sensors

MDPI

*Article*

# Segmentation of Glottal Images from High-Speed Videoendoscopy Optimized by Synchronous Acoustic Recordings

**Bartosz Kopczynski [1], Ewa Niebudek-Bogusz [2], Wioletta Pietruszewska [2] and Pawel Strumillo [1,*]**

[1] Institute of Electronics, Lodz University of Technology, 90-924 Lodz, Poland; bartosz.michal.k@gmail.com
[2] Department of Otolaryngology, Head and Neck Oncology, Medical University of Lodz, 90-001 Lodz, Poland; ewa.niebudek-bogusz@umed.lodz.pl (E.N.-B.); wioletta.pietruszewska@umed.lodz.pl (W.P.)
* Correspondence: pawel.strumillo@p.lodz.pl

**Abstract:** Laryngeal high-speed videoendoscopy (LHSV) is an imaging technique offering novel visualization quality of the vibratory activity of the vocal folds. However, in most image analysis methods, the interaction of the medical personnel and access to ground truth annotations are required to achieve accurate detection of vocal folds edges. In our fully automatic method, we combine video and acoustic data that are synchronously recorded during the laryngeal endoscopy. We show that the image segmentation algorithm of the glottal area can be optimized by matching the Fourier spectra of the pre-processed video and the spectra of the acoustic recording during the phonation of sustained vowel /i:/. We verify our method on a set of LHSV recordings taken from subjects with normophonic voice and patients with voice disorders due to glottal insufficiency. We show that the computed geometric indices of the glottal area make it possible to discriminate between normal and pathologic voices. The median of the Open Quotient and Minimal Relative Glottal Area values for healthy subjects were 0.69 and 0.06, respectively, while for dysphonic subjects were 1 and 0.35, respectively. We also validate these results using independent phoniatrician experts.

**Keywords:** vocal disorders; laryngeal high-speed video; image segmentation; acoustic recordings of voice; signal processing; multimodal sensing



**Citation:** Kopczynski, B.; Niebudek-Bogusz, E.; Pietruszewska, W.; Strumillo, P. Segmentation of Glottal Images from High-Speed Videoendoscopy Optimized by Synchronous Acoustic Recordings. *Sensors* **2022**, *22*, 1751. https://doi.org/10.3390/s22051751

Academic Editors: Enrico G. Caiani and Sheryl Berlin Brahnam

Received: 30 December 2021
Accepted: 15 February 2022
Published: 23 February 2022



## 1. Introduction

Regular assessment of the health of the human voice is important for the accurate detection of voice disorders with varied etiology. Exposure to the risk factors of voice disorders is increasing in the contemporary world. It is estimated that about a third of workers in industrialized societies use voice as their main work tool. UK figures report that over five million workers are routinely affected by voice impairment, at an annual cost of around £200 million [1]. In recent decades, constant advancements in technology and virtualization of life have rendered voice crucial for communication, particularly in the case of individuals for whom it is a primary tool of trade and who are exposed to excessive vocal loading, e.g., actors, singers, coaches, teachers, call-center workers, etc. Professional voice users report to otolaryngological and phoniatric outpatient clinics with common problems. Due to voice overload, the vocal folds may be affected and deformed by pathological abnormalities causing malfunction of the entire speech apparatus [2]. Incorrect phonation caused by excessive muscular activity may lead to loss of voice. The most common effect of abnormal phonation (hyper-phonation) is pathological changes that appear in the form of nodules, polyps, and the weakening of the arytenoid or thyroarytenoid muscles [3].

Precise assessment of voice disorders with the aid of modern technology enables a structural and functional assessment of the larynx. Vibrations of the vocal folds play an essential role in voice production [4]. During the periodic oscillation of the vocal folds, the area between the vocal folds, called the glottal area, changes, which results in periodic

interruption of the expiratory airflow through the glottis. Oscillation disturbance affects voice quality. Therefore, an accurate assessment method of vocal fold vibrations is crucial for the early diagnosis and treatment of various pathologies of the larynx [5]. Innovative instrumental methods are steadily gaining importance in otolaryngological and phoniatric studies of voice disorders [6].

Currently, the diagnosis of voice disorders can be facilitated by computer-based processing methods that enable the computation of many diagnostically meaningful parameters [7]. The most common diagnostic methods rely on acoustic voice measurements during sustained production of vocal sounds, termed phonation [8–10]. The parameters characterizing voice quality can be computed from microphone recordings of the produced voice [10,11], subglottal neck-surface accelerometer-based force measurements [12–14], or with an electroglottograph—an apparatus measuring the amount of electricity that passes through the larynx [15].

Recently, it has been stressed that direct visualization of laryngeal glottal structures and phonatory function in the clinical setting is essential to assess larynx pathologies. Therefore, in the last decades, an increasing number of scientific studies have reported new developments in advanced methods of digital processing and analysis of images of vibrating vocal folds [2,16,17]. There are three basic techniques of image collection: laryngovideostroboscopy, videokymography, and laryngeal high-speed videoendoscopy (LHSV).

Videokymography is a high-speed imaging method depicting one horizontal line transverse to the glottis. The successively collected line of pixels stacked (from top to bottom) into a matrix is presented in real-time on a standard monitor revealing a graphical representation of the spatial position of the vocal folds over time [18,19].

The most common visualization method used in clinical practice is laryngovideostroboscopy (LVS) [11], although it does have significant limitations because visualizing a single vibration cycle of the vocal folds requires recordings taken from tens or hundreds of images from a sequence of consecutive cycles. Thus, vibration disorders of intermittent nature cannot be adequately detected, which has a detrimental effect on the quantitative analysis of LVS images.

LHSV is the tool that provides the most precise insight into the function of the larynx during sound production. High-speed digital imaging of the oscillating vocal folds enables visualization of the true frame-by-frame movement of the vocal folds during sound production. However, this imaging technique requires the application of an advanced and expensive system that allows thousands of images to be recorded per second.

Recently, extensive research has been carried out in the field of a quantitative assessment of the glottal cycle using laryngeal high-speed digital imaging [16,20]. However, the main research problem encountered in these studies has been the development of automatic image analysis methods for segmenting the images of the larynx so that the boundaries of the vocal folds and consequently the glottal area could be reliably detected in each LHSV frame. If the segmentation process is inaccurate, the time- and size-related parameters characterizing the kinematics of the vibrating vocal folds will have little clinical relevance for the otolaryngologist and phoniatrician [21,22]. A detailed list of these parameters is provided in our previous study [11]. For a more in-depth review of related work, see Section 2.

In this work, we propose:

- an original method for automatic segmentation of the laryngoscopic images registered with LHSV based on the fusing of time-synchronized data modalities coming from the acoustic measurements of the produced voice and LHSV recordings.
- incorporation of the spectral domain data of the acoustic signal to control and optimize the segmentation algorithm of the LHSV images of the larynx during phonation.

We propose a novel approach to segmenting images of the moving objects to find other, non-medical applications. We demonstrate that it is effective in segmenting images of the vibrating vocal folds, and phoniatricians positively evaluated our results. The main

advantage of our method is that it allows automatic segmentation of LHSV images without the need for trial and error in the search for optimal segmentation parameters.

The paper is structured as follows. In Section 2, we review recent studies related to the analysis of LHSV images. Section 3 describes the apparatus used to record LHSVs and introduces the medical cases of voice pathologies considered in this study. Section 4 explains the proposed method of automatic segmentation of LHSV images and presents the results that verify it in Section 5. The potential of the presented approach is appraised in the discussion in Section 6. Finally, overall conclusions from the presented study are formulated in Section 7.

## 2. Related Works

In recent years, a majority of the image analysis techniques of the vocal folds phonation process have relied on LHSV recordings [23–25] because the information from images of the oscillating vocal folds recorded at a rate of approximately 4000 frames per second (fps) provides a greater tracking precision of the movements of the vocal folds [26]. In particular, thanks to a very short image acquisition process, the problem of the movement of the laryngoscope with respect to the larynx, which tends to complicate algorithms for the analysis of the laryngostroboscopic images, is minimized. That is a great technological advancement compared to laryngovideostroboscopic techniques, which involve long image acquisition time, e.g., up to 20 s, and reconstruction of a single vibration cycle of the vocal folds from many consecutive vibration periods. Moreover, laryngovideostroboscopic techniques enable the visualization of real-time kinematics of transient vibration disruptions that tend to accompany some important voice disorders [27].

However, whichever image acquisition technique is applied, quantitative analysis of laryngovideoscopic images requires complex image analysis methods and additional adjustment, e.g., the settings of the parameters used in the segmentation algorithms need to be established to achieve reliable results [28]. In particular, the first step in the quantification of vocal vibration kinematics is a segmentation of the glottal area, i.e., the region between the vocal folds, in each consecutive image of an LHSV sequence. Should that segmentation step be flawed, any further parameters characterizing geometric and time-related parameters of the VF movement will be inaccurate.

The development of a reliable image segmentation method of the larynx is a major challenge for automated computer algorithms for the following reasons [29]:

- a single view two-dimensional projection of a three-dimensional anatomic structure of the glottis is recorded; in particular, not all details of the elastic deformations of the vocal folds during the vibration cycle can be viewed,
- the point light source coming from the laryngoscope illuminates different anatomical regions of the glottis with nonuniform intensity,
- the position of the laryngoscope with respect to the glottis is different in each laryngeal examination, e.g., the distance and the viewing angle of the camera,
- the ground truth information about the glottal area can be collected only from subjective inspections and manual delineations of an expert doctor or a phoniatrician; for a very large number of images, the task is effort and time-consuming, and what is more, a special user interface needs to be developed to enable clinicians to precisely delineate the glottal area.

One of the most advanced approaches to the analysis of LHSV images was presented in [30]. The authors employed the Kalman filter to estimate the kinematic sequence of each of the vocal folds' edges to predict the contact force during their collision. Researchers in [31], on the other hand, defined the region of interest (ROI) containing the image of the glottal area by analyzing the average intensity variations both in the columns and in the rows of the images.

In [32], a novel method was proposed for automatic glottis segmentation in endoscopic high-speed videos. ROI detection was done using the Fourier descriptors and a thresholding method combined with a level set algorithm, incorporating the prior glottis shape. The

level set method is a numerical technique closely related to the active contour framework used to trace shapes of selected figures and identify dependencies among them based on the energy minimization criterion. Another advanced segmentation method utilizing the level set based curve evolution for vocal folds segmentation was proposed in [33]. However, the authors noted that the method required subjective parameter tuning and was unsuitable for fully automated analysis of the vocal fold movements during phonation.

