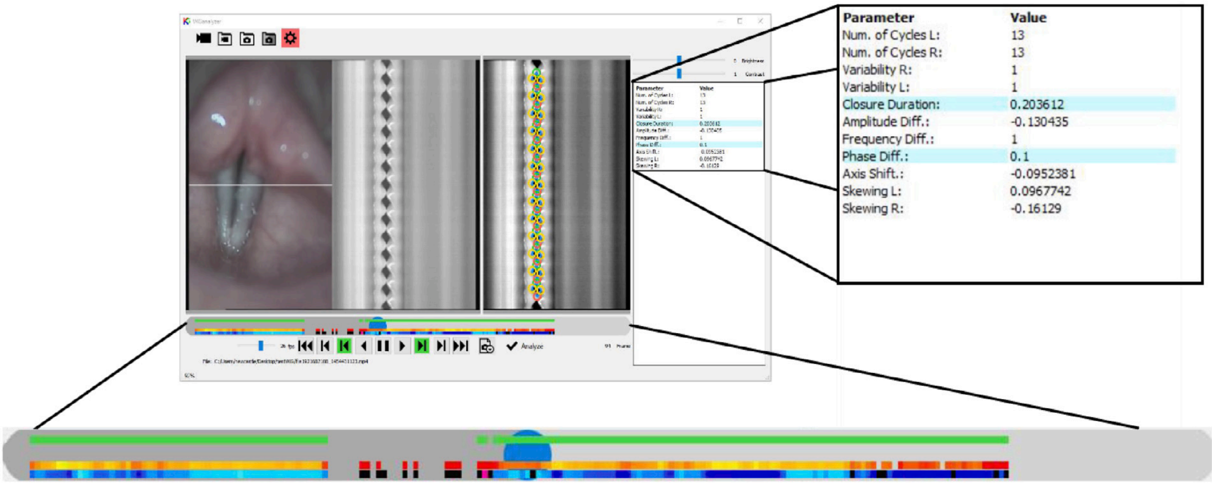


Fig. 8. Examples of the segmentation comparison study. The plus sign, star, diamonds, and circles denote the key points of opening, closing, left, and right lateral extremes, respectively. The magenta color codes positions selected by examiners, the white color codes the average of all examiners' positions, and the green color denotes the points automatically estimated by the algorithm. In image a), the selected areas are also magnified so that the pixelization of the images is clearly visible.



Parameter	Value
Num. of Cycles L:	13
Num. of Cycles R:	13
Variability R:	1
Variability L:	1
Closure Duration:	0.203612
Amplitude Diff.:	-0.130435
Frequency Diff.:	1
Phase Diff.:	0.1
Axis Shift.:	-0.0952381
Skewing L:	0.0967742
Skewing R:	-0.16129

Fig. 9. Program layout: the left pane shows the dual image as produced by the camera system; next to it, there is the VKG image after contrast and brightness normalization and image de-noising, together with the detected glottal contours and key points; the right pane shows the calculated features of the VKG recording; the bottom pane displays the slide-bar for video scrubbing, as well as the color-coded feature visualization bars.

The second verification study aimed to compare the automatically calculated vibratory features (**layer 3-4** Fig. 2) with those that were manually evaluated by human examiners using the pictogram-based VKG visual evaluation tool [31]. The visual evaluation data previously gathered by Hampala [24] were used for this purpose, partially forming the third dataset, i.e., the *Attributes Validation Dataset*. The original trial involved 50 VKG images obtained from 50 patients with various voice disorders, again showing the largest possible range of healthy and pathological vibratory patterns to which we added another 200 VKG images from 40 healthy subjects. Ten evaluators manually labeled the 50 images of pathological patients using the VKG visual evaluation tool [31], paying attention to 33 vibratory features per image. Eight of these evaluators performed the visual analysis twice for test–retest comparison purposes. This resulted in the total of 29 700 (50 × 18 × 33) manual evaluations. Nine out of the 33 features, thus 8 100 manual evaluations were selected for the purpose of this study. The 200 images of healthy patients were evaluated by three experts annotating the same nine features forming another set of 5400 (200 × 3 × 9) manual evaluations.

The resulting compilation of 250 visually-evaluated clinical images was then subjected to the objective image analysis by the VKG Analyzer tool. The feature values quantified by the software tool were discretized into the visually-based categories using the conversion tables defined in Section 3.5. The individual test–retest comparison results were divided into three categories – *Correct, Partially correct or indecisive, and Incorrect* [24]. Since the evaluation is fundamentally subjective, the label *Partially correct/indecisive* was introduced. In our study, it means that the algorithm misclassified the result to the neighboring category. E.g., “slightly larger amplitude difference” instead of “larger amplitude difference”, etc. This decision was a result of previous observations, that this level of error is common within human evaluation even for repeated evaluation by the same expert [24].

