Biomedical Signal Processing and Control 78 (2022) 103854

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Analysis of Laryngeal High-Speed Videoendoscopy recordings – ROI detection

Tomáš Ettler *, Pavel Nový

Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic

ARTICLE INFO

Keywords:
Vocal cords
Glottis
Laryngeal High-Speed Videoendoscopy
Laryngotopography
Image segmentation
Discrete Fourier transform

ABSTRACT

Accurate detection of the glottis in video sequences obtained during the vocal cords examination using Laryngeal High-Speed Videoendoscopy (LHSV) is a prerequisite for the calculation of parameters describing vocal cord kinematics. This work presents the knowledge and methods related to the determination of the region of interest (ROI), which is one of the important steps in the processing of LHSV video sequences with the aim of automatic glottis detection. ROI is defined as the area between the vocal folds and anterior and posterior commissures.

A number of methods have been published on this topic, which are used mainly in experimental LHSV video processing systems. To determine the ROI, we decided to use a method based on frequency analysis of oscillations of the vocal cord anatomical structures, which, as we know, has not been used in this context yet. The oscillation is represented by the change of brightness of the corresponding pixels in the LHSV images. The ROI is then successfully detected even in the relatively heterogeneous structure of tissues and fluids and for videos of various qualities, including luminance reflections, where the movement of the vocal cords can be detected.

These methods extend the currently used system using a thresholding method for ROI detection and improve the success rate from 69% to 89%. These methods were tested on the LHSV video corpus, which contains 412 video sequences with different recording quality, diagnoses, and age groups of patients, obtained from ENT clinical practice.

1. Introduction

Laryngeal High-Speed Video Endoscopy (LHSV) is one of the standard vocal cord examination methods today. Because LHSV video observation alone or the use of analytical tools in proprietary software may not always be enough to evaluate vocal cord kinematics, supporting tools for vocal cord behavior analysis are being developed. They are based on methods of processing and analysis of individual vocal cord images, generally static analysis of a 2D image signal, or image sequences, generally, analysis of a 2D image signal distributed over time. To evaluate the behavior of the vocal cords, a set of parameters was defined. Some of them have a geometric basis and describe the detected glottis obtained by segmentation of the LHSV image. Such parameters can then be used to detect changes and irregularities during one or more periods of vocal cord movement. Using these parameters, the criteria used to evaluate the quality of vocal cord kinematics are further defined [1–3,19].

The basic task of LHSV video processing is glottis detection using segmentation tools. This segmentation is performed for each frame of the LHSV recording. A number of methods are used for such segmentation, their detailed and current overview is given in the publication [4] and another overview of methods is also mentioned in the publication [5]. We use two methods for glottis segmentation, Max-Min-Thresholding and Cluster Analysis (K-means), see [5,17,18].

The LHSV vocal cord images contain a relatively heterogeneous tissue structure, fluids or light artifacts may be present, images may be blurred, or their quality may generally vary. Therefore, it is advantageous to limit image processing to a region of interest (ROI) for subsequent successful glottis detection. The ROI is anatomically defined by the position of the anterior and posterior commissures and the maximum range of the vocal folds as shown in Fig. 1.

The complete processing of LHSV videos and the analysis of the parameters is schematically described in Fig. 1. We can specify the following five key points of processing:

(a) vocal cord examination and LHSV video recording;
(b) determination of the ROI, this step may also include an image preprocessing phase;

* Corresponding author.
E-mail address: thritton@kiv.zcu.cz (T. Ettler).

https://doi.org/10.1016/j.bspc.2022.103854
Received 3 March 2022; Received in revised form 12 May 2022; Accepted 4 June 2022
Available online 15 June 2022

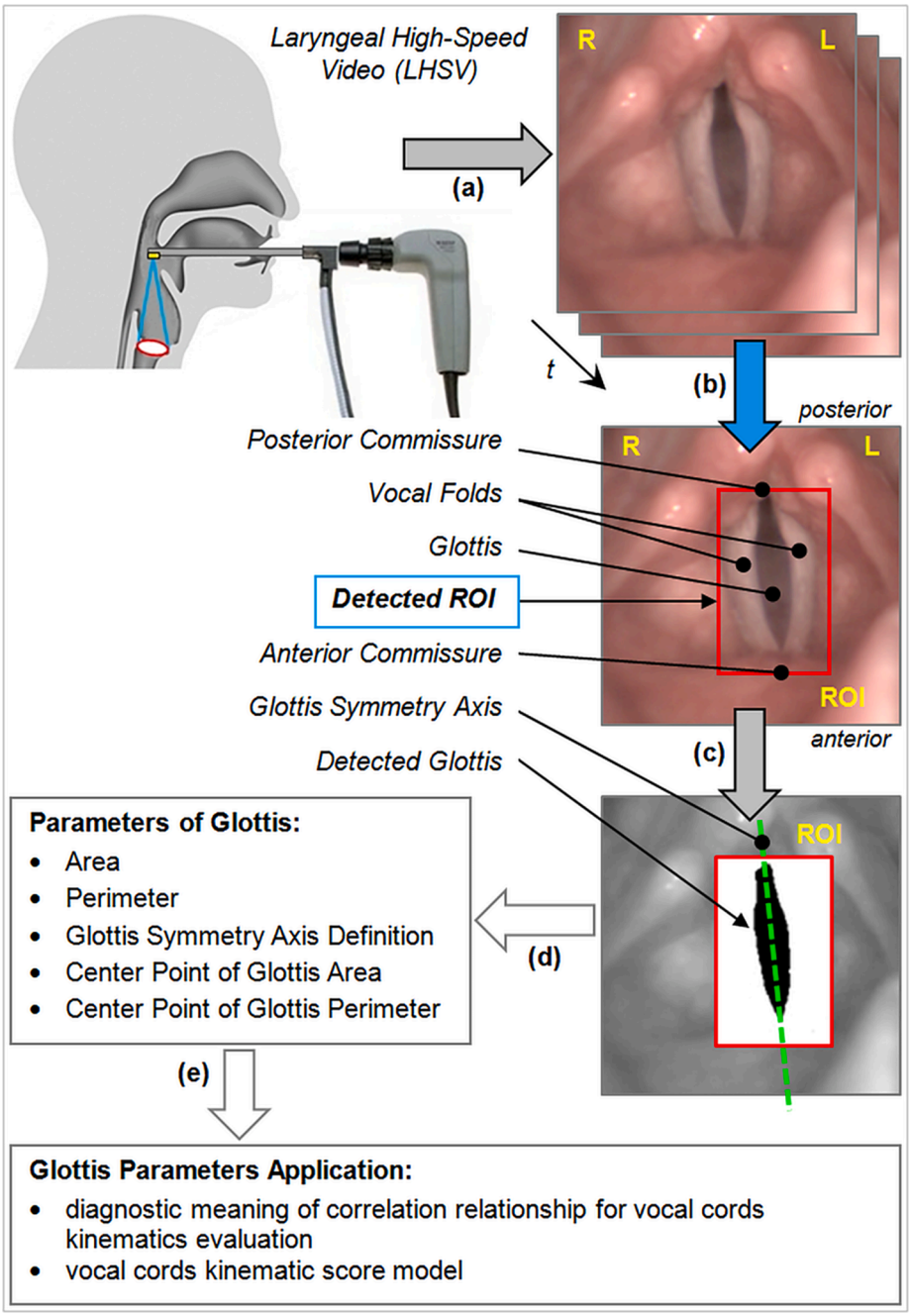

Fig. 1. Schema of the LHSV recordings processing, the similar approach as in [22]. (a) Recording of LHSV sequence. (b) ROI detection (image preprocessing + ROI detection methods application). (c) Glottis segmentation and detection, finding the glottis symmetry axis. (d) Computation of glottis parameters. (e) Vocal cords kinematics evaluation.

(c) glottis detection and glottis symmetry axis estimation;
(d) computation of defined glottis parameters;
(e) computation of criterion functions and analysis with the aim of detecting non-standard vocal cord behavior; vocal cords kinematics evaluation.

This article deals only with key point (b), Fig. 1(b), and presents new methods for determining the ROI area, including a comparison with the currently used method, called the *Thresholding method*, see below. Introduced methods use only LHSV data without prior knowledge of fundamental vocal cords oscillation frequency.

There are many publications describing automatic ROI detection in LHSV videos. One of the methods is described in [6,8] which is based on a region-growing algorithm with a selected seed point. Another approach using morphological operations was presented in [7]. Publications [9,12] mentioned ROI detection based on subtraction methods using more frames in the sequence, similar approach was used in [5], called *Thresholding method*, which was used for comparison with the new methods introduced in this article. Methods based on intensity variations of pixels are described in [10,11,13], motion estimation is then used in [14] and [15]. Publication [16] describes the salient region method using topological structural information. Other methods like Deep Neural Networks and the overall summary of methods used for ROI detection in LHSV recording or video frame were presented in [4].

It should be noted that according to the available information, the proprietary software provided with the LHSV systems allows only

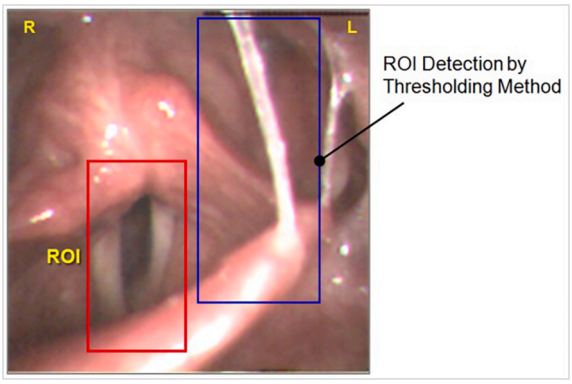

Fig. 2. Case report with a fluid where the *Thresholding method* of ROI detection failed.

manual ROI settings.