It should be noted that most of the developed methods that have been proposed for segmentation of glottal images are designed for specific image recording conditions and work properly only for local databases of videos collected at institutions, hospitals, or health centers, and thus require manual validation (especially in the case of new registrations). Designing an algorithm that would yield satisfactory results for any given laryngoscopic video registration of the vocal folds during voice production (phonation) is a complex task.

There is a group of image segmentation approaches that use continuity conditions derived from image sequences and extend the analysis into the time domain, e.g., by adapting a Geodesic Active Contour model defined in three dimensions and formulating continuous and variational energy minimization problems. The 3D surface is automatically evaluated through an algorithm optimizing the forces derived from the image and the surface shape (curvature and continuity), which minimizes the hypothetical energy functional [34]. Other methods adopt a Canny edge detector preceded by a 3D mean curvature filtering process [26].

Several methods have been proposed to validate the vocal folds image segmentation results with ground truth, based on calculating a metric of similarity between human- and machine-generated results [35]. The main problem related to objective evaluation is the necessity of generating ground truth, which is subjective and requires considerable time and effort [35].

A recent paper [36] tested different configurations of deep convolutional long-short-term memory networks were tested for automatic segmentation of the glottis and vocal folds in endoscopic LHSVs. The best-performing network was selected for extensive testing on a large set of LHSVs. Interestingly, the long-short-term memory architecture allowed the modeling of the spatial and temporal features of the vibrating vocal folds. This machine learning approach enabled fully automated quantification of the vibrations of the vocal folds. Nevertheless, the network required 13,000 LHSV frames to train the network. High segmentation precision was achieved, resulting in Dice coefficients values used for quantifying the segmentation results exceeding 0.85.

Other very recent work concerns the determination of the effect of incorporating features derived from vocal fold vibration transitions into acoustic boundary detection [37], comparative analysis of rapid videolaryngoscopy images and sound data [38], and a computer model for the study of unilateral vocal fold paralysis [39]. Interestingly, a method for detecting COVID-19 by analyzing vocal fold vibrations has also been proposed [40]. The presented literature review concludes that most image segmentation algorithms require a particular validation procedure to prove their accuracy. However, we have noted that a properly selected image segmentation technique combined with correlated acoustic analysis allows us to objectify the delineated contour of the vocal folds and provide a compliance parameter, which is crucial for quantitative image-based segmentation results of the glottal area. In our seminal work, we proposed such an image segmentation method [41]. The method is based on comparing the segmentation result with the synchronously collected acoustic registration during the patient's phonation of the vowel /i:/. This paper expands on the first study and validates the results on a set of LHSV recordings for normophonic and dysphonic voices. Previous work was tested on only a few cases and did not include a broader discussion of the results and clinical interpretation by phoniatricians. In this study, we also fully automate this method by automatically detecting the glottal folds region of interest (RoI).

## 3. Materials and Recordings of LSHV

The laryngeal recordings were carried out at the Department of Otolaryngology, Head and Neck Oncology, Medical University of Lodz.

Twenty-two subjects participated in the study, 7 males and 15 females (see Table 1). Eleven of the participants had normophonic voice (denoted N1–N11), whereas 11 were patients (denoted D1–D11) with voice disorders (dysphonia) caused by glottal insufficiency due to incomplete glottal closure. The age of the normophonic group (7 females and 4 males) ranged from 27 to 65 years, with a mean age of 46. The dysphonic group consisted of 8 females aged 26–64 (with the mean of 47 years) and 3 males aged 57–71 years (the mean = 65). Among the dysphonic patients, only the two oldest men experienced no professional vocal loading.

**Table 1.** Summary of patients participating in the study by gender and normophonic/dysphonic subjects.

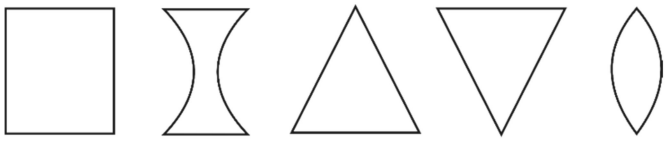
| | Patients | Normophonic | Dysphonic |
|---|---|---|---|
| Males | 7 | 4 | 3 |
| Females | 15 | 7 | 8 |
| Total | 22 | 11 | 11 |

For all subjects, after a routine Ear, Nose, and Throat (ENT) examination, the imaging recordings of the larynx were performed using an LHSV system. In the normophonic patients, the LHSV examination showed no significant deviations in the regularity and symmetry of vocal folds vibrations, mucosal wave, and glottal closure (Figure 1).

(**a**) (**b**)

**Figure 1.** Images of the glottis for the normophonic subject N3 for the maximum opening (**a**) and maximum closing (**b**) of the vocal folds correspondingly.

However, in three of the examined normophonic women, the imaging of the larynx revealed slightly incomplete glottal closure only in the 1/3 posterior part of the glottis, which did not affect their voice quality. In the dysphonic patients, disturbances of vocal fold vibrations and incomplete closure of the glottis during phonation were observed. The Glottal Closure Types (GTs) were described according to the guidelines of the Committee on Phoniatrics of the European Laryngological Society (ELS) [5], in the following way: type A is rectangle/longitudinal glottal closure, B—hourglass, C—triangle, D—V-shaped, and E—spindle-shaped. An illustration of these types of glottal closure is shown in Figure 2.



(**A**) (**B**) (**C**) (**D**) (**E**)

**Figure 2.** Classification of glottal closure types: (**A**) rectangle/longitudinal, (**B**) hourglass, (**C**) triangle, (**D**) V-shaped, (**E**) spindle-shaped.

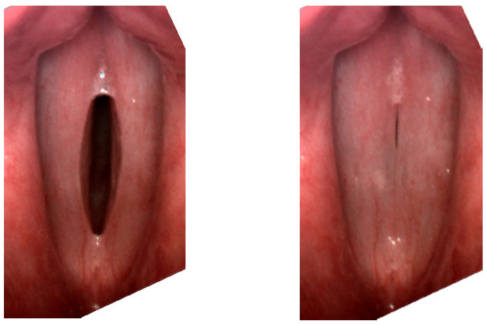
In the dysphonic subjects, the most commonly occurring abnormality was the spindle-shaped glottal closure. In 8 of the dysphonic patients, the spindle-shaped glottal gap along the entire membranaceous part of the glottis during the closed phase of the glottal cycle was observed (see, e.g., Figure 3). They complained of several voice-related problems: permanent hoarseness, vocal fatigue, and periodic voicelessness/aphonia. In one of the patients who reported periodic vocal fatigue, the longitudinal glottal closure (little incompleteness of glottal closure in the 1/2 posterior part of the glottis during the closed phase of the glottal cycle) was observed (Figure 4). One of the dysphonic subjects presented only a minimal spindle-shaped glottal gap in 1/3 middle part of the glottis (Figure 5).

(**a**) (**b**)

**Figure 3.** Images of the glottis for dysphonic patient D8 with severe glottal insufficiency for the maximum opening (**a**) and maximum closing (**b**) of the vocal folds correspondingly.

(**a**) (**b**)

**Figure 4.** Images of the glottis for dysphonic patient D5 with longitudinal glottal insufficiency for the maximum opening (**a**) and maximum closing (**b**) of the vocal folds correspondingly.



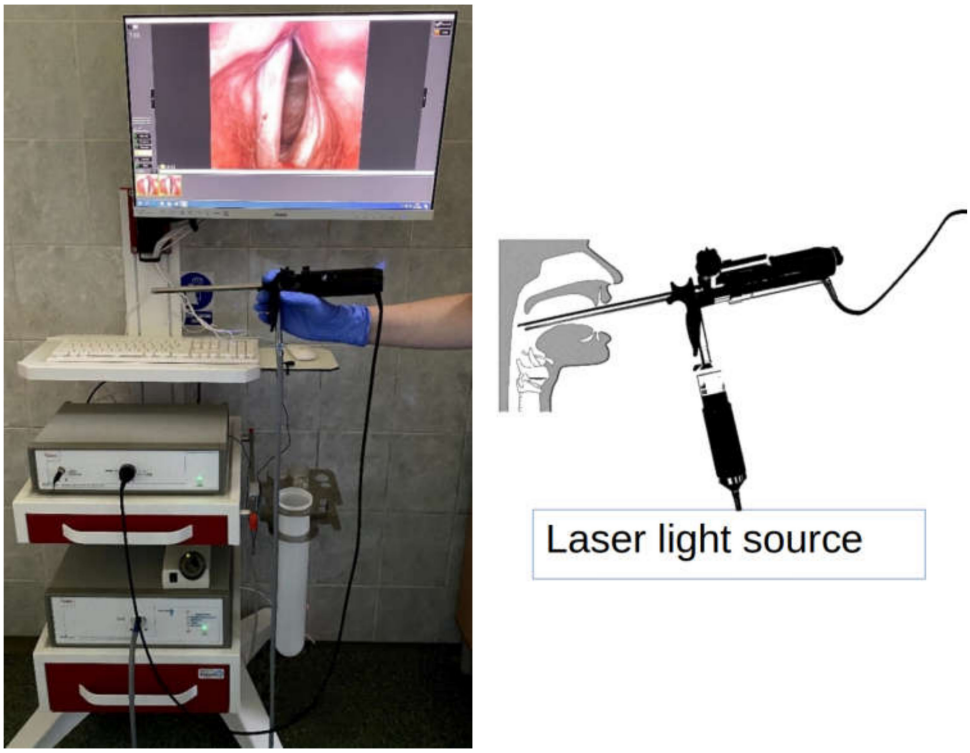
(**a**) (**b**)

**Figure 5.** Images of the glottis for dysphonic patient D4 with minimal spindle-shaped glottal insufficiency for the maximum opening (**a**) and maximum closing (**b**) of the vocal folds correspondingly.

The vocal fold function was assessed during sustained phonation of vowel /i:/ at a pitch and loudness comfortable for the subject. Simultaneously to the LHSV imaging, a synchronized acoustic recording of the voice produced during phonation was done. The recordings were repeated several times for each of the examined subjects.

The laryngeal images were recorded with a high-speed laryngeal camera from Diagnova Technologies with a 2/3-inch progressive CCD sensor with a camera shutter connected

to an external microphone that synchronously recorded the acoustic wave generated by the vocal folds during the phonation. The image capture rate of the camera was 3150 images per second. The images were digitized at a resolution of 480 × 400 pixels. The inherent geometric lens distortions were corrected with calibration methods based on pixel coordinate remapping [35]. The light source was a 15 W laser with special spectral characteristics to achieve excellent visualization of the glottal tissue. The light from the illuminator was transferred to the endoscopic optics via an optical fiber. The camera was equipped with an electronically controlled lens allowing for manual or automatic image focusing. The Fiegert-Endotech ø12.4/7.2 endoscope used in the laryngeal recordings together with the assembled complete laryngeal high-speed system is shown in Figure 6a,b shows a diagram of how the laryngoscope is positioned in the larynx during the examination. Simultaneously to the LHSV recording, a synchronized acoustic recording of the voice produced during the sustained phonation of vowel /i:/ was done. The microphone we used for the voice recordings was an electret microphone MK602762PC featuring 20 Hz–16,000 Hz bandwidth. The relative distance between the microphone and the subject during the recordings was approximately 30 cm. The acoustic wave signal was sampled at a sampling rate of 22,050 Hz [36].