4. Results

4.1. VKG Analyzer tool — representation

The developed user interface (UI) of the tool (addressing **objective I**) is shown in Fig. 9. The emphasis was on visualization clarity and ease of use so clinicians could easily use the VKG Analyzer tool during routine patient examinations. The user interface is divided into 4 areas. The main part of the UI is used to visualize the kymographic image. The left top part shows the original recorded image. In Fig. 9 it is the dual image produced by the VKG camera providing the standard and VKG views, but it could also be a DKG or SVKG image. Next to it, there is the processed kymographic image together with the detected borders and estimated features. The top right pane shows a list of computed vibration attributes.

The bottom part of UI is designed to visualize the video timeline with color-coded values of relevant vibration attributes. A user can select a set of parameters for visualization in the right pane (see Fig. 9). This is helpful particularly for evaluations of video recordings performed with the VKG camera. The interactive visualization timeline slide-bar helps clinicians to find instances of interest directly, eliminating the time-consuming process of frame-by-frame visualization and analysis of the whole video recording. Additionally, the green line at the top of the bar indicates the information-rich video frames where oscillations were detected and which were marked during the preselection phase. All the processed data can be stored for later analysis, and the stored records can be analyzed repeatedly. Furthermore, the analyzed data, e.g., the analyzed frames with the segmented contours and the extracted parameters can be exported and used for further external analyses.

4.2. Segmentation precision

The results of the first validation study addressing **objective II**, which focuses on the accuracy of the automatic segmentation tool, are revealed in Tables 10 and 11 showing the mean and standard deviations of the key point positions with respect to the human ground truth. These data are also visualized in Fig. 10.

In spatial domain (left–right accuracy), the mean difference between the software-detected individual key points and their average manual locations was always less than one pixel (refer to Table 10, last row, and to the horizontal differences between the average manual and software results shown in the individual graphs of Fig. 10). The smallest manual vs. automatic average difference was found for the right lateral peak (0.05 pixels) and largest one for the left medial peak (−0.73 pixels). The manual vs. automatic differences show considerable standard deviations, however (up to ±1.08 pixels for the right lateral peak), revealing that the software vs. average manual positions differed across different vibratory cycles. This variability is, nevertheless, comparable to the uncertainty of the manual location of the key points (the largest standard deviation was ± 0.92 pixels for the left lateral peak, see Table 10, second last row), thus suggesting that the software inaccuracy is similar to that of the manual evaluations. Considering all the key points together, the average difference between their automatic and manual locations was 0.12 ± 0.79 pixels (Table 10, last row, last

Table 10
Manual and automatic segmentation accuracy in spatial domain (left–right accuracy), expressed in pixels. Average manual locations of the key points (refer to Fig. 6)) were used as the reference (zero) points. Mean differences from the reference points and their variability (i.e., standard deviation) are shown for the individual raters (rows 1-6) and for the automatic (SW row) segmentation results for each key point. The mean row shows the uncertainty (i.e., standard deviation) of the manual location of the reference points. Last column provides the results for all the key points pooled together. For the up-down accuracy, see Table 11 and for graphical representation of these results, see Fig. 8 and Fig. 10.

	L lateral	L medial	Opening	Closing	R lateral	R medial	All Points
1	0.13 ± 0.66	−0.08 ± 0.43	−0.03 ± 0.45	0.01 ± 0.43	−0.44 ± 0.72	−0.01 ± 0.60	−0.07 ± 0.62
2	−0.39 ± 0.71	−0.29 ± 0.45	−0.01 ± 0.47	−0.04 ± 0.44	0.13 ± 0.73	0.32 ± 0.37	−0.05 ± 0.63
3	−0.68 ± 0.97	−0.19 ± 0.52	−0.22 ± 0.44	−0.22 ± 0.46	0.33 ± 0.77	0.42 ± 0.51	−0.09 ± 0.79
4	1.17 ± 1.24	−0.13 ± 0.66	0.17 ± 0.45	0.27 ± 0.56	−0.57 ± 1.03	−0.09 ± 0.41	0.14 ± 1.08
5	0.51 ± 0.98	1.15 ± 1.07	0.01 ± 0.51	−0.05 ± 0.49	−0.09 ± 0.81	−0.79 ± 0.64	0.12 ± 0.81
6	−0.75 ± 0.80	−0.47 ± 0.55	0.09 ± 0.42	0.04 ± 0.43	0.63 ± 0.86	0.14 ± 0.36	−0.05 ± 0.83
Mean	**±0.92**	**±0.65**	**±0.46**	**±0.47**	**±0.83**	**±0.49**	**±0.81**
SW	**−0.2 ± 0.91**	**−0.73 ± 0.34**	**0.13 ± 0.51**	**−0.17 ± 0.49**	**0.05 ± 1.01**	**0.29 ± 1.08**	**−0.12 ± 0.79**

Table 11
Manual and automatic segmentation accuracy in temporal domain (up–down accuracy), expressed in pixels. The organization of the Table is identical to that in Table 10. For graphical representation of these results, see Fig. 8 and Fig. 10.