In the system for automatic ROI detection, see Fig. 1(b), we use our former method of thresholding, the *Thresholding method*, see [5,17]. This method is constructed on the principle of point operations over the sequence of frames in the LHSV video recording. This is based on the difference between the frame with the most closed vocal cord and the frame with the most open vocal cord to detect the part of the image where the brightness (or color components) changes the most. The ROI is then detected by thresholding of differential image, see [23], and using the Connected Component Labeling method, see [24], to find the largest continuous area. Although this ROI detection method achieves very good results for images with different quality, see [3,5,17,18], it fails in some cases. For example, when there is a significant movement of the camera, the occurrence of fluid (mucus, saliva) that moves randomly during phonation, or in the case of light reflections. An example of such failure can be seen in a case report with an occurrence of fluid in the left part of the vocal cord structure, see Fig. 2. This case study is further analyzed, see Fig. 3.

Because a well-defined ROI is a prerequisite for subsequent successful glottis detection, we decided to test a different approach and use the method using frequency analysis of the brightness (or color components) change of the pixel to evaluate the oscillation of the anatomical structure corresponding to this point. This method is called laryngotopography, see [20,21,25–27]. Based on the defined decision rules, the tested pixel is then included in the ROI or not. The methods created in this way will be called *DFT ROI detection methods*.

2. Methods

High-speed video laryngoscopy is an optical indirect rigid laryngoscopic examination technique, which is complemented by a high-speed camera with a frame rate of at least 1000 fps and a minimum resolution of 256 × 256 pixels. The device is thus able to capture and store a video sequence of the real movement of the vocal cords during their all oscillation phases. This allows detailed analysis of the movement of the vocal folds and the development of the glottis shape during the opening and closing phases. During the examination, patient phonates the vowel "i:", which make the supraglottical space for the camera the most accessible.

The LHSV video sequences that we process are recorded by the HSV HRES ENDOCAM 5562[1] system, with parameters of 4000 fps, 256 × 256 pixels. The video recording has a length of 600–1000 frames, i. e. 0.15–0.25 s.

During the examination, an acoustic signal is recorded, the system also allows the recording of an electroglottography signal (EGG)[2] using an external device. We do not use acoustic recording or EGG recording for ROI detection, we only work with visual information in LHSV images.

2.1. Principles used for DFT ROI detection methods

According to the physiology of voice formation and the characteristics of vocal cords behavior, we can use an approach for ROI detection, which is based on the oscillation of anatomical structures in LHSV frames. The movement of the vocal cords is repeated regularly and the periodicity contained in the movement of vocal folds corresponds to the fundamental vocal cord frequency F_0 (or near to F_0). This periodicity is subsequently reflected in the brightness change of individual pixels in the LHSV frames (value of brightness Y or the values of the color components R, G, B), see [20,21,25].

To detect the periodicity in the pixels of LHSV images, we use the discrete Fourier transform (DFT), defined for the finite number of samples of the equidistantly sampled input signal, and the method of spectral analysis of the signal using the DFT-amplitude spectrum. This input signal is the brightness or color component value at the pixel (x,y) of LHSV images.

The principle of *ROI DFT detection methods*, based on the analysis of the oscillation of anatomical structures, is presented in Fig. 3. This figure shows the development of the brightness change Y of LHSV monochrome images together with the DFT amplitude spectrum for selected pixels (x,y), see details (a), (d), (e). It also shows the overall distribution of pixels with the detected fundamental frequency F_0 and other, parasitic frequencies F_x, which do not correspond to the oscillations of the vocal cords, see detail (b), as well as the distribution of DFT-amplitude spectrum values (power spectrum), see detail (c).

In Fig. 3 (recording of the vocal cords with dg. recurrent laryngeal n. paresis in left), there is a case report where the vocal cords and related structures are overlapped by a slowly moving fluid (saliva, see also Fig. 2). This fact leads to unsuccessful ROI detection using the *Thresholding method*. On the contrary, methods based on the analysis of the oscillation of individual anatomical structures allow filtering out the pixels, which oscillate at frequencies F_x outside the defined frequency range. The pixel P_1 (x,y) anatomically corresponds to the edge of the right vocal fold, the pixel P_2 (x,y) belongs to the area representing the fluid in the image, and the pixel marked P_3 (x,y) belongs to the structure of the outer edge of the false vocal cords on the left. There is a noticeable difference between the oscillating anatomical structure P_1 (x,y) of the glottis, where the frequency corresponds to the fundamental frequency F_0, and the almost static anatomical structure of the false vocal cord, where it is not possible to determine the frequency of brightness change, see point P_3 (x,y).

There is also a noticeable difference in the estimation of the fundamental frequency F_0 at point P_1 (x,y) and the parasitic frequency F_x, which is caused by the movement of the fluid, see point P_2 (x,y).

For a description of the ROI detection methods, we will introduce the following notation and assumptions:

IMV … video sequence of LHSV frames;
(x,y) … pixel coordinates of video sequence IMV;
IMV_i … i frame of video sequence IMV, where $i = 0, 1, 2, \ldots, N-1$;
N … number of frames in video sequence IMV;
$pv(x,y,i)$ … pixel in frame IMV_i on position (x,y), $pv(x,y,i) \in IMV_i$.

We further understand the sequence of pixel brightness values $pv(x,y,i)$ for $\forall\, i = 0, 1, 2, \ldots, N-1$ as a finite sequence of real input samples, continuous in amplitude and discrete in time, the sampling period corresponds to the technically defined frame rate [fps] by the LHSV system. Therefore, DFT-Fourier spectrum $PV(k)_{(x,y)}$ and values of DFT-

[1] HRES Endocam, Richard Wolf GmbH, https://www.richardwolf.com, used on department of ENT, University Hospital in Pilsen.

[2] Two-Channel Electroglottograph, Model EG2-PCX2, https://www.glottal.com/Electroglottographs.html.

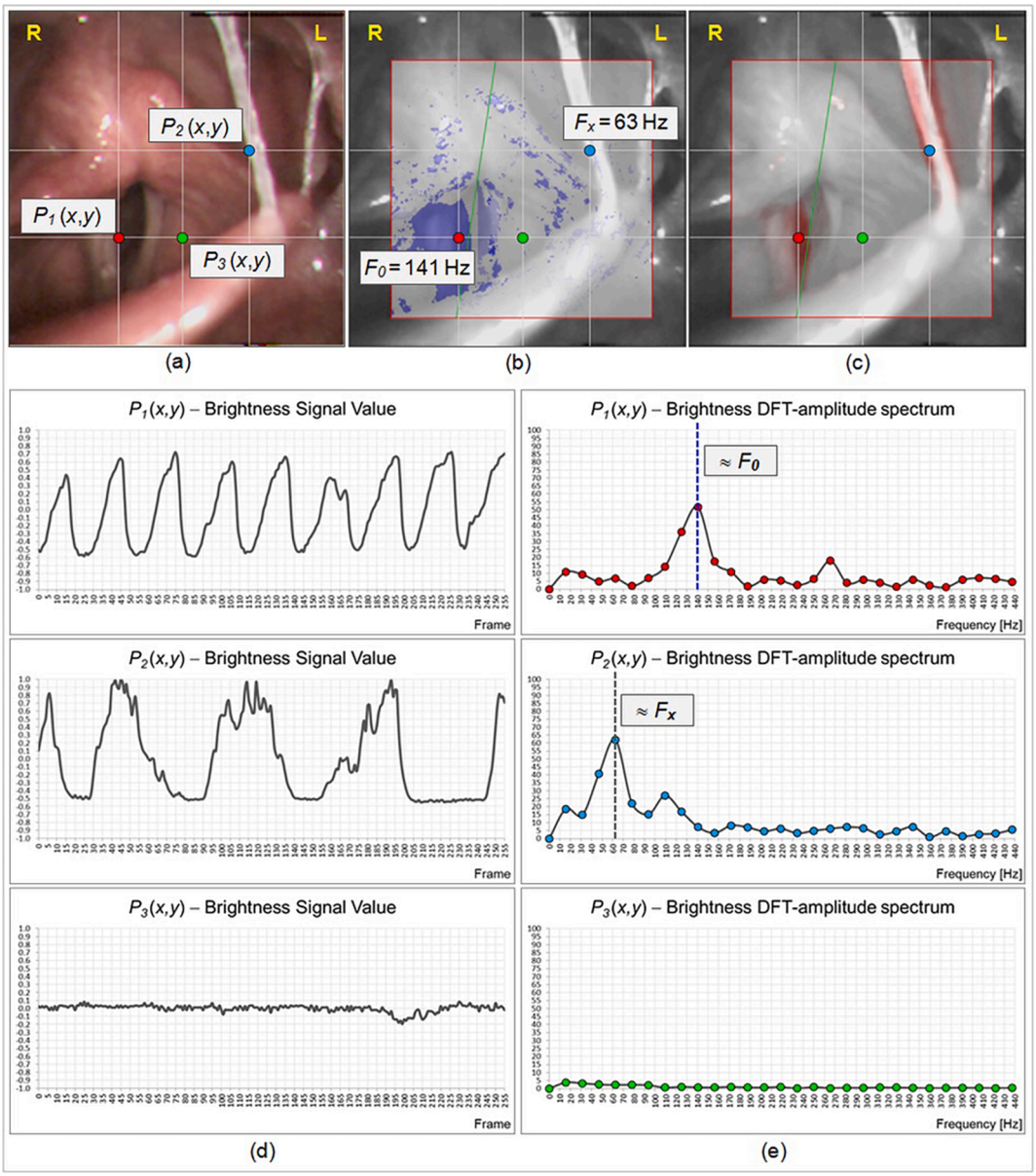

Fig. 3. Pixel brightness change and DFT-amplitude spectrum values for three points. $P_1(x,y)$ is in the glottis area, $P_2(x,y)$ is in the area with fluid and $P_3(x,y)$ is the anatomical structure of the outer edge of the false vocal cords on the left. There is an apparent difference in the most significant frequency in the glottis area ($\approx F_0$) and in the fluid area ($\approx F_x$). (a) Frame of video sequence *IMV* with marked points of individual selected anatomical structures. (b) Distribution of pixels with detected fundamental frequency F_0 and parasitic frequency F_x. (c) Distribution of DFT-amplitude spectrum values (Power spectrum). (d) The development of the brightness change Y of LHSV monochrome images for selected pixels. (e) DFT amplitude spectrum for selected pixels.

amplitude spectrum $|PV(k)|_{(x,y)}$ can be computed for such sequences of defined brightness values of $pv(x,y,i)$.