(**a**) (**b**)

**Figure 6.** Photograph of the LHSV recoding system with the 70-degree rigid scope, attached light source, and a microphone. The box on the lowest shelf of the rack is the endoscope's light source, and the box on the middle shelf is a high-speed camera offering acquisition of up to 4000 images per second (**a**), a diagram showing the position of the laryngoscope during laryngeal examination (**b**).

For the recording rate of the high-speed laryngeal images at 3150 frames per second, the camera captured approximately 200 images during one oscillation cycle of the VFs. This image frame rate was approximately 7 times slower than the sampling rate (22,050 Hz) of the acoustic signals, i.e., 7 acoustic audio samples were recorded during the acquisition of one LHSV image. In further analysis, the acoustic signal was down-sampled (as further explained in Section 4) to properly match its sampling rate to the image acquisition rate.

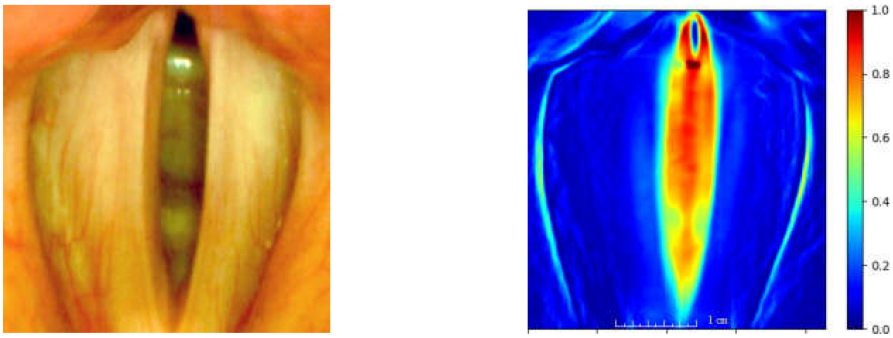
## 4. Automatic Segmentation Method of LHSV Images

The recorded LHSV RGB image sequences were converted to grayscale images using the standard formula [42]: Grayscale = 0.299R + 0.587G + 0.114B, where R, G, B are the red, green, and blue color components, respectively. Then, each image frame from the LHSV sequence was rotated so that the main axis along the glottal area was positioned vertically.

An important pre-processing step of the analysis of laryngeal images is to identify the region of interest (ROI), i.e., the region containing the vocal folds. We applied an efficient way of locating the ROI based on calculating the total image variation quantity as proposed in [31]. This quantity is obtained by calculating the sum of the absolute differences of image brightness in successive image frames. Thus, rapid changes in image brightness (e.g., due to moving vocal folds) will yield large values of this quantity. We define this quantity as the total variation map $TV(x, y)$ calculated over a sequence of frames as follows:

$$TV(x,y) = \sum_{t=0}^{N-1} |I(x,y,t+1) - I(x,y,t)| \quad (1)$$

where: $I(x, y, t)$—is the intensity function of the image at spatial coordinates $x$, $y$, and $t$ denotes the frame index $t = [0, 1, \ldots, N-1]$ of $N$ analyzed images from the LHSV sequence. Points of $TV(x,y)$ map assume large values for those image locations where there is a large variability of image brightness for consecutive images of an LHSV sequence. The map serves to locate the ROI for further image analysis. Figure 7 presents the obtained heat map based on the established ROI.




(**a**) (**b**)

**Figure 7.** The image of the glottis of a normophonic subject (**a**) and the corresponding total variation image (**b**), as defined in Equation (1), represented as a heat map (the larger the variation, the warmer the color of the map).

The most important element of the quantitative assessment of the phonatory process is the automated localization of the vocal fold edges during voice production. When VF edges are correctly detected, a glottovibrogram, Glottal Area Waveform (GAW), Glottal Gap Waveform (GGW) can be constructed. These representations provide a complete characterization of the kinematics of the vocal folds boundaries. From these representations, numerous geometric parameters of the glottal area shape and its variation over time can be calculated. The definitions of the parameters used in this study can be found in Appendix A. Additionally, a more complete set of parameters characterizing glottal area geometry used to quantify other laryngeal pathologies is defined in our previous work [11].

The image processing pipeline is shown in Figure 8 and is as follows.

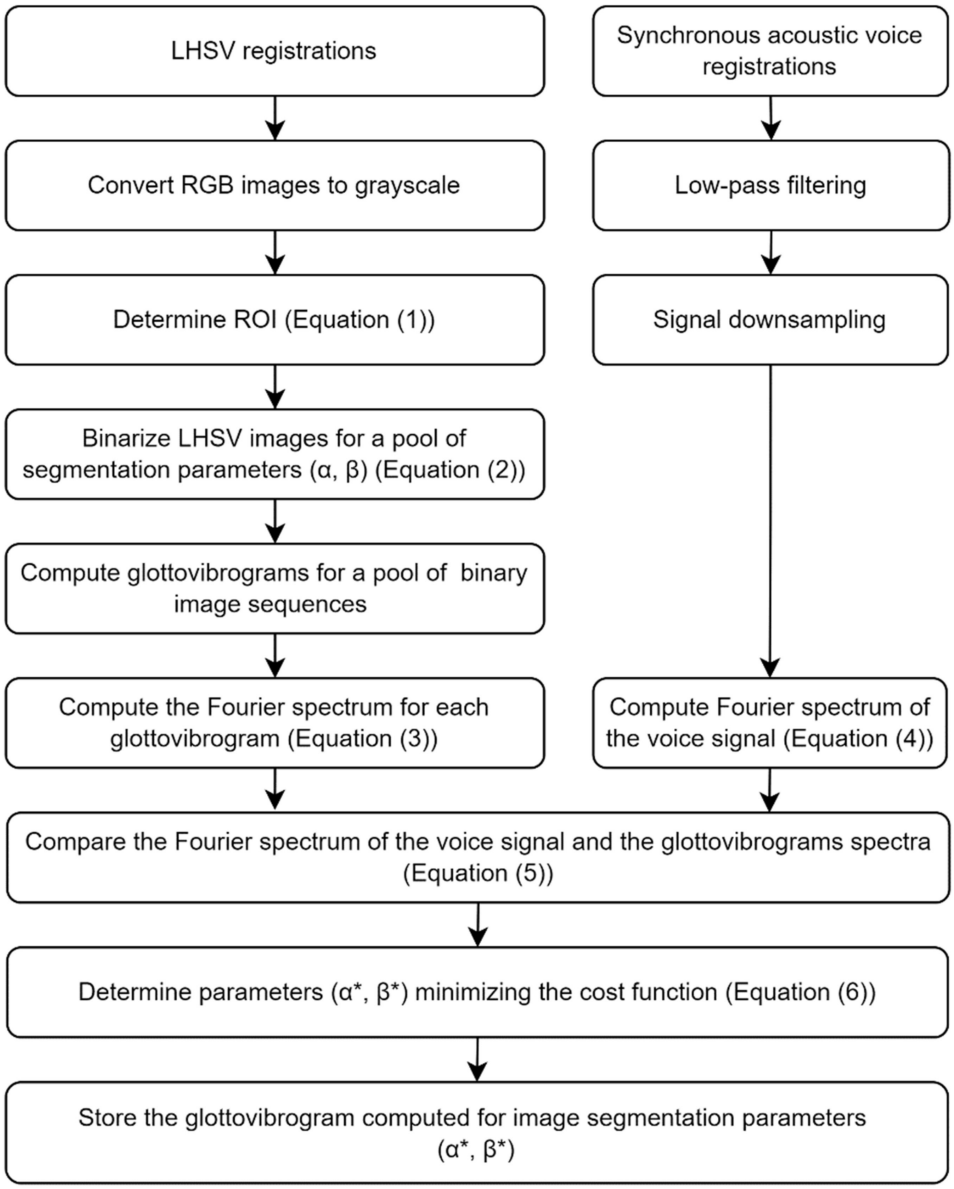



**Figure 8.** The processing pipeline of recorded LHSV images and synchronously recorded voice signal during sustained phonation of vowel /i:/.

The color image of the glottis (Figure 9a) captured by the high-speed camera is converted into a greyscale image, then the region of interest is selected, and the edges of the vocal folds are detected (Figure 9b). Based on this, the glottal area, i.e., the space between the vocal folds, is determined (Figure 9c). The GAW is the signal representing instantaneous variations of the glottal area in time (Figure 9e). From the GAW, one can calculate geometric and time-related parameters characterizing the oscillation process of the VF, e.g., the minimum and maximum values of the glottal area [11].

The GGW is the signal representing instantaneous variations in the width of the gap between the VFs computed at predefined levels of the glottis. From the GGW, one can calculate the closing and opening periods of the VFs during a vibration cycle [11].

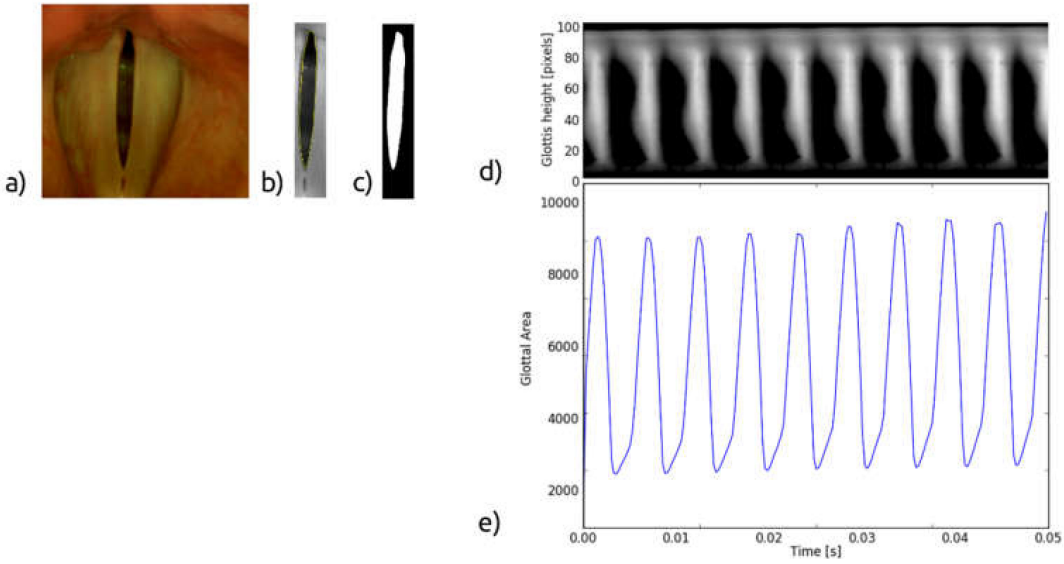



**Figure 9.** Representations of the LHSV image: (**a**) laryngeal image of the glottis, (**b**) detected contour of the glottal boundary, (**c**) glottal area, (**d**) the glottovibrogram, (**e**) the glottal area waveform.