	L lateral	L medial	Opening	Closing	R lateral	R medial	All Points
1	0.02 ± 1.10	−0.01 ± 1.57	−0.5 ± 0.94	−0.63 ± 0.92	−0.13 ± 0.98	−0.06 ± 1.71	−0.22 ± 1.06
2	0.86 ± 1.07	−0.19 ± 1.74	0.05 ± 1.22	−0.04 ± 1.12	0.4 ± 1.01	−0.45 ± 2.72	0.11 ± 1.24
3	−0.62 ± 1.24	1.36 ± 1.92	−0.28 ± 1.35	0.04 ± 1.04	−0.24 ± 0.89	0.91 ± 2.18	0.2 ± 1.25
4	0.45 ± 1.33	0.84 ± 1.96	1.72 ± 1.56	−0.55 ± 1.26	0.23 ± 1.23	1.48 ± 2.35	0.7 ± 1.60
5	−0.41 ± 1.39	−2.25 ± 2.75	0.41 ± 1.10	−0.3 ± 0.96	−0.16 ± 1.07	−1.85 ± 2.39	−0.76 ± 1.34
6	−0.3 ± 0.89	0.25 ± 1.44	−1.41 ± 1.09	1.47 ± 1.22	−0.11 ± 0.88	−0.04 ± 1.95	−0.02 ± 1.45
Mean	**±1.18**	**±1.94**	**±1.23**	**±1.09**	**±1.02**	**±2.24**	**±1.33**
SW	**0.18 ± 1.19**	**0.03 ± 1.35**	**0.87 ± 1.19**	**−1.15 ± 1.04**	**0.9 ± 1.44**	**0.43 ± 1.63**	**0.21 ± 1.48**

column), revealing that the performance of the automatic segmentation is very similar to the manual one, even though there is variability across individual vibratory cycles and different key points.

In the time domain (up-down accuracy), the differences between the software and manual locations of the key points were mostly larger than those in the spatial (left–right) domain (Table 11). Also, the standard deviations were larger here, revealing larger variability of the differences between the manual vs. automatic locations of the key points as well as larger uncertainty of the manual location of the key points. This is visually reflected also in the plots of Fig. 10, mostly showing the error bars to be longer in vertical than in horizontal direction. The largest uncertainty was found for the manual location of the medial peaks (±1.9 and ±2.2 pixels for the left and right medial peak, respectively, see Table 11, second last row). The largest differences of the automatic positions from the manual averages were found for the Opening and Closing Point (0.9±1.2 and −1.2±1.0 pixels, respectively), and for the Right Lateral Peak (0.9±1.4 pixels, Table 11, last row). Considering all the key points together, however, the average difference between the manual and automatic locations was only 0.2 pixels with the standard deviation of ±1.48 pixels (Table 11, last row, last column) again revealing that the performance of the automatic segmentation is, on average, similar to the manual one.

4.3. Precision of attributes

The results of the second validation study addressing **objective II** and comparing the estimated vibration attributes to the clinician visual assessments are depicted in Fig. 11. For the healthy subjects' data, 91% of cases were in agreement with the human assessment. For the disordered patients, the software tool's performance agreed with the manual annotation assessments in more than 84% of cases.

5. Discussion

The goals of this work were to develop and test a user-friendly software tool for automated analysis of clinical videokymographic recordings (**objective I**) and to perform a rigorous validation of the implemented segmentation and feature extraction algorithms (**objective II**).

Both objectives were fulfilled. The developed VKG Analyzer tool facilitates selecting and exporting individual frames from the video recordings (see layer 2 in Fig. 2) and provides the means for automatically segmenting the vibrating glottal contours and for detecting key points of vocal fold vibration (layer 3 in Fig. 2). These data can then be subjected to automated feature extraction (layer 4 in Fig. 2).

Because videokymographic data have a different structure than standard laryngoscopic images, novel algorithms had to be developed and tested in order to facilitate proper kymographic image processing. During algorithm design, software implementation and validation, a number of noteworthy issues arose, which are being discussed in the following paragraphs.

5.1. Segmentation method

The approach utilized here differs from the previously explored image segmentation algorithms. While numerous segmentation methods have been developed to process high-speed videolaryngoscopic images [49], these cannot be utilized in VKG recording processing because the input images have different formats and meanings. To achieve the best segmentation results, we have experimented with the Active Contours (Snakes) approach (used for example in [28,50]) but ultimately opted not to use it due to the higher time demands and dependency on good initialization. Other methods we experimented with were the Region Growing methods [48,51,52], Graph-cuts [53], classical thresholding approaches like Otsu thresholding [54], watershed [55], and others. The Region Growing methods were found to be slow and dependent on good initialization. Graph-cut algorithms were promising initially, but in the end, they were hard to initialize correctly. Finally, the standard thresholding methods were fast but did not produce satisfactory results. The problem of correct initialization of certain methods is a circular one. Usually, the initialization consists of identifying pixels inside the glottal opening, but when known, the segmentation is not needed in the first place.