Let apply:

$$PV(k)_{(x,y)} = \sum_{i=0}^{N-1} pv(x,y,i) \cdot e^{-j\frac{2\pi}{N} i k} \quad \text{...Fourier spectrum} \tag{1}$$

for $\forall k = 0,1,2,\ldots,N-1$;

$\forall (x,y) \in IMV_i$ and $IMV_i \in IMV$ for $\forall i$, where $i = 0, 1, 2, \ldots, N-1$.

In terms of signal spectral analysis, we will focus on the values of the DFT-amplitude spectrum, where each point (x,y) of the *IMV* video sequence will be represented by a vector of values $|PV(k)|_{(x,y)}$, $k = 0,1,2,$

$$|PV(k)|_{(x,y)} = \sqrt{\left\{ Re\left[PV(k)_{(x,y)} \right] \right\}^2 + \left\{ Im\left[PV(k)_{(x,y)} \right] \right\}^2} \quad \text{...DFT-amplitude spectrum} \tag{2}$$

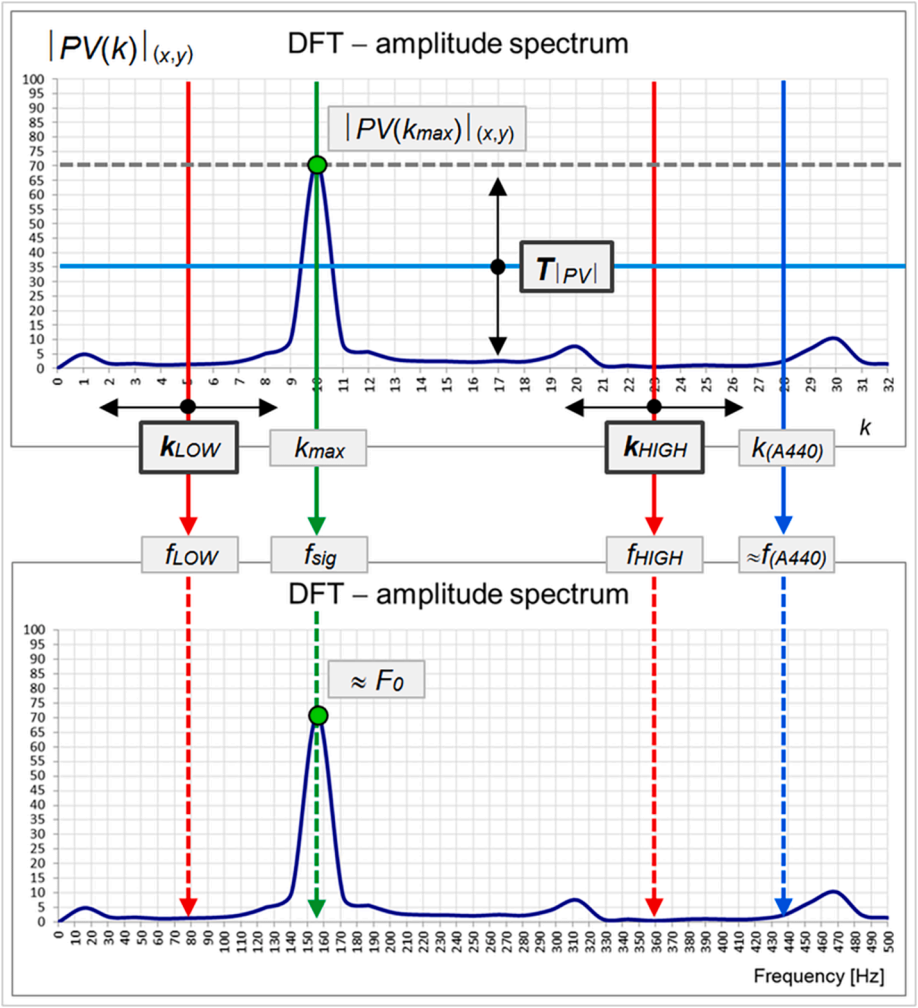

Fig. 4. Visualization of filter criteria of pixels (x,y) according to frequency $f_{sig}(x,y)$ and amplitude value $|PV(k_{\max})|_{(x,y)}$ in DFT-amplitude spectrum, where k_{LOW} and k_{HIGH} are indices in the DFT spectrum corresponding to the limited frequencies f_{LOW} and f_{HIGH}, see Eq. (10) and $T_{|PV|}$ is a threshold of DFT-amplitude spectrum value $|PV_{\max}|_{(x,y)}$, see Eq. (11).

…,$N-1$.

To determine the periodicity, i.e. the estimate of the value of the brightness change frequency Y (or color components R,G,B) at the point (x,y) of the *IMV* video sequence, we look for the maximum value of the DFT-amplitude spectrum for $\forall k = 1,2,\ldots,N/2$. The value of the DFT-amplitude spectrum for $k = 0$ corresponds to the DC component of the input signal, the choice of $N/2$ corresponds to the properties of the spectrum of the real sequence $pv(x,y,i)$.

For the index k_{max}, which corresponds to the most significant frequency of the change in brightness Y (or color components) in pixel (x, y), the following applies:

$$|PV(k_{\max})|_{(x,y)} = \max_{\forall k = 1,2,\ldots,N/2} \left\{ |PV(k)|_{(x,y)} \right\} \tag{3}$$

To calculate the frequency $f_{sig}(x,y)$ [Hz] of the brightness change we use the relation:

$$f_{sig}(x,y) = k_{\max} \frac{f_{sampl}}{N} = k_{\max} \Delta f, \tag{4}$$

where Δf is DFT grid (*frequency bin*):

$$\Delta f = \frac{f_{sampl}}{N} \tag{5}$$

f_{sampl} … LHSV sampling rate, i.e. frame rate [fps].

For the correct detection of pixels (x,y) as points that belong to the ROI, we also determine the conditions of the minimum and maximum detected frequencies f_{LOW} and f_{HIGH} and the minimal value of DFT-amplitude spectrum value $|PV(k)|_{(x,y)}$. If these criteria are met, we will consider the periodic movement of the anatomical structure at the position (x,y) to be significant and the point (x,y) will belong to the ROI, see Fig. 4.

Based on such assumptions, two methods for classification of pixels (x,y), if they are in ROI or not, were proposed: *DFT General method* and *DFT Geometrized method*.

2.2. DFT general method of ROI detection

The general algorithm is based on the assumption that each pixel at position (x,y) in the *IMV* video sequence is evaluated and decided whether is included it in the ROI or not.

Initial parameters of *IMV* video sequence:

- Frame resolution … 256 × 256 [px],
- Frame rate … 4000 [fps],
- Number of frames (N) … 256.

For every pixel (x,y) of LHSV recording, there is a sequence of discrete samples $pv(x,y,i)$, where $i = 0, 1, 2, \ldots, N-1$. This signal is normalized before further processing, and since it generally does not have to satisfy the condition of periodicity, see Eq. (6), we choose to

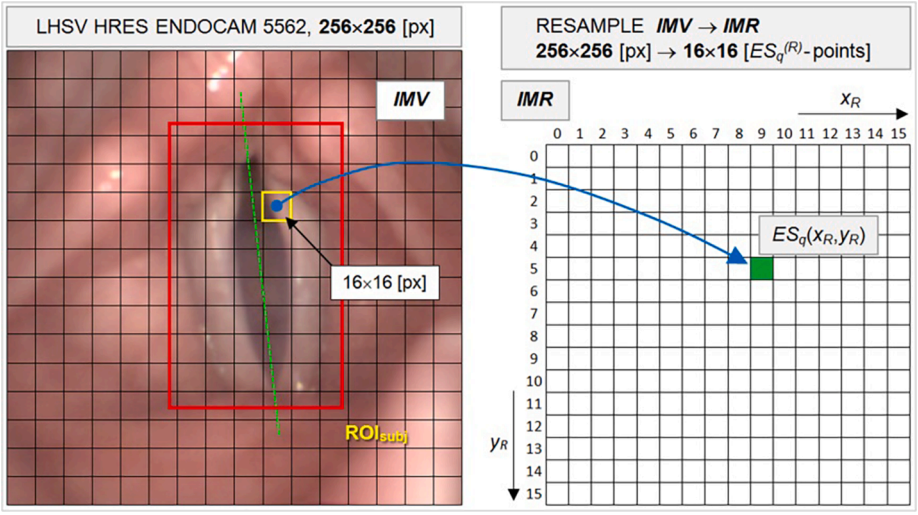

Fig. 5. Example of resampling the original *IMV* [LHSV HRES ENDOCAM 5562] frame (256x256) to an *IMR* frame of size (16x16), where one pixel (x_R,y_R) in the *IMR*, elementary square point $ESq(x_R,y_R)$, corresponds to the square which includes the area of 16x16 pixels (x_E,y_E) in the *IMV* frame.

adjust the signal $pv(x,y,i)$ by multiplying by the time window $w(i)$.