The glottovibrogram, on the other hand, is a spatio-temporal map illustrating time variations of the width of the glottal gap at different levels of the glottis. The glottovibrogram shown in Figure 9d depicts a map in which the columns represent time and rows correspond to the glottal width along the anterior-posterior length of the glottis. The instantaneous glottal gap width is represented by pixel brightness in the glottovibrogram map.

In this work, we propose an automatic method for detecting VF edges based on the combined analysis of the data derived from LHSV recordings and synchronously recorded acoustic signals. We show that the segmentation algorithm can be optimized by the spectral data of the acoustic signal without the need to refer to ground truth information.

The underpinning idea of the method shown in Figure 10 is to pool candidate segmentations of the glottal images for a large set of segmentation parameters. Then, select the best segmentation result by finding the best match between the pool of Fourier amplitude spectra computed from the glottovibrograms and the Fourier amplitude spectrum computed for the synchronously recorded acoustic signal.

The applied segmentation method of the glottal image is based on a simple image thresholding method, as follows:

$$I_O(x,y) = \begin{cases} 0 \; if med_\alpha(x,y) - I(x,y) < \beta \\ 1 \; if med_\alpha(x,y) - I(x,y) \geq \beta \end{cases} \quad (2)$$

where:

$x$, $y$—pixel coordinates of the monochrome image,

$I$—image recorded during the phonation process,

$I_O$—binary image containing the thresholding result,

$med_\alpha(x,y)$—median value computed at image coordinates $x$, $y$ for pixels in a block size $\alpha \times \alpha$,

$\alpha$—the first segmentation parameter, i.e., the block size of the median filter,

$\beta$—the second segmentation parameter.

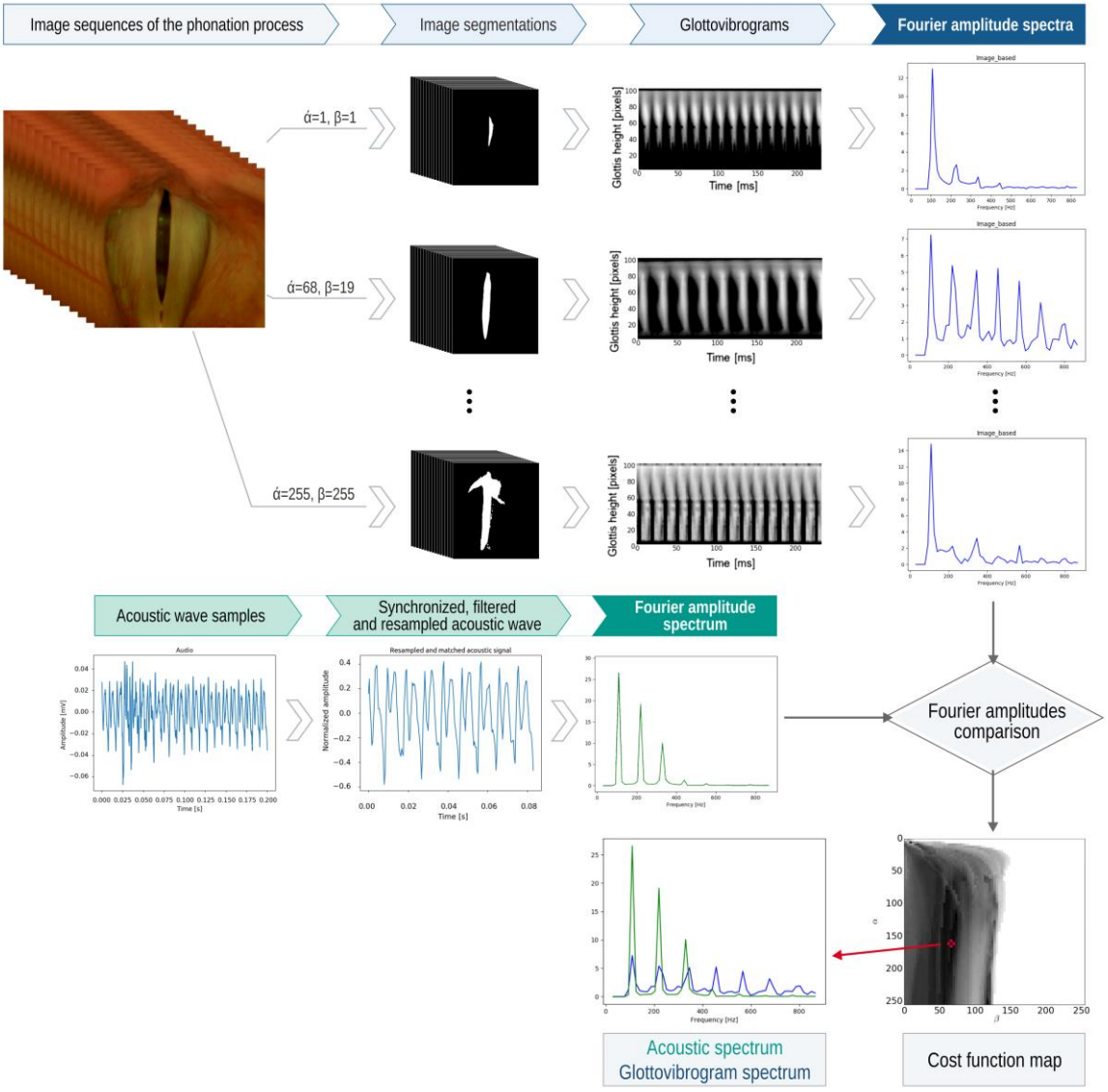



**Figure 10.** Diagram explaining the designed method of the segmentation of glottal images where, in the search for the best segmentation results, the Fourier spectra derived from the pool of segmented LHSVs are compared to the Fourier spectra of the acoustic recording.

Parameter $\alpha$ specifies the block size of the median two-dimensional filter, and parameter $\beta$ acts as a threshold value for the subtraction result between the image pixel $I(x,y)$ and median filtered pixel $med(x,y)$ at coordinates $x$, $y$. Note that the parameter $\alpha$ determines the strength of the median filter $med_{\alpha}(x,y)$, i.e., the larger the filter window size (larger $\alpha$), the stronger the smoothing effect the filter will have. Then, according to Equation (2), from this filtering result, the image content $I(x,y)$ is subtracted. We can interpret this operation as removing the constant component from the image, computed for the image window size defined by parameter $\alpha$. The remaining image data, i.e., devoid of the constant component, is thresholded at a level determined by the parameter $\beta$. The outcome of this segmentation method is a binary image consisting of pixels that are assigned values 0 (corresponding to the minimum pixel brightness) and 1 (corresponding to the maximum pixel brightness).

This segmentation method is applied for a pool of segmentation parameters $\alpha$ and $\beta$. As a result, we obtain $M = \alpha \times \beta$ segmentation results in Figure 10. Parameters $\alpha$, $\beta$ assume integer values in the range of [1, 255]. The task is to select from the $M$ segmentation results the one that best fits the glottal area.

We propose the following multistep automatic procedure for selecting the best segmentation result for an LHSV recording consisting of $N$ images of the glottis (refer to a graphical illustration of this method in Figure 10):

1. Compute a pool of $M$ sequences of binary images; each sequence consists of $N$ binary images obtained by segmenting LHSV images by applying the selected parameter combination ($\alpha$, $\beta$) of the segmentation procedure as defined in Equation (1).
2. For each of $M$ sequences for the selected parameters ($\alpha$, $\beta$), compute the glottovibrogram $g_{\alpha,\beta}(t,l)$, where $t$—is the discrete time coordinate (the horizontal axis) and $l$—is the level along the glottal length (the vertical axis).
3. For each of $M$ glottovibrograms, compute the Fourier spectrum along $L$ rows of the glottovibrogram and sum the results:

$$F_{\alpha,\beta}(f) = \frac{1}{N}\sum_{l=0}^{L-1}\sum_{t=0}^{N-1} g_{\alpha,\beta}(t,l)e^{-j\frac{2\pi t}{N}f} \quad (3)$$

where:
$g_{\alpha,\beta}(t,l)$—the point of the glottovibrogram map computed for a parameter set ($\alpha$, $\beta$),
$F_{\alpha,\beta}(f)$—Fourier coefficients of the glottovibrogram,
$f$—frequency,
$N$—the number of analyzed consecutive LHSV images,
$L$—the number of levels at which the glottal length is represented, i.e., the number of rows of the glottovibrogram.

4. Compute the Fourier spectrum of the acoustic recording $s(t)$ performed synchronously with the LHSV recording:

$$S(f) = \frac{1}{N}\sum_{t=0}^{N-1} s_d(t)e^{-j\frac{2\pi t}{N}f} \quad (4)$$

where:
$S(f)$—Fourier coefficients of the acoustic recording,
$s_d(t)$—downsampled (decimated) acoustic signal (as explained below),
$f$ frequency,
$N$ the number of acoustic samples (after down-sampling).

Note that the down-sampling of the acoustic signal is necessary before the glottovibrogram Fourier spectra and the acoustic spectra computed in Equations (3) and (4) can be compared. The acoustic signal is down-sampled by a factor of 7, i.e., from the sampling rate of 22,050 to the sampling rate of 3150, which is equal to the acquisition frame rate of the LHSV sequence. Before down-sampling, the acoustic signal is low-pass filtered using 4-th order Butterworth filter with a cut-off frequency $f_c$ = 1500 Hz to meet the sampling theorem condition that the maximum frequency components of the sampled signal cannot exceed half of the sampling rate.

5. For each combination of parameter values ($\alpha$, $\beta$) compute the cost function $d_{\alpha,\beta}$ to compare the modulus of the glottovibrogram spectra and the modulus of the acoustic spectra:

$$d_{\alpha,\beta} = \sum_{f=0}^{N/2-1} \left| \left|F_{\alpha,\beta}(f)\right| - \left|S(f)\right| \right| \quad (5)$$

where:
$|\cdot|$—denotes the modulus of the Fourier coefficients.