Our final solution is based on a handcrafted segmentation algorithm for finding the best threshold for segmentation. This approach has proven to be both robust and fast. In contrast to the Snakes algorithm (≈0.5 s per frame [50]), our implementation runs in real-time (< 0.04 s per frame). The image pre-processing and segmentation methods use

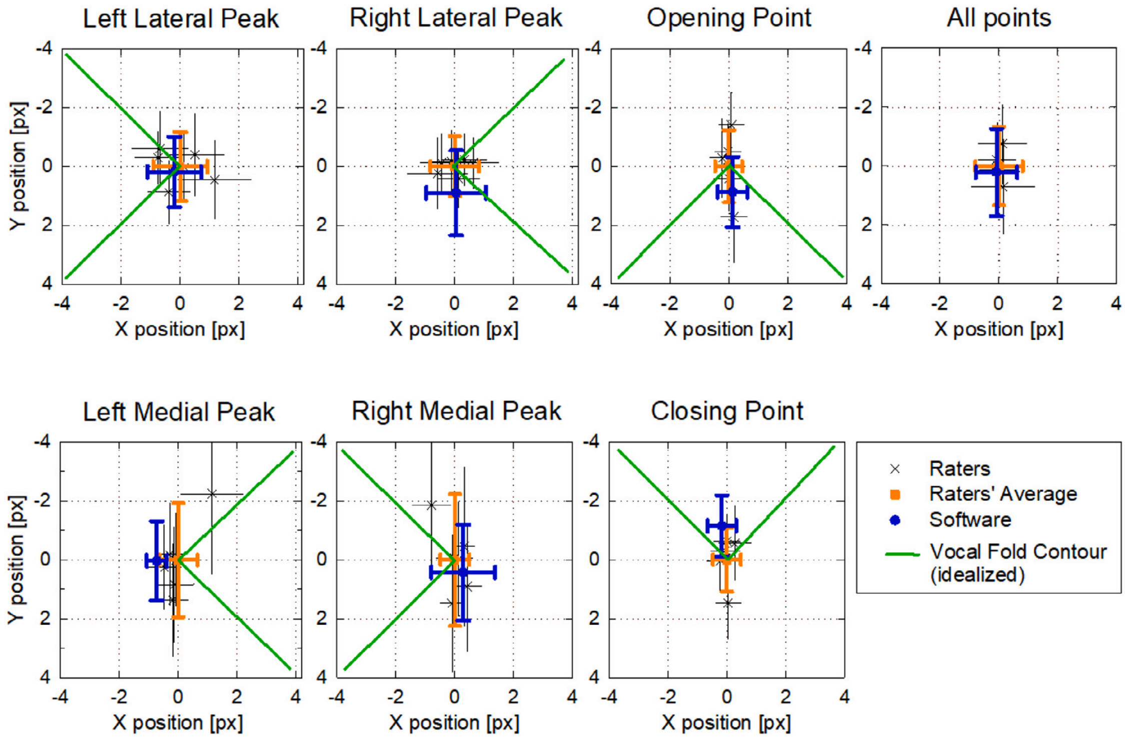


Fig. 10. Segmentation study — graphical representation of the results in Tables 10 and 11. The depicted crosses show the relative mean positions and their sizes in the *x*-axis and *y*-axis represent the standard deviations in the spatial and temporal domains, respectively. The overall human average value is colored orange and represents the ground truth, or golden standard, to which the software results (in blue) were compared. The black crosses represent the six human annotators.

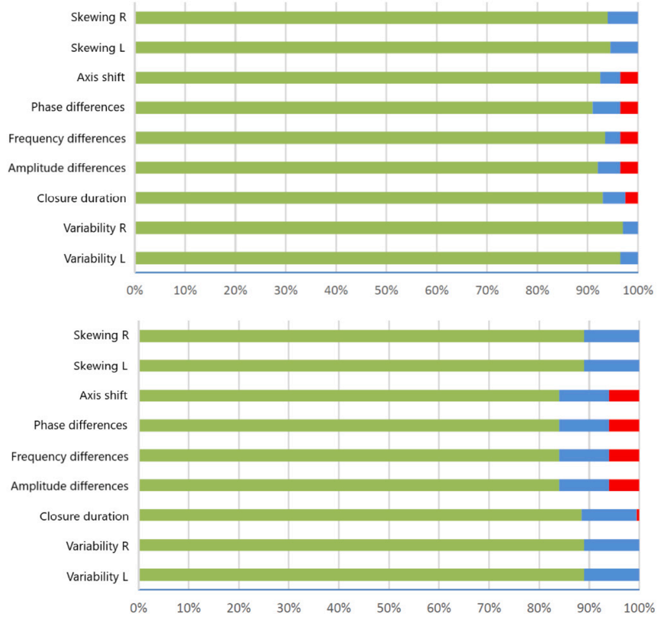


Fig. 11. Results of automatic features extraction by our program compared to visual evaluation by human experts for healthy (top) and disordered patients (bottom). The bars show percentage representation of correct results (green), partial correct (blue) and incorrect results (red) as evaluated by machine vs. human experts.

parameters and thresholds that needed to be determined empirically, however. To fine-tune these parameters, we used the *Training Dataset* (defined in Section 2.1) and performed a parameter search optimizing the resulting algorithm performance.