For simplicity and due to the required accuracy of the frequency resolution Δf, we choose a rectangular window[3] for $N_w = 256$.

$$N = rN_{period},\ r \in \boldsymbol{N}\ \ldots\ natural\ numbers,\qquad N = 2^m,\ m \in \boldsymbol{N}, \tag{6}$$

where N_{period} is the period of a discrete signal.

Then we can write for the normalized signal multiplied by the time window, see [26]:

$$pv^{w}_{norm}(x,y,i) = \frac{pv(x,y,i)}{\frac{1}{N}\sum_{i=0}^{N-1} pv(x,y,i)} - 1 \tag{7}$$

For a normalized signal, see Eq. (7), we calculate Fourier spectrum and DFT-amplitude spectrum values according to the formulas Eq. (1) and Eq. (2):

$$PV(k)_{(x,y)} = \sum_{i=0}^{N-1} pv^{w}_{norm}(x,y,i)\cdot e^{-j\frac{2\pi}{N}ik} \tag{8}$$

$$|PV(k)|_{(x,y)} = \sqrt{\left\{Re\left[PV(k)_{(x,y)}\right]\right\}^2 + \left\{Im\left[PV(k)_{(x,y)}\right]\right\}^2} \tag{9}$$

for $\forall k = 0,1,2,\ldots,N-1$.

For each pixel (x,y) of the *IMV* video sequence, we further determine the maximum value of the DFT-amplitude spectrum and determine the index corresponding to this maximum $k_{max}(x,y)$. This index $k_{max}(x,y)$, according to Eq. (3) and Eq. (4), corresponds to the frequency of change (periodicity) of the brightness component Y (or color components) at the point (x,y). We obtain for $\forall(x,y)$ an array of maximum amplitudes $|PV(k_{\max})|_{(x,y)}$ and their corresponding indices $k_{max}(x,y)$.

Because we work with the periodicity of the brightness Y (or color components) change, which corresponds to the anatomical structures of the vocal cords moving with the fundamental vocal cord frequency F_0, we can limit the frequency spectrum to the frequencies f_{LOW} and f_{HIGH} (Fig. 4) for further analysis and processing.

The values of the frequencies f_{LOW} and f_{HIGH} are based on the physiology of voice production and experience with LHSV signal processing. The purpose of this frequency limitation is to filter the changes in brightness caused by possible movement of the laryngoscope, light reflections during recording, movements of fluids, and movements of anatomical structures outside the area of vocal folds. These frequency limits are determined according to the F_0 range, which can be achieved during medical examination with a rigid laryngoscope and according to the real frequency distribution and DFT-amplitude spectrum values at pixels from the oscillating vocal cord area [21]. The publication [26] states the value of $f_{LOW} = 70$ Hz, the value of f_{HIGH} is limited from above by the standard concert pitch (A440), i.e. $f_{HIGH} = 440$ Hz (the idea was also demonstrated in Fig. 2 and Fig. 3). To verify the correctness of the f_{LOW} and f_{HIGH} settings, we performed a statistical analysis of the F_0 fundamental frequency values for the used data corpus. F_0 values are included in LHSV metadata and we also did our analysis of audio recordings, which are part of the LHSV recordset. Here we used the Autocorrelation function (ACF) and the methods of frequency analysis, Harmonic product spectrum (HPS) and Cepstrum analysis. The range of values of the fundamental frequency F_0 for the used LHSV corpus is $F_0 \in \langle 92\text{ Hz}; 420\text{ Hz}\rangle$.

For the limiting frequencies determined in this way, we then calculate the corresponding indices in the DFT spectrum, i.e. k_{LOW} and k_{HIGH}, see Eq. (10):

$$\begin{aligned} F_0 &\in \langle f_{LOW}; f_{HIGH}\rangle, \\ k_{LOW} &= f_{LOW}\frac{N}{f_{sampl}}, \\ k_{HIGH} &= f_{HIGH}\frac{N}{f_{sampl}}. \end{aligned} \tag{10}$$

In the next step, we calculate a histogram of the frequencies of occurrence of individual indices $k_{max}(x,y)$ for $\forall(x,y)$ with restriction to a defined interval of allowable frequencies $\langle k_{LOW}, k_{HIGH}\rangle$, i.e. for $\forall\, k_{max}(x,y) \in \langle k_{LOW}, k_{HIGH}\rangle$, and the minimal achieved value $|PV_{\max}|_{(x,y)}$, for $\forall(x,y)$, i.e. we include in the calculation of the histogram those pixels (x,y) for which the following will apply:

$$\{k_{\max}(x,y) \in \langle k_{LOW}, k_{HIGH}\rangle\}\ AND\ \left\{|PV(k_{\max})|_{(x,y)} \geqslant T_{|PV|}\right\} \tag{11}$$

where $T_{|PV|}$ is a threshold of value $|PV_{max}|_{(x,y)}$.

In the histogram of the selection adjusted in this way, we then select the most occurring index $k_{max}(x,y)$, which is understood as an index of representative periodicity and is an estimate of F_0 in *IMV*. This index is labeled as k_{ref} and the resulting amplitude array $|PV(k_{ref})|_{(x,y)}$ is compiled according to the original DFT-amplitude spectrum $|PV(k)|_{(x,y)}$ using this index k_{ref}.

To determine if the pixel (x,y) belongs to the searched ROI area, we

[3] It is possible to use Kaiser window {β = 0.5}, see [26], or e.g. Hamming window {256}, see [25,27].

use the threshold criterion:

$$\left| PV\left(k_{ref}\right) \right|_{(x,y)\,\in\,ROI} \geqslant T_{ROI} \cdot \max_{\forall\,(x,y)} \left\{ \left| PV\left(k_{ref}\right) \right|_{(x,y)} \right\} \tag{12}$$

where the threshold value $T_{ROI} \in (0;1)$.

The setting of the threshold values T_{ROI} and $T_{|PV|}$ was a result of the extensive testing of the *DFT methods*.

2.3. DFT geometrized method of ROI detection

Our goal is to detect the ROI, which delimits the area of all pixels (x, y) in the *IMV* video sequence corresponding to the anatomical structures of the vocal cords so that subsequent glottis detection can be performed. This approach can be geometrized. The aim is to decide on the inclusion of larger geometric units in the ROI and thus achieve the inclusion of the corresponding anatomical structure (anterior, posterior commissure, left and right vocal folds) with certain robustness. For this purpose, we choose to resample the original LHSV HRES ENDOCAM 5562, 256 × 256 [px] images to a geometric grid (the dimensions can be selected):

256 × 256 [px] → 128 × 128 [px], 64 × 64 [px], 32 × 32 [px], 16 × 16 [px] or 8 × 8 [px].

The individual squares, which are formed by applying the grid to resample, are called Elementary Squares $ESq(x_R,y_R) \equiv ESq^{(R)}$. Each of them is represented by a single point (x_R,y_R) in the resampled frame of the video sequence, see example in Fig. 5, the value of these points is calculated from the values of the brightness of the original pixels $(x_E,y_E) \in ESq^{(R)}$, see Eq. (13). Thus, the point (x_R,y_R) is a defined representative of the elementary square $ESq^{(R)}$ and according to the detected frequency of its oscillation and the value of the DFT-amplitude spectrum, the belonging of elementary square $ESq^{(R)}$ to the ROI is assessed, together with all pixels $(x_E,y_E) \in ESq^{(R)}$. This process eliminates the influence of noise in the DFT result and also decreases the number of computing operations.

For elementary square $ESq(x_R,y_R) \equiv ESq^{(R)}$ applies according to the above:

IMR … resampled video sequence *IMV*;

IMR_i … i frame from resampled video sequence *IMR*, where $i = 0, 1, 2, \ldots, N-1$;

(x_E,y_E) … the pixel coordinates of the *IMV* that belongs to the elementary square $ESq^{(R)}$ after the resampling;

$pv(x_E,y_E,i)$ … the brightness of a pixel belonging to an elementary square $ESq^{(R)}$ with (x_E,y_E) coordinates in frame IMV_i;

(x_R,y_R) … the pixel coordinates of the *IMR* that represents the elementary square $ESq^{(R)}$;

$pv_R(x_R,y_R,i)$ … the brightness value of the pixel that represents the elementary square $ESq^{(R)}$ in resampled frame *IMRi*.

Due to the nature of the ROI detection problem and the selected resampling geometry, we choose as the value of the representative pixel of the elementary square $ESq^{(R)}$ the average brightness, which we calculate form all pixels (x_E,y_E) belonging to this elementary square. We determine this value of the representative point in each frame of the video sequence.

$$pv_R(x_R, y_R, i) = \frac{1}{N_{ESq^{(R)}}} \sum_{\forall (x_E,y_E)} \sum_{\in ESq^{(R)}} pv\,(x_E, y_E, i) \tag{13}$$

where $N_{ESq(R)}$ is a number of pixels (x_E,y_E), which belong to $ESq^{(R)}$.