6. Find a parameter set ($\alpha$*, $\beta$*) for which the cost function is minimum:

$$argmind_{\alpha,\beta} = (\alpha^*, \beta^*) \quad (6)$$

7. Select the best segmentation result of the glottal image according to the criterion (6), obtained for parameters ($\alpha^*$, $\beta^*$).

The values of the cost function $d_{\alpha,\beta}$ computed for a set of parameters ($\alpha$, $\beta$) are shown in Figure 10 in the form of a grayscale image where the value of the cost function is represented by pixel brightness. An asterisk denotes the minimum of the cost function. The example of the best fit of the Fourier amplitude spectra is shown at the bottom of Figure 10.

We should note that the proposed method involves a high computational cost due to the optimization process in which the best set of segmentation parameters ($\alpha^*$, $\beta^*$) is selected. This optimization method requires the computation of $N \times N = 255 \times 255 = 65{,}025$ segmentations of each image frame from the LHSV recording. Then, for a series of segmented images, the corresponding glottovibrograms must be constructed. Their Fourier spectra have to be calculated. Then, these spectra have to be compared one by one with the spectrum of the recorded acoustic signal, and the best fit between them has to be selected. We estimate the proposed method's computation time to segment a single LHSV recording consisting of 256 images to be approximately 5 min for a PC equipped with an Intel i7 processor. However, mapping the proposed algorithm, which consists of multiple independent computational threads, to the Graphical Processing Units (GPU) would significantly mitigate this shortcoming of the algorithm.

The image processing and analysis algorithms and acoustic signal processing algorithms were developed in Python and C++ programming languages using open libraries, i.e., NumPy, SciPy, Matplotlib, and OpenCV. For time-critical methods (e.g., computing the median of a subset of pixel values), the C++ programming language was used to create Python bindings. We used the Spyder Integrated Development Environment for programming in Python.

## 5. Results

The proposed method was tested on the LHSV recordings collected from 11 individuals with normochromic voices and 11 with pathological voices, i.e., glottal insufficiency. During the LHSV recordings, the requirement was to record the voice signal simultaneously during phonation of vowel /i:/. Both the video and acoustic recordings were pre-processed according to the procedures described in Section 4 to make them suitable for computing the Fourier spectra, i.e., the pool of glottovibrograms was computed for a set of candidate segmentation parameters ($\alpha$, $\beta$) and the acoustic recordings were down-sampled to match the sampling rate of the signal ($fs$ = 22,050 Hz) to the frame rate of the LHSVs ($f_v$ = 3150 Hz).

In Figure 11, we show an example segmentation results for the normophonic subject N2 obtained for six random selections of parameter values ($\alpha$, $\beta$) and one segmentation obtained for a parameter set ($\alpha^*$, $\beta^*$), i.e., that minimizes the cost function defined in Equation (5).

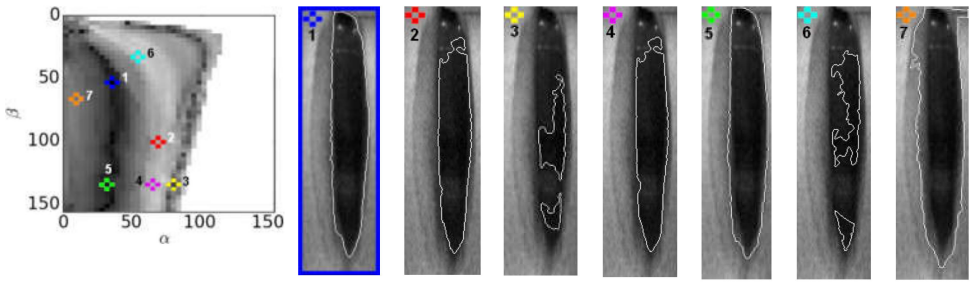



**Figure 11.** Plot of the cost function map $d_{\alpha,\beta,}$ (left panel) and example image segmentation results (right panel) obtained for the normophonic subject N2. The segmentation results are obtained for parameters ($\alpha$, $\beta$) and assigned different numbers in the cost function plot. The best segmentation result is shown in a thick box on the left side of the right panel and marked with the number 1.

### *5.1. Phoniatrician Validation of the Obtained Results*

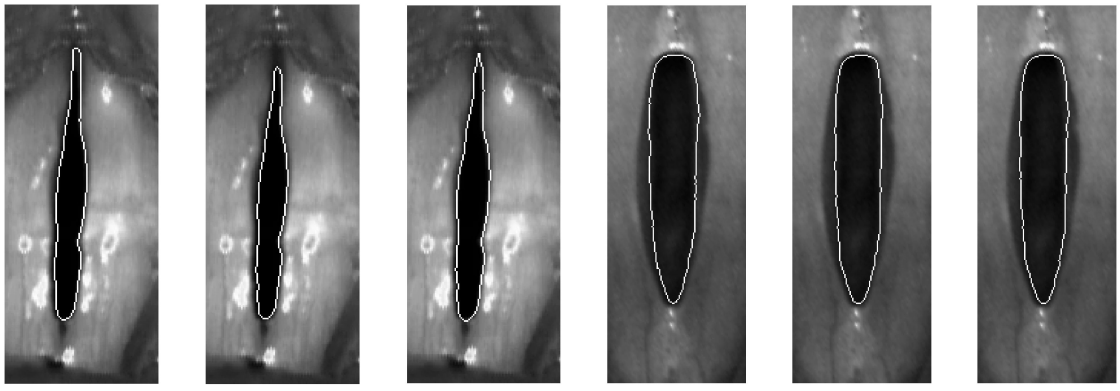
The complexity of the vocal fold anatomical structure makes it difficult to provide objective ground-truth annotations that would enable quantitative evaluation of the established vocal fold edge positions during phonation.

Our attempt to use a graphics tablet to delineate vocal fold boundaries on LHSV images was labor-intensive and not very precise. The drawn lines in many segments had to be corrected, and the result was not satisfactory in most cases. Thus, this method of obtaining ground truth from phoniatricians for annotation of vocal fold boundaries did not work.

Therefore, a different approach using the capabilities of the proposed image segmentation method was used, in which a large pool of candidate segmentation was computed. Our automatic segmentation method searched for the minimum of the cost function defined by Equation (5) to determine the optimal segmentation result. We asked phoniatricians to perform a similar task on a preselected set of image segmentation results, i.e., to select, according to their clinical experience, the segmentation results that best match the vocal fold boundaries. Then we compared our results with the indications of phoniatricians. Below is a more detailed explanation of our approach to validating the segmentation results.

For each of the 22 examined LHSV recordings for both groups of individuals (normophonic subjects and patients with glottal insufficiency), we prepared a set of 60 different segmentation results obtained for different parameter values ($\alpha$, $\beta$) where only one segmented image was obtained for the parameters ($\alpha^*$, $\beta^*$), i.e., the one that was selected as the best according to the minimum condition of the cost function Equation (5). Then, for the set of 60 segmented images computed for each of the LHSV recordings, we asked two independent expert phoniatricians to select the best three image segmentation results corresponding to the best detection of the location of the vocal fold edges. The set of 60 segmented images of the glottis selected for evaluation was obtained for 60 pairs of segmentation parameters ($\alpha$, $\beta$) selected from the cost function map (see sample map in the left panel of Figure 11). The coordinates of these parameters in the cost function form a matrix of $6 \times 10$ regularly spaced points in the rectangular neighborhood of the parameters ($\alpha^*$, $\beta^*$) corresponding to the minimum of the cost function $d_{\alpha,\beta}$.

Importantly, for each of the recordings, the phoniatricians' selection of the best three segmentation results included the segmentation obtained for the parameters ($\alpha^*$, $\beta^*$). Another notable observation is that the selection done by each of the phoniatricians differed very little, regardless of whether they were concerned for the normophonic subjects or patients with glottal insufficiency (see Figure 12, for an example of segmentation results selected by phoniatricians).



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

**Figure 12.** Example segmentations of the glottic images selected by the phoniatricians: images (**a**–**c**) are for the normophonic subject N10; (**d**–**f**) is for the patient I5 with glottal insufficiency; images (**a**) and (**d**) are the results obtained for the optimum segmentation parameter set ($\alpha^*$, $\beta^*$) minimizing cost function (5).

### *5.2. Calculation of Geometric and Time-Related Parameters for the Segmented LHSV Images*

The designed algorithms described in this work make it possible to determine the position of VFs edges in terms of function minimization tasks. The obtained and validated segmentation results make it possible to compute several indices that quantitatively characterize the kinematics of the vocal fold vibrations (the definition of the indices is given in Appendix A). The presented solution is the basis for an objectified and quantitative analysis. The values of the computed indices for the examined subjects are summarized in Table 2.

**Table 2.** Geometric and time-related parameters for the segmented LHSV images in normophonic subjects and dysphonic subjects with glottal insufficiency.

| | Patient Number | Closing Quotient | Open Quotient | Speed Quotient | MRGA [1] |
|---|---|---|---|---|---|
| Normophonic | N1 | 0.36 | 0.81 | 1.25 | 0.23 |
| | N2 | 0.36 | 0.61 | 0.72 | 0.05 |
| | N3 | 0.37 | 0.65 | 0.76 | 0.00 |
| | N4 | 0.48 | 0.81 | 0.69 | 0.09 |
| | N5 | 0.42 | 0.67 | 0.62 | 0.01 |
| | N6 | 0.49 | 0.78 | 0.57 | 0.22 |
| | N7 | 0.40 | 0.76 | 0.90 | 0.01 |
| | N8 | 0.13 | 0.26 | 1.08 | 0.00 |
| | N9 | 0.45 | 0.69 | 0.53 | 0.17 |
| | N10 | 0.38 | 0.66 | 0.74 | 0.06 |
| | N11 | 0.49 | 0.77 | 0.59 | 0.20 |
| Dysphonic | D1 | 0.49 | 1 | 1.04 | 0.33 |
| | D2 | 0.53 | 1 | 0.89 | 0.60 |
| | D3 | 0.54 | 1 | 0.85 | 0.52 |
| | D4 | 0.49 | 1 | 1.04 | 0.52 |
| | D5 | 0.55 | 0.92 | 0.67 | 0.01 |
| | D6 | 0.47 | 0.92 | 0.96 | 0.03 |
| | D7 | 0.47 | 0.95 | 1.02 | 0.30 |
| | D8 | 0.48 | 1 | 1.08 | 0.69 |
| | D9 | 0.49 | 1 | 1.04 | 0.40 |
| | D10 | 0.52 | 1 | 0.92 | 0.35 |
| | D11 | 0.53 | 0.93 | 0.77 | 0.05 |
| *p*-values | | $1.3 \times 10^{-3}$ | $7.1 \times 10^{-5}$ | 0.04 | 0.01 |

[1] Minimal Relative Glottal Area.