5.2. *Segmentation accuracy*

To test the correctness and robustness of our segmentation method, the automatic segmentation results were compared to the key point coordinates segmented manually. Considering the results across all the key points together, there were negligible differences between their automatic locations and the raters' manual average (recall Fig. 10, plot for All points). Furthermore, comparison of the error bars in the same plot reveals that the variability of the manual-to-automatic differences was very similar to the uncertainty of the manual location of the key points, suggesting that the performance of the automatic segmentation algorithm is comparable to the manual segmentation.

Nevertheless, there is a tendency of the software to locate the Opening Points about 1 pixel later, and the Closing Points about 1 pixel earlier, than the raters (Fig. 10, plots "Opening point" and "Closing point"). Taking into account the time running towards the bottom of the VKG image, this makes the duration of the open phase to be slightly shorter than when evaluated manually. This case is also reflected in annotations of the Opening and Closing points in Fig. 8(a) suggesting that manual annotators considered slightly different threshold between the vocal fold and the glottis — the software tends to locate the glottal boundary at slightly darker pixels inside the glottis than the raters. This tendency is detectable also in the plots for the Lateral and Medial peaks in Fig. 10 showing analogous, but much smaller shifts of the automatic key point locations to the right or to the left side, always towards the glottis (see the shifts of the blue versus the orange crosses in Fig. 10 horizontally). Nevertheless, considering the theoretical inaccuracy limit of 1 pixel, the observed differences between the manual and automatic evaluations smaller than c. 1 pixel are deemed acceptable. The tool can therefore be considered as a valid alternative to the manual procedures.

In this respect, it should be noted that manual annotators are not always consistent in their evaluations. Differences among the individual raters are visible in the spread of their annotations for the different key points (black crosses in the plots of Fig. 10). More specifically, Fig. 8(d) demonstrates the low precision of human experts particularly in the temporal (y-coordinate) domain for determining the Medial Peaks in cases of missing closure points. The low precision can be attributed to the roundedness of the contours making it difficult to locate the exact position of the peaks. The raters' uncertainty is reflected also in the large standard deviations (close to ±2 pixels) for the Left and Right Medial Peaks in Table 11 (second last row) and in the correspondingly large orange vertical error bars of the respective plots in Fig. 10.

The segmentation precision study underlines the strengths and weaknesses of our approach. The influence of the input data quality on the segmentation precision is shown in Fig. 8(c). The shift in left (right on image) lateral opening is caused by an imprecise segmentation threshold estimation due to the low image contrast. For purposes of the study, the contrast of the input image was unchanged, although the software tool allows a manual correction of contrast and brightness. The algorithm performed as expected for images with sufficient input image contrast (example can be viewed in Fig. 8(a)).

5.3. Feature extraction

Our VKG Analyzer tool implemented a larger amount of parameters than preceding tools for VKG analysis [28,29]. To enable easier interpretation of the numerical results for the clinicians, we derived empirical ranges for relating the numerical results of the quotients to descriptive categories defined in [31]. This made it possible to perform a comparative study, that aimed to address the precision and objectivity of extracted characteristics.

The result of the study (Fig. 11) show good software agreement with human examiners, namely in more than 91% of the cases for healthy patients (top graph) and more than 84% of the cases for disordered patients (bottom graph), depicted by the green segments of the graphs. We find this result satisfactory.

Additionally, the study revealed that in many cases, for the same image, the same examiner evaluated some attributes differently when the tests were performed several days apart. This experiment underlines the subjectivity of the task, and consequently, the difficulty of obtaining objective ground truth. To incorporate inconsistencies of the human evaluations into the study, we marked the mis-classifications to the neighboring categories as "Partially Correct" (see Fig. 11 blue segments of the graphs). A disadvantage of this approach is that it considers misclassifications on different attributes as equally significant, although different attributes have different interpretations and importance. Nevertheless, this approach allows good insight into the accuracy of the visual as well as visual versus automatic image assessment.

5.4. Additional software features

A noteworthy feature of the presented software is that it is designed to process not only VKGs, but also DKG and SVGK images. Furthermore, it allows to export the extracted glottal contours to a file in order to be analyzed by another means. These exported contour data, created by our tool, have already been successfully used in other detailed studies providing good applicability of the developed software framework [26,56,57].