We apply a general ROI detection algorithm to the video sequence *IMR* created by resampling, see Eq. (6) to Eq. (12), which can be abbreviated in the following steps:

1) Normalization of the signal multiplied by the time window:

$$pv^{w}_{norm}\,(x_R, y_R, i) = \frac{pv_R\,(x_R, y_R, i)}{\frac{1}{N}\sum_{i=0}^{N-1} pv_R\,(x_R, y_R, i)} - 1 \tag{14}$$

2) Computation of the Fourier spectrum and DFT-amplitude spectrum values for normalized signal, see Eq. (14);

3) Determination of the maximum value of the DFT-amplitude spectrum and the index of this maximum $k_{max}(x_R,y_R)$ for each point (x_R,y_R) of the video sequence *IMR*;

4) Computation of the histogram of the frequencies of occurrence of individual indices $k_{max}(x_R,y_R)$ for $\forall\,(x_R,y_R)$ with limitation to the defined interval of allowable signal frequencies $\langle k_{LOW}, k_{HIGH}\rangle$ and at the same time the minimal achieved value $|PV_{\max}|_{(xR,yR)}$, for $\forall\,(x_R,y_R)$;

5) Selecting the index k^R_{ref} that corresponds to the index with the highest frequency of occurrence; k^R_{ref} will be an estimate of F_0 in *IMR*;

6) Construction of the resulting array of amplitudes that correspond to the index k^R_{ref};

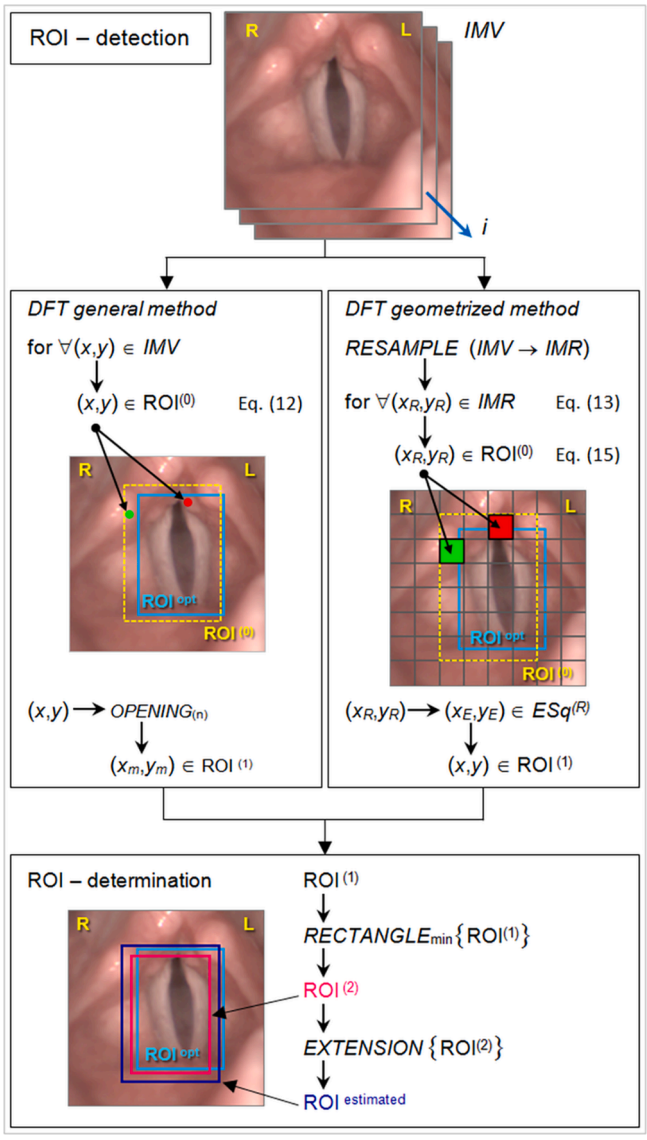

Fig. 6. Principle diagram of the procedure for obtaining the resulting ROI estimate ($ROI^{estimated}$) by *DFT General* and *DFT Geometrized* methods. In the case of the *DFT general method*, $ROI^{(0)}$ represents the initial assignment of pixels (x,y) to the ROI according to the criterion function Eq. (12). The $ROI^{(1)}$ area is then the result of a pixel filtering operation using the $OPENING_{(n)}$ method. For the *DFT geometrized method*, the $ROI^{(0)}$ is a set of individual elementary squares, which satisfy the criterion function Eq. (15). $ROI^{(1)}$ is then the set of those pixels (x,y) that belong to the individual elementary squares of $ROI^{(0)}$. The $ROI^{(0)}$ and $ROI^{(1)}$ are sets of pixels that generally do not form a rectangular area. The resulting area estimate $ROI^{estimated}$ is obtained by delimiting $ROI^{(1)}$ with a minimum circumscribed rectangle (=$ROI^{(2)}$) and then extended by a defined number of pixels to ensure covering whole vocal cords.

Table 1
LHSV data corpus structure used for testing ROI detection methods.

LHSV data corpus (No. 412)						
diagnosis	number of persons tested			video recordings		
	total sum	men	women	total sum	men	women
cystis vocal	1	0	1	7	0	7
vocal polyp	17	9	8	37	15	22
chordectomy	1	0	1	5	0	5
papillom	3	3	0	7	7	0
vocal nodules	9	3	6	19	5	14
carcinoma	5	5	0	5	5	0
granuloma	1	1	0	2	2	0
Reinke's edema	7	1	6	12	1	11
recurrent laryngeal n. paresis	40	8	32	107	18	89
tonsillectomy	0	0	0	0	0	0
hemangioma	1	0	1	1	0	1
thyroidectomy	23	2	21	39	4	35
vocal fold leukoplakia	1	1	0	4	4	0
chronic laryngitis	0	0	0	0	0	0
healthy vocal cords	21	1	20	48	2	46
dg. is not determined	52	28	24	119	58	61
total	182	62	120	412	121	291

7) Pixels (x_R,y_R), belonging to the target ROI, can be chosen by thresholding criterion:

$$\left| PV_{ref} \right|_{(x_R,y_R)\in ROI} \geqslant T^{R}_{ROI} \cdot \max_{\forall (x_R,y_R)} \left\{ \left| PV_{ref} \right|_{(x_R,y_R)} \right\} \quad (15)$$

where threshold value $T^{R}_{ROI} \in (0;1)$;

8) For all pixels with coordinates $(x_R,y_R) \in IMR$, which belong to the ROI according to the threshold criterion, see Eq. (15), we assign the corresponding pixels $(x_E,y_E) \in ESq(x_R,y_R)$; this determines the resulting set of pixels (x,y) from the *IMV* that belong to the ROI.

2.4. Determining of the results

We will approach the definition of the resulting ROI area in two ways depending on whether we used the *general* or the *geometrized* ROI detection method to determine the points $(x,y) \in IMV$ belonging to the ROI, see Fig. 6.

In the case of the *DFT General method*, after applying the criterion Eq. (12) we obtain the distribution of pixels $(x,y) \in IMV$ which meet the criteria of the frequency range F_0 and the minimum value of DFT-amplitude spectrum and form the first estimate of the ROI $\approx$ ROI$^{(0)}$. Due to the heterogeneous characteristics of *IMV* images, pixels (x,y) are also detected outside the anatomical structures of the vocal cords, e.g. in the form of isolated pixels. To filter such pixels, we use a combination of morphological transformations such as OPENING(n) (*Dilatation after Erosion*, symmetrical structural element 3 × 3) to depth *n*. The result is the elimination of unwanted artifacts in the form of isolated pixels (or small groups of pixels) and protrusions outside the anatomical structure of the vocal cords.

Setting the depth value *n* of the transformation $OPENING_{(n)}$ depends on the image data and is part of the algorithm heuristics. The result of this modification by the morphological transformation is a set of pixels $(x_m,y_m) \in IMV$, which corresponds to the ROI$^{(1)}$ area estimate. This set of pixels is surrounded by a minimum rectangle ROI$^{(2)}$, which represents an estimate of the ROI area. The minimum rectangle ROI$^{(2)}$ size is further adjusted - extended by the given number of pixels [px] in all directions to include the glottis even in special cases of limited vocal fold movement, e.g. for recurrent laryngeal n. paresis diagnosis.

The resulting rectangle ROI^{estimated} is then understood as an estimation of the ROI, which delimits the anatomical structure of the vocal cords between the anterior and posterior commissures and the maximum range of the vocal folds with a certain degree of robustness.

Table 2
Summary results of ROI detection methods for LHSV data corpus No.412.

ROI Detection	Thresholding method		DFT general method		DFT geometrized(8x8) method	
CORRECT	285	69.17%	347	84.22%	368	89.32%
FAILURE	116	28.16%	32	7.77%	35	8.50%
FAILURE TIGHT	11	2.67%	33	8.01%	9	2.18%
TOTAL	412	LHSV data corpus No. 412				

Table 3
The overall success rate of correct ROI detection using a combination of methods *Thresholding, DFT General, and DFT Geometrized (8x8)*.

Overall success methods:Thresholding, DFT general, DFT geometrized (8x8)		
Total number of recordings in LHSV data corpus No. 412	412	100.00%
at least one of the three methods is successful	398	96.60%
- of which all tested methods were successful	246	59.71%
- of which two of the three methods tested are successful	110	26.70%
- of which one of the three tested methods is successful	42	10.19%
neither method is successful	14	3.40%

DFT Geometrized method, created by resampling the *IMV* to the *IMR*, comprises, according to relation Eq. (13), filtering of the isolated pixels by averaging the brightness of the area $ESq^{(R)}$. Therefore, no further processing of the *IMR* is necessary before defining the baseline ROI$^{(1)}$ estimate. The initial set of pixels corresponding to ROI$^{(0)}$ is thus formed in the *IMR* by all pixels (x_R,y_R) that belong to the *IMR* and meet the condition of criterion Eq. (15). The ROI$^{(1)}$ estimate then consists of pixels $(x,y) \in IMV$, which correspond to all pixels (x_E,y_E) for which applies $(x_E,y_E) \in ESq(x_R,y_R)$.

As with the *DFT General method*, the ROI estimate in the form of the minimum circumscribed rectangle of the ROI$^{(1)}$ is extended by the given number of pixels in all directions to obtain a robust ROI estimate.

3. Results and discussion

A data set composed of 412 LHSV video sequences with the following parameters was used to test the achieved results, see Table 1.

As can be seen from the corpus structure according to Table 1, the corpus contains a diverse set of diagnoses and video recordings that are the source of a variety of anatomical structures and case reports in LHSV images. This diversity was useful for detailed testing of own ROI detection methods regardless of diagnoses and subsequent glottis segmentation methods. There are also multiple records of individual patients in the corpus, usually in the case of monitoring the progress of the disease or treatment, e.g. before and after microsurgery. From our point of view of testing ROI detection methods, such video recordings are unique and independent.