The study confirms that it is possible to calculate quantitative parameters describing vocal fold vibratory characteristics based on the computer segmentation of LHSVs. The quotients Closing Quotient (CQ), Open Quotient (OQ), and Speed Quotient (SQ) were computed based on the obtained glottovibrograms, and their values depended directly on the accuracy of image segmentation. Table 2 presents the values of the CQ, OQ, SQ calculated for the middle part of the glottis. The rationale for taking that approach is that most of the dysphonic patients (subjects 9/11) presented with the largest incomplete glottal closure in the middle segment of the glottis, classified as type E: spindle-shaped GTs according to ELS. Thus, the segmentation results are essential for accurate quantification of the VF oscillations. Moreover, according to [42], vibrations in the medium segment of the glottis play a major role in normal voices. The pathology assessment at this point is the

most important in glottal insufficiency (complete lack of closure at this position). Thus, OQ calculated in the medium segment of the glottis is a meaningful parameter for this type of voice.

Please also see the box-and-whisker plot in Figure 13, showing the spread of the calculated quotient values. The boxes are drawn from first quartile Q1 to third quartile Q3 with a horizontal line within the box to denote the median. Out of the calculated quotients, the CQ and OQ assumed significantly different values in the normophonic and the dysphonic group (Figure 13). Moreover, the median values of OQ and MRGA quotients for healthy subjects were 0.69 and 0.06, respectively, while for dysphonic subjects were 1 and 0.35, respectively. Nevertheless, the differences for all calculated quotients for normophonic and dysphonic subjects were significant. See the bottom row in Table 2 with *p*-values calculated by applying the non-parametric Mann–Whitney U test to the calculated quotients for normophonic and dysphonic subjects. Note that all *p*-values are less than 0.05.

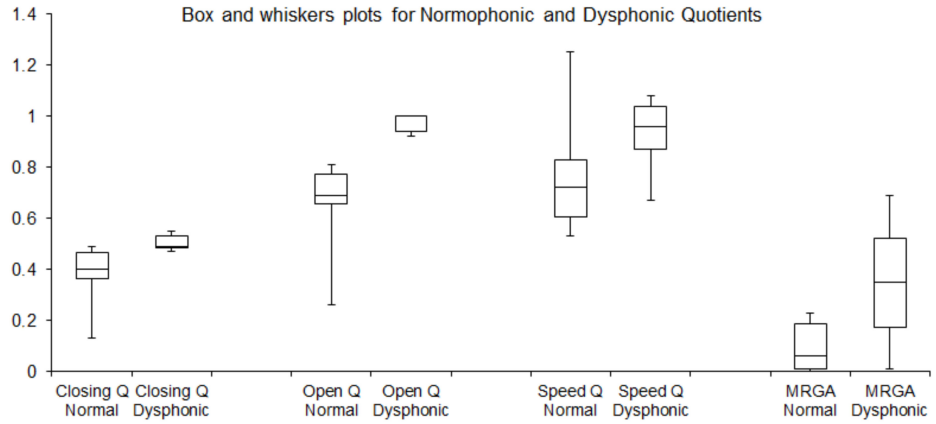



**Figure 13.** Box-and-whisker plots of calculated quotients for normophonic and dysphonic subjects. The upper and lower boundaries of the boxes indicate first quartile Q1 to third quartile Q3, respectively, while the boundary of the lower whisker denotes the minim value in the data set and the upper whisker boundary denotes the maximum value in the data set.

See Figure 14 illustrating clear discrimination of the two examined groups of subjects for MRGA, CQ, and OQ quotients. The collected data can be further used to build a larger database, e.g., from multiple phoniatric clinics, and apply machine learning algorithms [43] to discriminate normophonic and dysphonic patients robustly based on the calculated quotients.

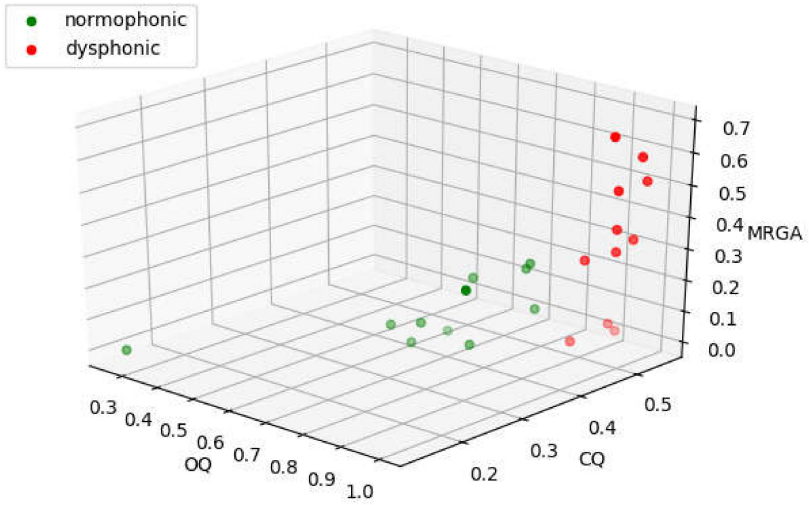



**Figure 14.** 3D plot for indices MRGA, OQ, CQ illustrating good discrimination of the normophonic subjects (green dots) and patients with glottal insufficiency (red dots).

The OQ assumed the highest values for the patients with the spindle-shaped glottal gap along the entire membranaceous part of the glottis. The OQ in those patients reached the value of 1.00, confirming that their vocal folds remained open throughout the phonation cycle. Similarly, for those subjects, the MRGA characterizing the ratio between minimum and maximum glottal area in the glottal cycle assumed large values (median 0.35), which confirmed a lack of the glottal closure. In the normophonic subjects, the MRGA reached small values (median 0.06), reflecting complete vocal fold closure along the entire length of the glottis.

## 6. Discussion

Computer image analysis techniques have brought about major advances in medical diagnosis based on quantified analysis of biomedical images of different modalities. In this respect, the analysis of biomedical images of moving tissue is particularly challenging. The human vocal folds vibrate with a frequency exceeding 200 Hz in the case of women. Real-time monitoring of this complex physiological phenomenon makes great demands on image recording systems. Recently, researchers' interest in laryngeal high-speed recordings has gained on the previously popular stroboscopic recordings as they required longer recording times and suffered from the inability to reproduce irregular phonatory functions of the vocal folds. Thanks to thousands of images recorded per second, high-speed cameras offer real-time insight into the movement of the oscillating vocal folds. Determination of vocal fold vibrations during the phonatory function of the larynx is a crucial element in the diagnosis of the clinical type of voice disorders. The LHSV technique is an innovative diagnostic tool used to visualize larynx kinematics. It offers an unprecedented quality of real-time visualization of VF phonatory movement [44].

Nevertheless, the task of image segmentation, whose goal is to detect and track the edges of the vocal folds, remains a difficult computational problem. Many approaches have been recently proposed for solving this problem, with those involving deep neural networks trained on laryngeal images of healthy subjects and patients with voice disorders showing the greatest promise. Although very successful such approaches require tens of thousands of training examples of laryngeal images to achieve image segmentation precision comparable to manual segmentations [38]. However, as the authors of this paper conclude, comparing these results with other approaches is not possible due to the lack of a suitable reference data set. One of the drawbacks of machine learning methods is that the results they produce lack explanatory power, and the segmentation decisions are hidden within the deep structure of the neural network. Nevertheless, advances in machine learning techniques towards explaining neural network decisions are ongoing and can certainly offer powerful tools for image recognition with explanatory features.

We have proposed a novel approach in which we control the image segmentation algorithm with data derived from acoustic recordings collected synchronously with the video capture of the phonatory process. It needs to be noted that the recorded acoustic signal is filtered by the vocal tract [45–48] and does not directly reflect the mechanical oscillation of the vocal folds. However, owing to the Fourier spectral representation, the fundamental frequency of the vibrations can be clearly outlined. During the phonation of vowel /i:/, the acoustic signal after computing its amplitude Fourier spectrum consists of the fundamental frequency and formants characterizing the acoustic properties of the vocal tract (as shown in Figure 10). It is worth noting that the harmonic corresponding dominates such a spectrum to the fundamental frequency.

The basis of our method is the assumption that the best fit between the amplitude Fourier spectra of the acoustic signal and the spectra derived from the glottovibrogram will occur when the sequence of segmented images reflecting vocal folds movements is represented by the harmonic identical to the frequency of vibration of the vocal folds, i.e., the fundamental frequency. In the case of dominance of other frequencies in the segmented images, i.e., different from the fundamental frequency, large values of the cost function

(5) were obtained, indicating incorrect segmentation of the vocal folds, i.e., detection of laryngeal anatomical structures that do not represent the movement of the vocal fold edges.

We recognize that the study is not without its limitations: the method has only been tested not on a large number of recordings, and only one type of voice disorder involving glottal insufficiency was considered in the study. However, the video material was carefully selected under the supervision of phoniatricians, who selected the most important and representative cases for our study. At present, due to the limitations of COVID, we cannot collect video material on a larger scale that could include representative groups, particularly in terms of gender, age, and health status.

Moreover, it should be noted that it is possible to develop image segmentation methods other than ours that might give even better segmentation results. Our primary intention was rather to show the potential of our original approach to the problem of segmenting images of moving objects for those cases where other sources of data of different modalities are available and can be used to optimize the image segmentation process.

In subjects with voice disorders, impairments in vocal fold oscillations affect the acoustic quality of their voice. It should be noted that there are three main vocal fold dynamical features that foster normophonic/euphonic voice: vocal fold oscillations are assumed to be: (1) symmetric, (2) periodic, and (3) exhibit a closed state during oscillations [6]. Incomplete glottal closure during the phonatory function of the larynx is associated with vocal fatigue and a breathy voice. However, it is assumed that slightly insufficient dorsal glottal closure should be regarded as normal, particularly in women [49]. The study confirms that three of the examined women with normophonic voice observed incomplete glottal closure in the 1/3 posterior part of the glottis with no effect on voice quality. Other types of glottal insufficiency were considered pathological. In the examined dysphonic subjects, the spindle-shaped glottal closure (type E according to the guidelines of ELS) was observed the most frequently (9 patients). Its distinctive features were the bowed shape of the vocal fold edges and a lack of glottal closure in the inter-membranaceous part of the glottis, resulting from the asthenic or atrophied vocal muscles and mucosa in the vocal folds. Such structural modifications lead to increased glottal air leakage and a breathy, weak voice. Professional vocal loading and aging (presbyopia) are considered to be the most frequent factors predisposing to this kind of glottic dysfunction [49–51]. Our results concur with their study. The patients in our study who experienced the following voice symptoms: vocal fatigue, weakness of voice, or voicelessness, mainly included professional voice users with long-term vocal loading. In turn, the two oldest male participants with no experience of professional vocal loading had clinically diagnosed presbyphonia.