5.5. Overall assessment

Results of both the validation studies indicate that the developed software is a valid, fast and robust automatic tool for vocal fold vibration analysis with minimal hardware requirements.

The comparison of the objectively measured attributes, which are automatically estimated by the developed software to visual assessments of ten evaluators makes this study unique. To the best of our knowledge, this is the first study that relates visual perception of such videokymographic features to objectively measured parameters. This rigorous and thorough validation ensures reliable application of the developed tool.

6. Summary

In the context of this study, we have developed and introduced a novel software tool for automated segmentation and feature extraction of all sorts of kymographic data (VKG, DKG, and even SVKG). The software is capable of automatically calculating the vocal folds' fundamental and derived vibration attributes. Additionally, it helps clinicians to focus on the information-rich sections of the VKG video recording by automatically pre-selecting such images from the recorded VKG examination session.

The software and its algorithms have been subjected to a rigorous validation at unprecedented scope, ensuring robust and reliable application in both a clinical and a research setting. Based on comparative results, the vibration attribute estimation demonstrated agreement with manual annotation in more than 91% (healthy patients) and 84% (disordered patients) cases. Owing to these outstanding validation results, the software is expected to become a robust and reliable state-of-the art tool for clinical and scientific examination of vocal fold vibrations and laryngeal function.

CRediT authorship contribution statement

Aleš Zita: Conceptualization, Methods research, Software, Writing. **Adam Novozámský:** Methods research, Software, Investigation, Writing. **Barbara Zitová:** Conceptualization, Supervision, Investigation, Writing. **Michal Šorel:** Methods research. **Christian T. Herbst:** Software, Writing. **Jitka Vydrová:** Investigation, Data acquisition, Medical expert and consultant, Validation. **Jan G. Švec:** Voice researcher and consultant, Data validation, Data acquisition, Writing.

Declaration of competing interest

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.bspc.2022.103878.

Acknowledgments

The research was supported by the Technology Agency of the Czech Republic projects no. TA04010877 (2014–2017) and no. TH04010422 "VKG 3.0" (2019–2021) and Czech Science Foundation, grant no. GA21-03921S (2021–2023). .

References

[1] W. Angerstein, G. Baracca, P. Dejonckere, M. Echternach, U. Eysholdt, F. Fussi, A. Geneid, T. Hacki, K. Karmelita-Katulska, R. Haubrich, et al., Diagnosis and differential diagnosis of voice disorders, in: Phoniatrics I, Springer, 2020, pp. 349–430.

[2] R.R. Patel, M.S. Harris, S.L. Halum, Objective voice assessment, in: Sataloff's Comprehensive Textbook of Otolaryngology: Head & Neck Surgery: Laryngology, Vol. 4, JP Medical Ltd, 2015, p. 155.

[3] R.R. Patel, S.N. Awan, J. Barkmeier-Kraemer, M. Courey, D. Deliyski, T. Eadie, D. Paul, J.G. Švec, R. Hillman, Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function, Am. J. Speech-Lang. Pathol. 27 (3) (2018) 887–905.

[4] D.M. Bless, R. Patel, N. Connor, Laryngeal imaging: stroboscopy, high-speed digital imaging, and kymography, in: The Larynx, Vol. 1, Plural Publishing San Diego, CA, Oxford, and Brisbane, 2009, pp. 181–210.

[5] J.G. Švec, H.K. Schutte, Kymographic imaging of laryngeal vibrations, Curr. Opin. Otolaryngol. Head Neck Surgery 20 (6) (2012) 458–465.

[6] C.A. Rosen, Stroboscopy as a research instrument: development of a perceptual evaluation tool, Laryngoscope 115 (3) (2005) 423–428.

[7] D.D. Mehta, R.E. Hillman, Current role of stroboscopy in laryngeal imaging, Curr. Opin. Otolaryngol. Head Neck Surgery 20 (6) (2012) 429.

[8] P. Woo, Stroboscopy and high-speed video examination of the larynx, in: R.T. Sataloff, M.S. Benninger (Eds.), Sataloff's Comprehensive Textbook of Otolaryngology: Head & Neck Surgery: Laryngology, Vol. 4, JP Medical Ltd, 2015, p. 193.

[9] D. Deliyski, Laryngeal high-speed videoendoscopy, in: K.A. Kendall, R.J. Leonard (Eds.), Laryngeal Evaluation: Indirect Laryngoscopy To High-Speed Digital Imaging, Thieme Medical, New York, 2010, pp. 245–270.

[10] G. Andrade-Miranda, Y. Stylianou, D.D. Deliyski, J.I. Godino-Llorente, N. Henrich Bernardoni, Laryngeal image processing of vocal folds motion, Appl. Sci. 10 (5) (2020) 1556.