Based on the analysis and testing of video recordings from LHSV data corpus No. 412, the following parameters were set:

- $N = 256$,
- $f_{sample} = 4000$ Hz,
- $\Delta f = 15.625$ Hz … see (5)
- $k_{LOW} = 5$, $f_{LOW} = 78.13$ Hz,
- $k_{HIGH} = 28$, $f_{HIGH} = 437.50$ Hz … see (10)
- $T_{ROI} = 0.50$ … see (12)
- $n = 2$ … see $OPENING_{(n)}$,
- $(x_E,y_E) \in ESq^{(R)} = 8 \times 8$ [px],
- $T^{R}_{ROI} = 0.50$ … see (15)
- signal normalization … see (7), (14)

Complete results of comparison tests for LHSV data corpus No. 412 with parameters mentioned above are shown in Table 2. Used methods for ROI detection are compared there:

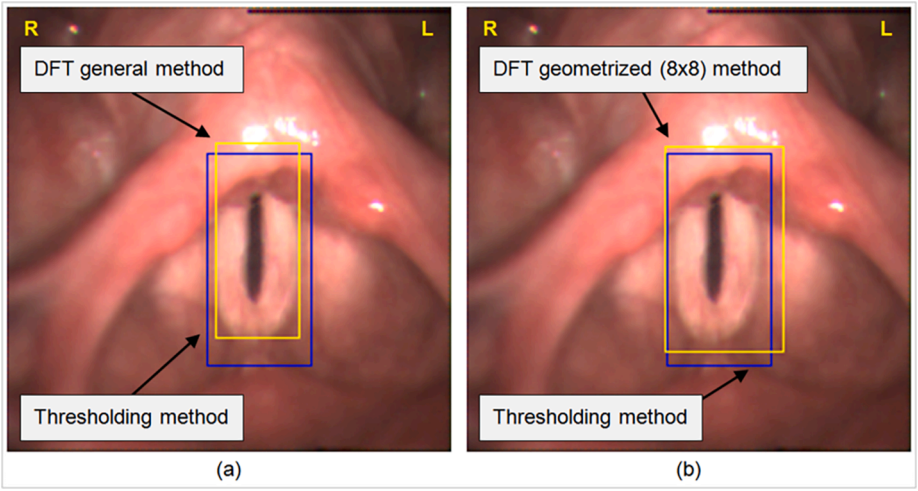

Fig. 7. Example of correct ROI estimation by all three methods.

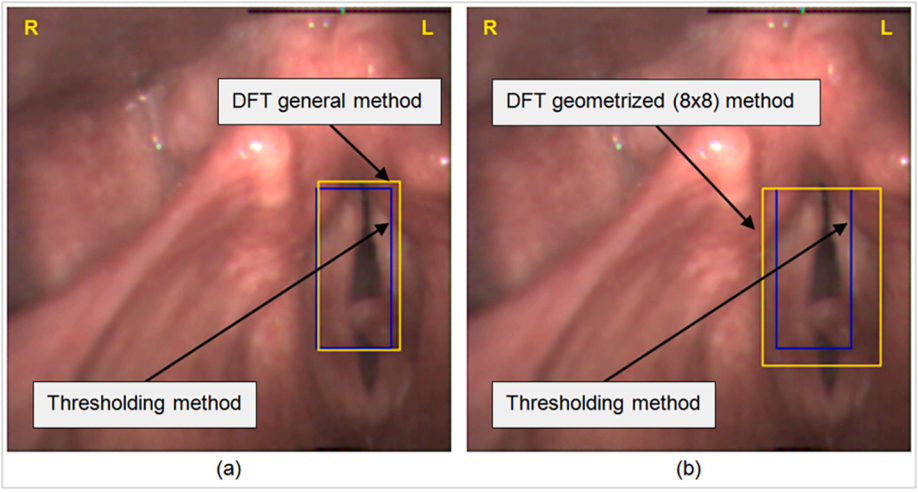

Fig. 8. Example of incorrect ROI estimation for all three tested methods.

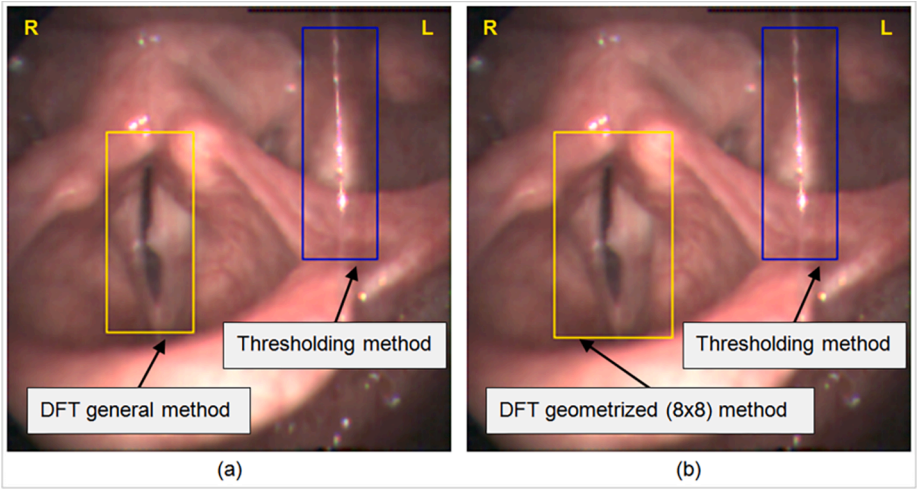

Fig. 9. Example of incorrect ROI estimation using the *Thresholding method* and correct detection by *DFT general* and *DFT geometrized (8x8)* methods.

Table 4
Compared ROI detection methods for LHSV data corpus No.412.

Compared ROI detection methods (a)			
Thresholding method → DFT general method			
TRUE – unchanged	TRUE → TRUE	255	61.89%
FALSE – unchanged	FALSE → FALSE	35	8.50%
ROI improved detection	FALSE → TRUE	92	22.33%
ROI worsened detection	TRUE → FALSE	30	7.28%
Compared ROI detection methods (b)			
Thresholding method → DFT geometrized (8x8) method			
TRUE – unchanged	TRUE → TRUE	265	64.32%
FALSE – unchanged	FALSE → FALSE	24	5.83%
ROI improved detection	FALSE → TRUE	103	25.00%
ROI worsened detection	TRUE → FALSE	20	4.85%
Compared ROI detection methods (c)			
DFT general method → DFT geometrized (8x8) method			
TRUE – unchanged	TRUE → TRUE	328	79.61%
FALSE – unchanged	FALSE → FALSE	25	6.07%
ROI improved detection	FALSE → TRUE	40	9.71%
ROI worsened detection	TRUE → FALSE	19	4.61%

Thresholding method … the former method of thresholding from [5,17],

DFT general method … method based on frequency analysis DFT in pixel *(x,y)*,

DFT geometrized (8x8) method … DFT method for (8 × 8) *IMV* pixels, which after resampling belong to the elementary square $ESq^{(R)}$.

The evaluation is subjective and is based on the assessment of the position and size of the ROI estimate with respect to the anatomical structures of the vocal cords. In this sense, we label the correct position and size of the ROI as TRUE and CORRECT, incorrect ROI estimate is labeled as FALSE. For evaluating incorrect results we distinguish between completely wrong ROI location, labeled as FAILURE, and cases where ROI location is correct but any part of vocal cords is not fully included in the ROI estimate, labeled as FAILURE-TIGHT. For example, the anterior or posterior commissure is not fully assigned to the ROI area, or the border of the ROI fits close to the left or right vocal fold. Such cases may not negatively affect further glottis segmentation. For summary results, see Table 2. DFT methods improve the success of ROI detection compared to the original thresholding method from 69.17% to 89.32%.

Because we assume the usage of a combination of all three tested

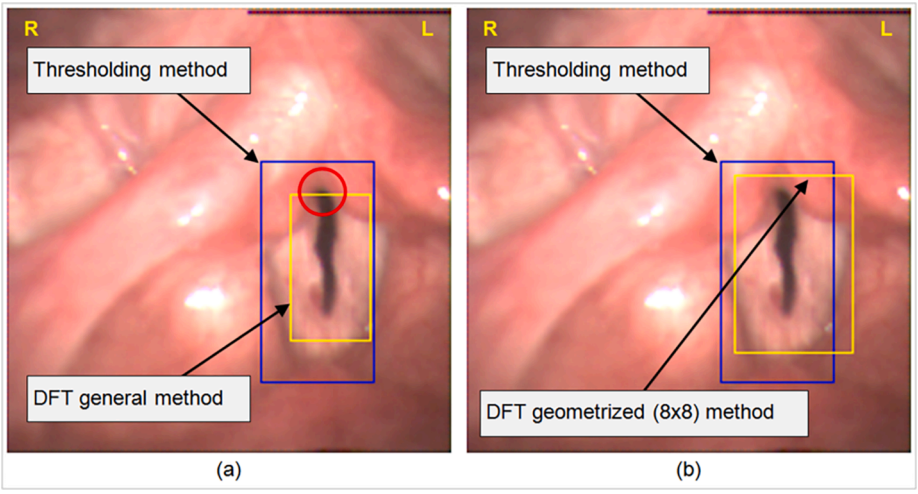

Fig. 10. Case report of failure of the *DFT general method* compared to the correct result of *Thresholding method*, Table 4(a). (a) Correct location of the ROI with a *tight* boundary of the anatomical structure in the area of the posterior commissure by the *DFT general method*. (b) Correct ROI estimation *by DFT geometrized (8x8) method*.