Furthermore, the application of LHSV enabled the determination of the parameters OQ, CQ, SQ, and MRGA from the variations of the GGW in consecutive LHSV images. These parameters, characterizing incomplete glottal closure, have been found meaningful in assessing vocal fold vibrations, particularly vocal apparatus' insufficiency, as reported in recent studies [52]. Such evaluations are important in clinical voice assessment, e.g., in objective diagnosis and monitoring the results of an administered therapy [53]. Determination of the computed parameters makes it possible to parametrize dysfunctions of the glottis, including asthenia of the internal muscles of the larynx affecting vibrations of the VFs. One of the most relevant indexes is the OQ representing the duration of the open phase in relation to the total glottal cycle. The OQ is considered a good measure for comparing normophonic subjects with patients suffering from glottal insufficiency.

In [54], it was reported that in 96% of the patients with occupational voice disorders, the value of the OQ was on average 0.98, while the mean OQ value in the normophonic subjects took the value of 0.68. Moreover, in the dysphonic subjects, the MRGA, i.e., the ratio between the minimum and the maximum glottal area in the glottal cycle, assumed higher values than in the normophonic group (mean value 0.35 vs. 0.09) and quantified the incompleteness of the glottal closure (glottal gap). These results are consistent with our study and our earlier study [11] conducted on another group of patients. Analysis of videolaryngostroboscopic images was used for characterizing incomplete glottal clo-

sure. In another study [12], which used a high-speed video system to collect data from healthy subjects only, the mean value of OQ was 0.66 ± 0.14 in females and 0.56 ± 0.1 in males. These results again are consistent with our study. Similarly, the results for the mean value of SQ = 0.85 ± 0.21 obtained for the healthy female subjects are consistent with our results SQ = 0.77 ± 0.21. Also, in [55], in line with the studies discussed above, an elevated mean value of the OQ obtained for the patients with voice disorders was OQ = 0.84 ± 0.16 compared to healthy subjects OQ = 0.84 ± 0.16. Interestingly, and in contrast to the reviewed studies, in a recent work reporting an open platform for laryngeal high-speed videoendoscopy [29], very high values of the OQ were obtained for healthy subjects taking a mean value of 0.998. However, it was noted in [56] that singers might develop a special mechanism for voice production, the so-called laryngeal mechanism, in which OQ can reach high values exceeding 0.9.

However, it should be noted that comparisons of different studies using LHSV imaging to quantify the vibration of the vocal folds should be made with caution [57]. In a recent work [7], which provided a comprehensive review of the computer methods for quantifying vocal folds vibration, the authors concluded that it is difficult to compare the effectiveness of different methods due to the lack of publicly available databases designed for benchmarking different laryngeal image analysis methods. This is because these studies use different laryngeal image datasets, different image acquisition equipment, different assessment methods, and individual performance metrics. Meaningful comparisons between different studies require publicly available datasets and establishing a set of guidelines, preferably developed by experts from different health centers. Finally, it is worth noting a novel computational approach for spatial segmentation of high-speed laryngeal videoendoscopy images in a connected speech presented in [58]. This approach aims to develop an LHSV-based measurement of the vibratory characteristics of the vocal folds based on natural speech production, as opposed to the traditionally used phonation examination protocol.

## 7. Conclusions

In this work, we have shown that it is possible to track the location of the edges of the vibrating vocal folds in LHSVs in a way that does not require manual validation or intervention by medical personnel. In particular, we have demonstrated that image segmentation techniques can be optimized by utilizing data derived from synchronously collected acoustic recordings during sustained phonation of vowel /i:/. By transforming the glottovibrogram to the frequency domain and mapping it onto a one-dimensional spectrum, we could compare it directly to the spectrum of the acoustic signal and build a relevant cost function. The minimum of the cost function was the criterion by which the best segmentation results were identified. Independent otolaryngology-phoniatrics experts successfully validated these results.

We would like to strongly emphasize that the main value of our contribution, beyond the segmentation method itself, is the automatic technique for optimizing image segmentation methods of video images of the moving vocal folds where segmentation parameters can be defined, e.g., threshold value, size of the filtered neighborhood, etc.

It is important to note that most of the developed methods for segmentation laryngeal images require manual tuning of many parameters to obtain acceptable image segmentation results. Our method allows automatic segmentation of LHSV images without trial and error in searching for optimal segmentation parameters. We hope that the proposed approach, which considers acoustic modality, may inspire other researchers to use this image segmentation technique.

The proposed method of segmentation of LHSVs enabled automated tracking of the vocal fold edges during phonation. On that basis, we computed the corresponding glottovibrograms and GGWs that allowed us to further calculate a number of indices quantifying pathological changes in the phonatory process. The current findings support the use of this analysis method in clinical practice, which promotes LHSV as a reliable laryngeal imaging tool. The calculated indices will allow clinicians to provide reliable measures to

objectively assess laryngeal phonatory function. Quantitative assessment of vocal fold vibratory disturbances characteristic for the glottal insufficiency may improve the diagnosis of occupational voice disorders or presbyphonia. In recent decades these laryngeal diseases have attracted clinical attention due to the increasing number of professional voice users and the aging population in the modern world.

As indicated in the conclusion section, the proposed optimization procedure for the image segmentation algorithm involves a high computational cost that can, however, be mitigated by mapping the computations to GPU hardware.

Finally, we hope that the proposed method can trigger studies that will follow the proposed path and further explore the approach in which the fusion of data from LHSVs and acoustic recordings is used to optimize image analysis techniques aiding clinicians in the diagnosis and quantification of voice disorders.

**Author Contributions:** Conceptualization, methodology, and software, B.K.; medical consultations, interpretation of the results and provision of the recordings, E.N.-B.; medical consultations, interpretation of the results and provision of the recordings, W.P.; supervision, validation, and preparation of the original manuscript, P.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded in part by the National Science Centre, Poland, grant Preludium, no. UMO-2016/21/N/ST6/02612 in years 2017–2019.

**Ethical Committee Approval:** Approval for this study was granted by the Ethical Committee of the Medical University of Lodz, Poland (no. RNN/96/20/KE 08/04/2020), and all patients gave informed consent to participate in the study.

**Data Availability Statement:** The acoustic recordings and segmentation results of the LHSV images reported in this work are available at: https://tulodz-my.sharepoint.com/:f:/g/personal/pawel_strumillo_p_lodz_pl/Eni2ELdrtmBOh_ez4ul-asIBmaZr_5pcHjaYjZH8R5b_YA?e=d1OCwL (last accessed on 20 December 2021).

**Conflicts of Interest:** Authors declare no conflict of interest.

## Appendix A. Definitions of the Parameters Characterizing Vocal Folds Kinematics

Here, we define the parameters we compute to compare vocal fold kinematics for the normophonic individuals and the patients with glottal incompetence. For a more detailed description of those and other parameters characterizing the process of VF oscillation, please also refer to our earlier publication [10].

1. The Open Quotient (OQ) is the proportion of the time during which the vocal folds are open within the phonation interval [11,56–58]:

$$\mathrm{OQ} = \frac{t_{oc} + t_{cc} + t_{co}}{T} \tag{A1}$$

where: $T$—is the phonation cycle interval, $t_{oc}$—is the VFs closing phase duration, $t_{cc}$—is the duration of the vocal folds' closed phase, and $t_{co}$—is the duration of the VFs' opening phase. Note that for glottal incompetence, this quotient equals unity because the duration of the closure of the VFs $t_{cc} = 0$ (see Figure A1b). On the other hand, for the normophonic voice, this quotient is less than unity because between the opening and closing phases, there is a non-zero $t_{cc}$ duration for which the VFs are closed (see Figure A1a).

2. The Closing Quotient (CQ) is the ratio between the duration of the closing phase and the phonation cycle interval [54,55]:

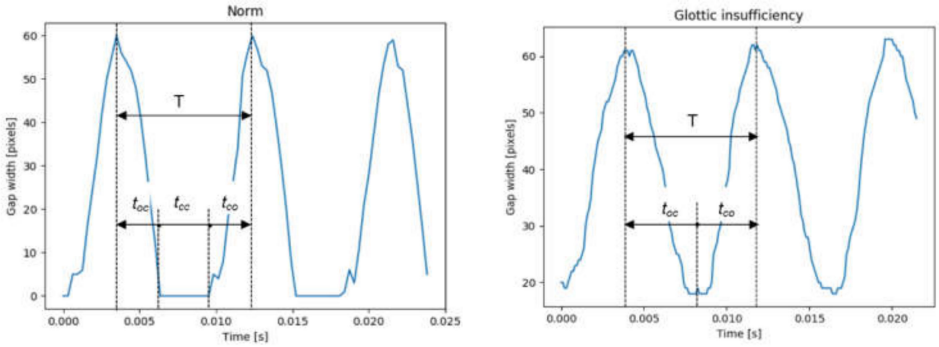
$$\mathrm{CQ} = \frac{t_{oc}}{T} \tag{A2}$$

3. The Speed Quotient (SQ) is the ratio between the duration of the opening phase ($t_{co}$) and the closing phase ($t_{oc}$) [11,55]:

$$\mathrm{SQ} = \frac{t_{co}}{t_{ocj}} \tag{A3}$$

4. The Minimal Relative Glottal Area (MRGA) is the ratio between the minimum area of the glottal area (for the closure of the VFs) to the maximum area of the glottal area (for the maximum opening of the VFs):

$$\mathrm{MRGA} = \frac{A_{min}}{A_{maxj}} \tag{A4}$$




(**a**) (**b**)

**Figure A1.** Example glottal width waveforms for a normophonic subject (**a**) and a patient with glottal insufficiency (**b**). The indicated time intervals are explained in the definitions of the parameters.