[11] A.M. Kist, P. Gómez, D. Dubrovskiy, P. Schlegel, M. Kunduk, M. Echternach, R. Patel, M. Semmler, C. Bohr, S. Dürr, et al., A deep learning enhanced novel software tool for laryngeal dynamics analysis, J. Speech Lang. Hearing Res. 64 (6) (2021) 1889–1903.

[12] P. Gómez, A. Kist, P. Schlegel, D.A. Berry, D.K. Chhetri, M. Döllinger, Bagls, a multihospital benchmark for automatic glottis segmentation, Sci. Data 7 (1) (2020) http://dx.doi.org/10.1038/s41597-020-0526-3.

[13] M.K. Fehling, F. Grosch, M.E. Schuster, B. Schick, J. Lohscheller, Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional LSTM network, Plos One 15 (2) (2020) e0227791.

[14] A. Yamauchi, H. Imagawa, H. Yokonishi, K.-I. Sakakibara, N. Tayama, Multivariate analysis of vocal fold vibrations in normal speakers using high-speed digital imaging, J. Voice (2021) http://dx.doi.org/10.1016/j.jvoice.2021.08.002.

[15] K.A. Kendall, High-speed digital imaging of the larynx: recent advances, Curr. Opin. Otolaryngol. Head Neck Surgery 20 (6) (2012) 466–471.

[16] Q. Qiu, H.K. Schutte, A new generation videokymography for routine clinical vocal-fold examination, Laryngoscope 116 (10) (2006) 1824–1828.

[17] Q. Qiu, H.K. Schutte, Real-time kymographic imaging for visualizing human vocal-fold vibratory function, Rev. Sci. Instrum. 78 (2) (2007) 024302.

[18] J.G. Švec, F. Šram, Videokymographic examination of voice, in: E.P.M. Ma, E.M.L. Yiu (Eds.), Handbook of Voice Assessments, third ed., Plural Publishing, San Diego, CA, 2011, pp. 129–146.

[19] J.G. Švec, H.K. Schutte, Videokymography: high-speed line scanning of vocal fold vibration, J. Voice 10 (2) (1996) 201–205.

[20] T. Wittenberg, M. Tigges, P. Mergell, U. Eysholdt, Functional imaging of vocal fold vibration: digital multislice high-speed kymography, J. Voice 14 (3) (2000) 422–442.

[21] Y. Isogai, Analysis of the vocal fold vibration by the laryngo-strobography-improvements of the analytic function, Larynx Jpn. 8 (1996) 27–32.

[22] M.W. Sung, K.H. Kim, T.Y. Koh, T.Y. Kwon, J.H. Mo, S.H. Choi, J.S. Lee, K.S. Park, E.J. Kim, M.Y. Sung, Videostrobokymography: a new method for the quantitative analysis of vocal fold vibration, Laryngoscope 109 (11) (1999) 1859–1863.

[23] P. Krasnodębska, A. Szkiełkowska, B. Miaśkiewicz, H. Skarżyński, Characteristics of euphony in direct and indirect mucosal wave imaging techniques, J. Voice 31 (3) (2017) 383–e13.

[24] V. Hampala, Vizuální hodnocení videokymografických snímků u hlasových poruch [Visual evaluation of videokymographic features in voice disorders], (Master's thesis), Palacký University, Olomouc, Czech Republic, 2011.

[25] H.S. Bonilha, D.D. Deliyski, J.P. Whiteside, T.T. Gerlach, Vocal fold phase asymmetries in patients with voice disorders: a study across visualization techniques, 21, (1) 2012, pp. 3–15.

[26] S.P. Kumar, K.V. Phadke, J. Vydrová, A. Novozámský, A. Zita, B. Zitová, J.G. Švec, Visual and automatic evaluation of vocal fold mucosal waves through sharpness of lateral peaks in high-speed videokymographic images, J. Voice 34 (2) (2020) 170–178.

[27] C. Manfredi, L. Bocchi, S. Bianchi, N. Migali, G. Cantarella, Objective vocal fold vibration assessment from videokymographic images, Biomed. Signal Process. Control 1 (2) (2006) 129–136, Voice Models and Analysis for Biomedical Applications.

[28] C. Manfredi, L. Bocchi, G. Cantarella, G. Peretti, Videokymographic image processing: objective parameters and user-friendly interface, Biomed. Signal Process. Control 7 (2) (2012) 192–201.

[29] C. Piazza, S. Mangili, F. Del Bon, F. Gritti, C. Manfredi, P. Nicolai, G. Peretti, Quantitative analysis of videokymography in normal and pathological vocal folds: a preliminary study, Eur. Arch. Oto-Rhino-Laryngol. 269 (1) (2012) 207–212.

[30] P.H. Dejonckere, J. Lebacq, L. Bocchi, S. Orlandi, C. Manfredi, Automated tracking of quantitative parameters from single line scanning of vocal folds: A case study of the 'messa di voce' exercise, Logopedics Phoniatr. Vocology 40 (1) (2015) 44–54.