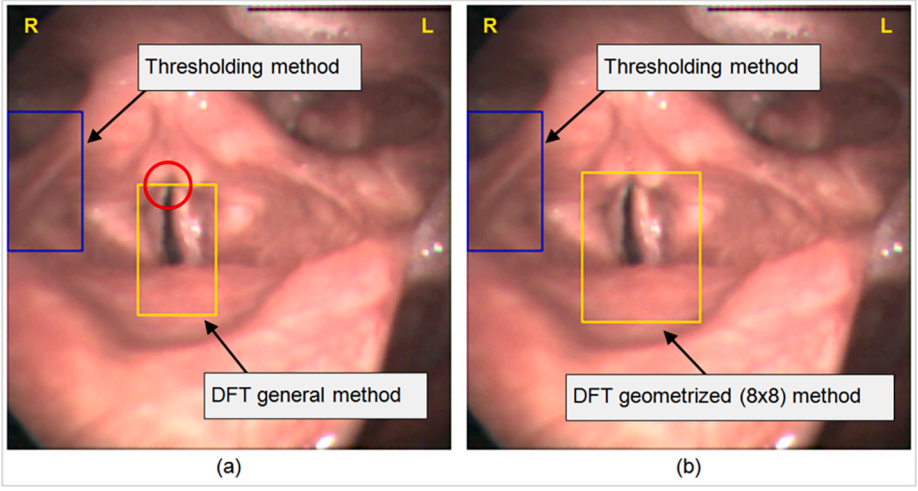

Fig. 11. Case report of ROI estimation using the *DFT general method*, where the ROI estimate did not improve against the *Thresholding method*, the result is also marked FALSE, Table 4(a). (a) Correct location of the ROI with a *tight* boundary of the anatomical structure in the area of the posterior commissure estimated by the *DFT general method*. (b) Correct ROI estimation by *DFT geometrized (8x8) method*.

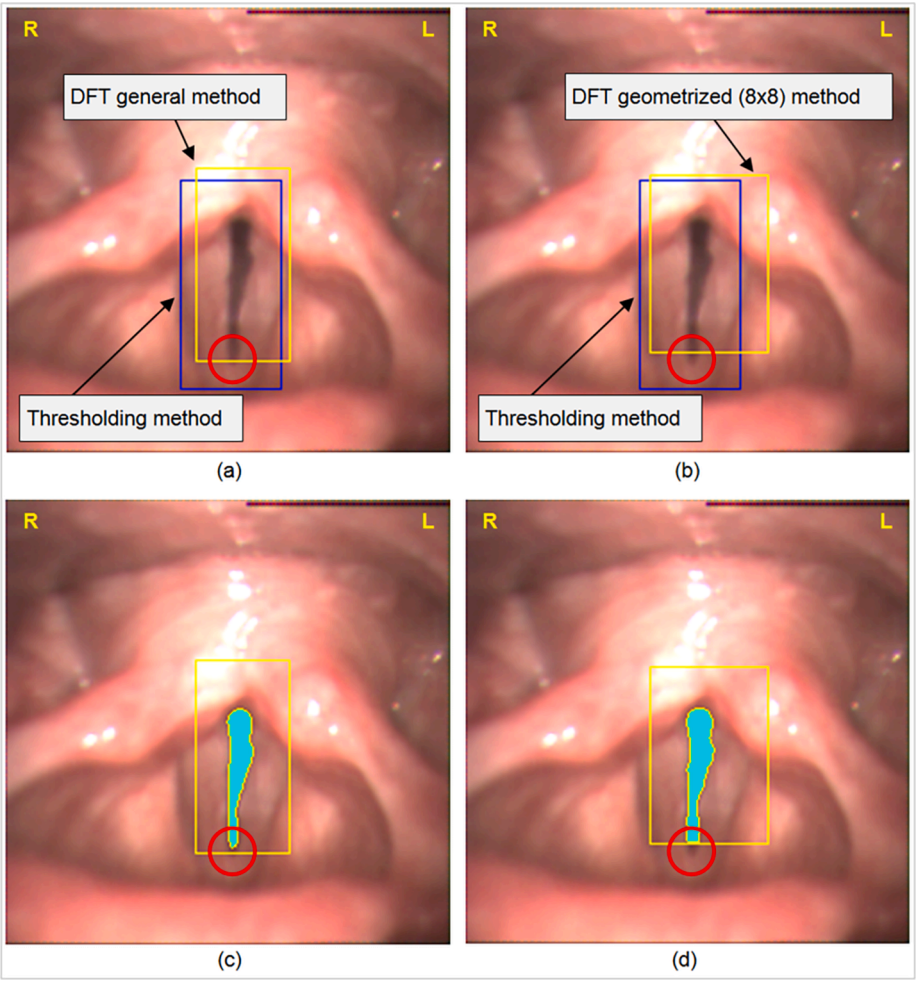

Fig. 12. Example of *DFT general* and *DFT geometrized (8x8) method* failure, results marked FALSE, compared to the *Thresholding method*, where the result is marked TRUE. (a) Correct ROI localization by *DFT general method*, a *tight* boundary in the anterior commissure area. (b) Correct ROI localization *DFT geometrized (8x8) method*, a *tight* boundary in the area of the anterior commissure. (c) The resulting glottis segmentation using ROI detected by the *DFT general method*. (d) The resulting glottis segmentation using ROI detected by the *DFT geometrized (8x8) method*.

methods in the system for ROI detection in the future, an overview of their overall success is also given, see Table 3. Out of a total of 412 tested videos, ROI was correctly estimated by at least one of the three ROI detection methods in 398 cases (96.60%). On the contrary, in 14 cases (3.40%) none of the methods achieved the correct result. Out of the total number of 398 successful ROI detections, correct ROI detection was achieved by three tested methods simultaneously in 246 video recordings (59.71%). The correct detection of ROI by two of three methods at the same time was achieved in 110 video recordings (26.70%) and by only one of the methods in 42 video recordings (10.19%).

Examples of ROI estimates according to Table 3 are in the following figures: Fig. 7, Fig. 8, and Fig. 9.

If we analyze 14 cases of video recordings when none of the three tested methods of ROI detection was successful (see Table 3), these were mostly cases where the vocal cords did not move, the movement was minimal, or a larger amount of fluid was moving in the image.

In the following, we focused on monitoring and analyzing changes in ROI detection results, always for method pairs in the following sessions, see Table 4:

(a) *Thresholding method → DFT general method,*
(b) *Thresholding method → DFT geometrized (8x8) method,*
(c) *DFT general method → DFT geometrized (8x8) method.*

When comparing the *Thresholding method* with the *DFT general method*, see Table 4(a), in 92 cases (22.33 %) of LHSV video recordings, the ROI estimate was improved (FALSE → TRUE), while in 30 cases (7.28 %) was worsened (TRUE → FALSE). However, after a detailed analysis of 30 cases of method failure in ROI estimate, it can be stated that in 16 cases (53 %) it is always the correct location of ROI, but due to the lower frequency of vocal structures in the anterior and posterior commissure area, the result was *tight* but not correct. Therefore, it is also subjectively classified as incorrect (FALSE), see case report Fig. 10(a), a marked part of the back commissure. It should be noted that by applying the *DFT geometrized (8x8) method* to the mentioned 16 video recordings with so-called *tight* ROI estimates by the *DFT general method*, we achieve a correct ROI estimate in 12 cases, see example Fig. 10(b).

We also analyzed 35 cases where the ROI estimate was not improved using the *DFT general method* (FALSE → FALSE). Also, in this case, there are almost half of the recordings, in 17 cases (48 %), where the ROI was correctly localized, but the ROI was *tight* or does not include all anatomical structures of vocal cords, like in Fig. 11(a). By applying the *DFT geometrized (8x8) method* to 17 videos with these *tight* ROI estimates, the estimate was improved in 13 cases, see example in Fig. 11(b).

Using the same methodology, we also compared the *Thresholding method* and the *DFT geometrized (8x8) method*, see Table 4(b). In 103 cases (25 %) of LHSV video recordings, the ROI estimate was improved using the *DFT geometrized (8x8) method* (FALSE → TRUE), while in 20 cases (4.85 %) of LHSV video recordings, the ROI estimate worsened (TRUE → FALSE). Among these 20 failures of ROI estimate, there are 5 cases (25 %) of video recordings, where the localization of ROI is correct, but again it is a so-called *tight* estimate, see Fig. 12(b).

There were 24 cases (5.83%) of cases where the use of the *DFT geometrized (8x8) method* did not improve the ROI estimate compared to the *Thresholding method* (FALSE → FALSE). 4 cases (17 %) can be classified as *tight* around commissures.

What effect has the so-called *tight* ROI estimate on the resulting glottis segmentation by the *K-means cluster analysis method*, see [3,5,18], we can see in the examples on Fig. 12(c) and (d). It is clear that the designation FALSE, spec. FAILURE-TIGHT is in many cases strict because even for a *tight* ROI defined in this way, the resulting glottis is determined correctly.

4. Conclusion

To diagnose the vocal cords using high-speed laryngoscopy (LHSV) and calculate the required parameters, correct segmentation of the glottis in each frame of the video sequence is necessary. Because the images may contain several disturbing artifacts and anatomical structures that are not directly related to the vocal cords or glottis, it is appropriate to define a Region of Interest (ROI). The ROI, if properly defined, delimits the anatomical space around the left and right vocal folds and between the anterior and posterior commissures.

In this work, we present two methods *DFT general* and *DFT geometrized (8x8)*, which are based on frequency analysis of pixel brightness oscillation, corresponding to anatomical structures of vocal cords. We compare the results with the already used and published method, called the *Thresholding method*. The presented methods are tested on a data corpus of LHSV video recordings, containing 412 video sequences with the various recording quality, with different diagnoses, and different age groups of patients.