[31] J. Švec, M. Frič, F. Šram, H. Švecová, H. Schutte, Visually-based evaluation protocol for laryngeal videokymographic images, in: Proceedings AQL, 2006.

[32] J.G. Švec, F. Šram, H.K. Schutte, Videokymography in voice disorders: what to look for? Ann. Otol. Rhinol. Laryngol. 116 (3) (2007) 172–180.

[33] MATLAB Image Processing Toolbox, URL http://www.mathworks.com/products/image/.

[34] Open Source Computer Vision Library, URL http://opencv.org/.

[35] Qt, URL http://www.qt.io/.

[36] R.D. Hipp, Sqlite, 2020, URL https://www.sqlite.org/index.html.

[37] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vis. Graph. Image Process. 39 (3) (1987) 355–368.

[38] J. Serra, Image Analysis and Mathematical Morphology, Academic Press, London, 1988.

[39] N. Isshiki, Recent advances in phonosurgery, Folia. Phoniatr. (Basel) 32 (2) (1980) 119–154.

[40] Q. Qiu, H.K. Schutte, L. Gu, Q. Yu, An automatic method to quantify the vibration properties of human vocal folds via videokymography, Folia. Phoniatr. Logop. 55 (3) (2003) 128–136.

[41] J.G. Švec, F. Šram, H.K. Schutte, Videokymography, in: M.P. Fried, A. Ferlito (Eds.), The Larynx, third ed., Plural Publishing, San Diego, CA, 2009, pp. 253–274.

[42] P. Aichinger, F. Pernkopf, Synthesis and analysis-by-synthesis of modulated Diplophonic Glottal Area waveforms, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 914–926.

[43] D.D. Mehta, D.D. Deliyski, T.F. Quatieri, R.E. Hillman, Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings, J. Speech Lang. Hear. Res. 54 (1) (2011) 47–54.

[44] E. Dunker, B. Schlosshauer, Irregularities of the laryngeal vibratory pattern in healthy and hoarse persons, in: D.W. Brewer (Ed.), Research Potentials in Voice Physiology, State University of New York, Syracuse, NY, 1964, pp. 151–184.

[45] M. Hirano, Clinical Examination of Voice, Springer-Verlag, Wien, Austria, 1981.

[46] P. Moore, H. von Leden, Dynamic variations of the vibratory pattern in the normal larynx, Folia. Phoniatr. (Basel) 10 (4) (1958) 205–238.

[47] R. Timcke, H. von Leden, P. Moore, Laryngeal vibrations: measurements of the glottic wave. I. The normal vibratory cycle, AMA Arch. Otolaryngol. 68 (1) (1958) 1–19.

[48] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, M. Döllinger, Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos, Med. Image Anal. 11 (2007) 400–413, http://dx.doi.org/10.1016/j.media.2007.04.005.

[49] Y. Maryn, M. Verguts, H. Demarsin, J. van Dinther, P. Gomez, P. Schlegel, M. Döllinger, Intersegmenter variability in high-speed laryngoscopy-based glottal area waveform measures, Laryngoscope 130 (11) (2020) E654–E661.

[50] T. Shi, H.J. Kim, T. Murry, P. Woo, Y. Yan, Tracing vocal fold vibrations using level set segmentation method, Int. J. Numer. Methods Biomed. Eng. 31 (6) (2015) e02715.

[51] T. Wittenberg, P. Mergell, M. Tigges, U. Eysholdt, Quantitative characterization of functional voice disorders using motion analysis of high-speed video and modeling, in: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3, 1997, pp. 1663–1666, http://dx.doi.org/10.1109/ICASSP.1997.598831.

[52] J. Demeyer, T. Dubuisson, B. Gosselin, M. Remacle, Glottis segmentation with a high-speed glottography: a fully automatic method, in: 3rd Adv. Voice Funct. Assess. Int. Workshop, 2009.

[53] Y.Y. Boykov, M.-P. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in ND images, in: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vol. 1, IEEE, 2001, pp. 105–112.

[54] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.

[55] V. Osma-Ruiz, J. Godino Llorente, N. Saenz-Lechon, R. Fraile, Segmentation of the glottal space from laryngeal images using the watershed transform, Comput. Med. Imag. Graph.: Official J. Comput. Med. Imag. Soc. 32 (2008) 193–201, http://dx.doi.org/10.1016/j.compmedimag.2007.12.003.

[56] Z. Štanclová, Relationships between the vocal fold vibration parameters and voice intensity: A laryngeal high speed videoendoscopic study of a healthy woman. (In Czech), (Bachelor's thesis), Palacky University, Faculty of Science, Olomouc, the Czech Republic, 2021.

[57] H. Lehoux, L. Popeil, J.G. Švec, Laryngeal and acoustic analysis of chest and head registers extended across a three-octave range: a case study, J. Voice (2022) http://dx.doi.org/10.1016/j.jvoice.2022.02.014.