The individual ROI detection methods are thus tested on a diverse set of LHSV data, which contains recordings without an obvious finding on the vocal cords, recordings with smaller or more significant findings of pathology up to findings that are close to the vocal cord dysfunction. Significant pathology, when the dynamics of the vocal cords are significantly disturbed, includes for example asymmetry, immobile vocal cords, glottis affected by neoplasms, etc.

By comparing the results of the three methods for data from LHSV corpus No. 412, methods constructed on the principle of frequency analysis are more successful than the existing *Thresholding method*. Specifically, the *Thresholding method* achieves successful ROI detection in 69% of cases, the *DFT general method* in 84% of cases, and the *DFT geometrized (8x8) method* in 89% of cases.

ROI estimation methods are used as one of the components of the processing and analysis of LHSV videos. For this reason, in addition to the standard manual ROI setting, we assume usage of all three tested ROI estimation methods as possible “suggestions” for setting the resulting ROI.

In this case, the success rate of the combination of ROI detection methods *Thresholding*, *DFT general,* and *DFT geometrized (8x8)* is 96.60% (398 out of 412), where at least one of the ROI detection methods is successful (subjectively assessed concerning the anatomy of the vocal cords).

For bulk processing, the parameters of the estimated ROIs are then part of the metadata for individual LHSV videos. When testing and comparing methods on LHSV data corpus No. 412, all methods were unsuccessful in 3.40% (14 of 412). In most cases, these are static vocal cords or vocal cords with significantly limited movement, where the role of the moving structure is taken over by another part of the anatomical structures in the image, or movement of fluids or reflection of light.

CRediT authorship contribution statement

Tomáš Ettler: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Investigation. **Pavel Nový:** Visualization, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Grant No. SGS-2019-018 Data and Software Engineering for Advanced Applications.

The authors thank their colleagues from the Department of ENT, University Hospital in Pilsen, Jiří Pešta (biomedical engineer – in memory) and Monika Vohlídková (ENT doctor, phoniatrician, audiologist) for professional consultations, providing expert knowledge and cooperation in creating a data corpus.

References

[1] I.R. Titze, Principles of Voice Production. 2nd ed., National Center of Voice and Speech: Iowa City, IA, USA, 2000, ISBN: 0-87414-122-2, pp. 87–183.
[2] J. Švec, Studium mechanicko-akustických vlastností zdroje lidského hlasu, [Studies on the mechanic-acoustic properties of the human voice. Thesis. In Czech]. Palacký University, Faculty of Natural Sciences, Department of Experimental Physics, Olomouc, 1996.
[3] J. Pešta, J. Slípka, M. Vohlídková, T. Ettler, P. Nový, F. Vávra, Kinematika hlasivek - nové parametry hodnocení, [Vocal Cord Kinematics – New Evaluation Parameters. Journal Publication. In Czech]. Otorinolaryngologie a foniatrie, 65, c. 2, pp. 88-96, Praha, 2016. https://www.prolekare.cz/en/journals/otorhinolaryngology-and-phoniatrics/2016-2/vocal-cord-kinematics-new-evaluation-parameters-58651.
[4] G.A. Miranda, Y. Stylianou, D.D. Deliyski, J.I. Godino-Llorente, N.H. Bernordoni, Laryngeal Image Processing of Vocal Folds Motion. Appl Sci 2020, 10, 1556. https://www.mdpi.com/2076-3417/10/5/1556.
[5] T. Ettler, Detekce a hodnocení videozáznamu pohybu hlasivek z vysokorychlostní kamery, [Detection and Evalution of Glottis in High Speed Video Recording. Professional work for the state doctoral exam. In Czech]. University of West Bohemia, Faculty of Applied Sciences, Department of Computer Science and Engineering, Pilsen, 2017. https://www.kiv.zcu.cz/site/documents/verejne/vyzkum/publikace/technicke-zpravy/2017/Rigo_Ettler_2017-1.pdf.
[6] 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, https://doi.org/10.1016/j.media.2007.04.005.
[7] S.Z. Karakozoglou, N. Henrich, C. D'Alessandro, Y. Stylianou, Automatic glottal segmentation using local-based active contours and application to glottovibrography, Speech Commun. 54 (5) (2012) 641–654, https://doi.org/10.1016/j.specom.2011.07.010.
[8] A. Pinheiro, M.E. Dajer, A. Hachiya, A.N. Montagnoli, D. Tsuji, Graphical evaluation of vocal fold vibratory patterns by high-speed videolaryngoscopy, J. Voice 28 (2014) 106–111, https://doi.org/10.1016/j.jvoice.2013.07.014.
[9] A. Skalski, T. Zielinki, D. Deliyski, Analysis of vocal folds movement in high speed videoendoscopy based on level set segmentation and image registration. In Proceedings of the International Conference on Signals and Electronic Systems (ICSES), Kraków, Poland, 2008; pp. 223–226. https://doi.org/10.1109/ICSES.2008.4673399.
[10] T. Koç, T. Çiloğlu, Automatic segmentation of high speed video images of vocal folds, J. Appl. Math. 2014 (2014) 1–16.
[11] G. Andrade-Miranda, J.I. Godino-Llorente, L. Moro-Velázquez, J.A. Gómez-García, An automatic method to detect and track the glottal gap from high speed videoendoscopic images, Biomed. Eng. Online 14 (1) (2015), https://doi.org/10.1186/s12938-015-0096-3.
[12] M. Blanco, X. Chen, Y. Yan, A restricted, adaptive threshold segmentation approach for processing high-speed image sequences of the glottis, ENG 05 (10) (2013) 357–362.
[13] G.A. Miranda, J.I. Godino-Llorente, Glottal Gap tracking by a continuous background modeling using inpainting, Med. Biol. Eng. Compu. 55 (2017) 2123–2141, https://doi.org/10.1007/s11517-017-1652-8.
[14] A. Mendez, E.M. Ismaili Alaoui, B. Garcia, E. Ibn-Elhaj, J. Ruiz, Glottal Space Segmentation from Motion Estimation and Gabor Filtering. 31st Annual International Conference of the IEEE EMBS, Minneapolis, Minnesota, USA, pp. 5756–5759, 2009. https://doi.org/10.1109/IEMBS.2009.5332612.
[15] E.M. Ismaili Alaoui, A. Mendez, E. Ibn-Elhaj, B. Garcia, Keyframes detection and analysis in vocal folds recordings using hierarchical motion techniques and texture information. In Proceedings of the 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009; pp. 653–656. https://www.researchgate.net/publication/224114786. https://doi.org/10.1109/ICIP.2009.5413745.
[16] F. Schenk, P. Aichinger, I. Roesner, M. Urschler, Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours. Annals of the BMVA Vol. 2015, No. 1 pp 1–15. 2015. https://www.researchgate.net/publication/282731724_Automatic_high-speed_video_glottis_segmentation_using_salient_regions_and_3D_geodesic_active_contours. www.bmva.org/annals/2015/2015-0003.pdf.
[17] T. Ettler, Analýza vysokorychlostního záznamu kmitání hlasivek, [Analysis of Vocal Cord Oscillations from High Speed Videolaryngoscopy Recordings. Diploma Thesis. In Czech]. University of West Bohemia, Faculty of Applied

Sciences, Department of Computer Science and Engineering, Pilsen, 2012. https://www.kiv.zcu.cz/site/documents/verejne/vyzkum/publikace/technicke-zpravy/2017/Rigo_Ettler_2017-1.pdf.

[18] T. Ettler, P. Nový, Using cluster analysis for image processing in high speed video laryngoscopy. International Conference on Applied Electronics, Proceedings, ISBN 978-80-261-0891-7, Department of Applied Electronics and Telecommunications, University of West Bohemia, Pilsen, 2020.

[19] T. Ettler, P. Nový, Diagnostic meaning of correlation relationship. 19th Conference on Applied Mathematics Aplimat 2020, Proceedings, ISBN 978-80-227-4983-1, Institute of Mathematics and Physics, Faculty of Mechanical Engineering, Slovak University of Technology in Bratislava, Bratislava, 2020. http://evlm.stuba.sk/APLIMAT/indexe.htm, https://www.proceedings.com/53722.html, [CrossRef].

[20] KIPS_ Kay's Image Processing Software_Documentation, Color High-Speed Video System (Model 9170), KIPS (Model 9181) [online]. KayPENTAX. [accessed 6 March 2009].

[21] Ch. Baierova, Frekvenční analýza kmitů hlasivkové štěrbiny, [Frequency Analysis of Vocal Cord Oscillations. Diploma Thesis. In Czech], University of West Bohemia, Faculty of Applied Sciences, Department of Computer Science and Engineering, Pilsen, 2018. https://www.kiv.zcu.cz/~novyp/dip/dp_baierova.pdf.

[22] P. Schlegel, M. Semmler, M. Kunduk, M. Döllinger, C. Bohr, A. Schützenberger, Influence of analyzed sequence length on parameters in laryngeal high-speed videoendoscopy, Appl. Sci. 8 (12) (2018) 2666.

[23] N. Otsu, A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4310076.

[24] L. Shapiro, G. Stockman, Computer Vision, Chapter 3.4 Connected Components Labeling. Prentice Hall. pp. 69−73, 2002.

[25] S. Granqvist, P. Lindestad, A method of applying Fourier analysis to high-speed laryngoscopy, J. Acoust. Soc. Am. 110 (6) (2001) 3193–3197, https://doi.org/10.1121/1.1397321.

[26] P. Aichinger, I. Roesner, B. Schneider-Stickler, W. Bigenzahn, F. Feichster, A. K. Fuchs, M. Hagmüller, G. Kubin, Spectral analysis of laryngeal high-speed videos: case studies on diplophonic and euphonic phonation, in: 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), 2013, pp. 81–84.

[27] K.I. Sakakibara, H. Imagawa, M. Kimura, H. Yokonishi, N. Tayama, Modal Analysis of Vocal Fold Vibrations Using Laryngotopography. INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, 2010.