## References


1. Carding, P. Occupational voice disorders: Is there a firm case for industrial injuries disablement benefit? *Logop. Phoniatr. Vocol.* **2007**, *32*, 47–48. [CrossRef] [PubMed]
2. Woo, P. Objective Measures of Stroboscopy and High-Speed Video. *Adv. Otorhinolaryngol.* **2020**, *85*, 25–44. [CrossRef] [PubMed]
3. Behlau, M. The 2016 G. Paul Moore Lecture: Lessons in Voice Rehabilitation: Journal of Voice and Clinical Practice. *J. Voice* **2019**, *33*, 669–681. [CrossRef] [PubMed]
4. De Jong, F.I.C.R.S.; Kooijman, P.G.C.; Thomas, G.; Huinck, W.J.; Graamans, K.; Schutte, H.K. Epidemiology of voice problems in Dutch teachers. *Folia Phoniatr. Logop.* **2006**, *58*, 186–198. [CrossRef]
5. Dejonckere, P.H.; Bradley, P.; Clemente, P.; Cornut, G.; Crevier-Buchman, L.; Friedrich, G.; Van De Heyning, P.; Remacle, M.; Woisard, V. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). *Eur. Arch. Otorhinolaryngol.* **2001**, *258*, 77–82.
6. Patel, R.R.; Awan, S.N.; Barkmeier-Kraemer, J.; Courey, M.; Deliyski, D.; Eadie, T.; Paul, D.; Švec, J.G.; Hillman, R. 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.* **2018**, *27*, 887–905. [CrossRef]
7. Andrade-Miranda, G.; Stylianou, Y.; Deliyski, D.D.; Godino-Llorente, J.I.; Henrich Bernardoni, N. Laryngeal Image Processing of Vocal Folds Motion. *Appl. Sci.* **2020**, *10*, 1556. [CrossRef]
8. Chang, M.X.; Leonardus Willems, F. Human Speech Processing Apparatus for Detecting Instants of Glottal Closure. U.S. Patent No. 6,470,308, 22 October 2002.
9. Grygiel, J.; Strumiłło, P.; Niebudek-Bogusz, E. Application of Mel Cepstral processing and Support Vector Machines for diagnosing vocal disorders from voice recordings. In Proceedings of the Signal Processing Algorithms, Architectures, Arrangements, and Applications, SPA 2011, Poznan, Poland, 29–30 September 2011; pp. 1–4.
10. Mehta, D.D.; Van Stan, J.H.; Hillman, R.E. Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 659–668. [CrossRef]
11. Niebudek-Bogusz, E.; Kopczynski, B.; Strumillo, P.; Morawska, J.; Wiktorowicz, J.; Sliwinska-Kowalska, M. Quantitative assessment of videolaryngostroboscopic images in patients with glottic pathologies. *Logop. Phoniatr. Vocology* **2017**, *42*, 73–83. [CrossRef]
12. Lohscheller, J.; Švec, J.G.; Dollinger, M. Vocal fold vibration amplitude open quotient speed quotient and their variability along glottal length: Kymographic data from normal subjects. *Logop. Phoniatr. Vocology* **2013**, *38*, 182–192. [CrossRef]
13. Sujecka, J.; Świech, W.; Poryzała, P.; Borowska-Terka, A. A prototype system for quantitative assessment of voice fatigue: Design for accessibility. In *Ergonomics for People with Disabilities*; De Gruyter: Berlin, Germany, 2018. [CrossRef]

14. Lin, J.Z.; Espinoza, V.M.; Marks, K.L.; Zañartu, M.; Mehta, D.D. Improved Subglottal Pressure Estimation from Neck-Surface Vibration in Healthy Speakers Producing Non-Modal Phonation. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 449–460. [CrossRef] [PubMed]
15. Qin, X.; Wang, S.; Wan, M. Improving Reliability and Accuracy of Vibration Parameters of Vocal Folds Based on High-Speed Video and Electroglottography. *IEEE Trans. Biomed. Eng.* **2009**, *56*, 1744–1754. [CrossRef] [PubMed]
16. Bonilha, H.S.; Deliyski, D.D.; Whiteside, J.P.; Gerlach, T.T. Vocal fold phase asymmetries in patients with voice disorders: A study across visualization techniques. *Am. J. Speech-Lang. Pathol.* **2012**, *21*, 3–15. [CrossRef]
17. Gaber, A.G.H.; Liang, F.Y.; Yang, J.S.; Wang, Y.J.; Zheng, Y.Q. Correlation among the Dysphonia Severity Index (DSI), the RBH voice perceptual evaluation, and minimum glottal area in female patients with vocal fold nodules. *J. Voice* **2011**, *28*, 20–23. [CrossRef] [PubMed]
18. Švec, J.G.; Sundberg, J.; Hertegård, S. Three registers in an untrained female singer analyzed by videokymography, strobolaryngoscopy and sound spectrography. *J. Acoust. Soc. Am.* **2008**, *123*, 347–353. [CrossRef]
19. Švec, J.G.; Schutte, H.K. Videokymography: High-speed line scanning of vocal fold vibration. *J. Voice* **1996**, *10*, 201–205. [CrossRef]
20. Deliyski, D.D.; Hillman, E.R.; Mehta, D.D. Laryngeal High-Speed Videoendoscopy: Rationale and Recommendation for Accurate and Consistent Terminology. *J. Speech Lang. Hear. Res.* **2015**, *58*, 1488–1492. [CrossRef]
21. Zacharias, S.R.C.; Deliyski, D.D.; Gerlach, T.T. Utility of Laryngeal Highspeed Videoendoscopy in Clinical Voice Assessment. *J. Voice* **2017**, *32*, 216–220. [CrossRef]
22. Hewavitharanage, S.; Gubbi, J.; Thyagarajan, D.; Lau, K.; Palaniswami, M. Estimation of vocal fold plane in 3D CT images for diagnosis of vocal fold abnormalities. In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 3105–3108.
23. Titze, I.R. The Myoelatic Aerodynamic Theory of Phonation, Iowa City: National Center for Voice and Speech. 2006. Available online: https://www.worldcat.org/title/myoelastic-aerodynamic-theory-of-phonation/oclc/79872494 (accessed on 25 December 2021).
24. Gutierrez-Arriola, M.; Osma-Ruiz, V.; hon, N.S.; Godino-Llorente, J.I.; Fraile, R.; Arias-Londono, J.D. Segmentation of the Glottal Space from Laryngeal Images using the Watershed Transform. *Comput. Med. Imaging Graph.* **2008**, *32*, 193–201.
25. Skalski, A.; Zielinki, T.; Deliyski, D. Analysis of Vocal Folds Movement in High Speed Videoendoscopy Based on Level Set Segmentation and Image Registration. In Proceedings of the 2008 International Conference on Signals and Electronic Systems Krakow, Kraków, Poland, 14–17 September 2008; pp. 223–226.
26. Koç, T.; Çiloğlu, T. Automatic Segmentation of High Speed Video Images of Vocal Folds. *J. Appl. Math.* **2014**, *2014*, 818415. [CrossRef]
27. Sielska-Badurek, E.M.; Jędra, K.; Sobol, M.; Niemczyk, K.; Osuch-Wójcikiewicz, E. Laryngeal stroboscopy—Normative values for amplitude, open quotient, asymmetry and phase difference in young adults. *Clin. Otolaryngol.* **2019**, *44*, 158–165. [CrossRef] [PubMed]
28. Barbalata, C.; Mattos, L.S. Laryngeal Tumor Detection and Classification in Endoscopic Video. *IEEE J. Biomed. Health Inform.* **2016**, *20*, 322–332. [CrossRef] [PubMed]
29. Kist, A.M.; Dürr, S.; Schützenberger, A.; Schützenberger, A.; Döllinger, M. OpenHSV: An open platform for laryngeal high-speed videoendoscopy. *Sci. Rep.* **2021**, *11*, 13760. [CrossRef] [PubMed]
30. Díaz-Cádiz, M.E.; Peterson, S.D.; Galindo, G.E.; Espinoza, V.M.; Motie-Shirazi, M.; Erath, B.D.; Zañartu, M. Estimating Vocal Fold Contact Pressure from Raw Laryngeal High-Speed Videoendoscopy Using a Hertz Contact Model. *Appl. Sci.* **2019**, *9*, 2384. [CrossRef]
31. Andrade-Miranda, G.; Godino-Llorente, J.I. ROI detection in high speed laryngeal images. In Proceedings of the IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, China, 29 April–2 May 2014; pp. 477–480. [CrossRef]
32. Gloger, O.; Lehnert, B.; Schrade, A.; Völzke, H. Fully Automated Glottis Segmentation in Endoscopic Videos Using Local Color and Shape Features of Glottal Regions. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 795–806. [CrossRef]
33. Shi, T.; Kim, H.J.; Murry, T.; Woo, P.; Yan, Y. Tracing vocal fold vibrations using level set segmentation method. *Int. J. Numer. Methods Biomed. Eng.* **2015**, *31*, e02715. [CrossRef]
34. Schenk, F.; Aichinger, P.; Roesner, I.; Urschler, M. Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours. *Ann. BMVA* **2015**, *2015*, 1–15.
35. Pinheiro, A.; Dajer, M.E.; Hachiya, A.; Montagnoli, A.N.; Tsuji, D. Graphical Evaluation of Vocal Fold Vibratory Patterns by High-Speed Videolaryngoscopy. *J. Voice* **2014**, *28*, 106–111. [CrossRef]
36. Fehling, M.K.; Grosch, F.; Elke Schuster, M.; Schick, B.; Lohscheller, J. Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. *PLoS ONE* **2020**, *15*, e0227791. [CrossRef]
37. Vojtech, J.M.; Cilento, D.D.; Luong, A.T.; Noordzij, J.P., Jr.; Diaz-Cadiz, M.; Groll, M.D.; Buckley, D.P.; McKenna, V.S.; Noordzij, J.P.; Stepp, C.E. Acoustic Identification of the Voicing Boundary during Intervocalic Offsets and Onsets Based on Vocal Fold Vibratory Measures. *Appl. Sci.* **2021**, *11*, 3816. [CrossRef]
38. Pietruszewska, W.; Just, M.; Morawska, J.; Malinowski, J.; Hoffman, J.; Racino, A.; Barańska, M.; Kowalczyk, M.; Niebudek-Bogusz, E. Comparative analysis of high-speed videolaryngoscopy images and sound data simultaneously acquired from rigid and flexible laryngoscope: A pilot study. *Sci. Rep.* **2021**, *11*, 20480. [CrossRef] [PubMed]

39. Li, Z.; Wilson, A.; Sayce, L.; Avhad, A.; Rousseau, B.; Luo, H. Numerical and Experimental Investigations on Vocal Fold Approximation in Healthy and Simulated Unilateral Vocal Fold Paralysis. *Appl. Sci.* **2021**, *11*, 1817. [CrossRef] [PubMed]
40. Ismail, M.A.; Deshmukh, S.; Singh, R. Detection of COVID-19 Through the Analysis of Vocal Fold Oscillations. In Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, USA, 6–11 June 2021; pp. 1035–1039. [CrossRef]
41. Kopczynski, B.; Strumillo, P.; Just, M.; Niebudek-Bogusz, E. Acoustic Based Method for Automatic Segmentation of Images of Objects in Periodic Motion: Detection of vocal folds edges case study. In Proceedings of the Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA), Xi'an, China, 7–10 November 2018; pp. 1–6. [CrossRef]
42. Gonzales, R.C.; Woods, R.E. *Digital Image Processing*, 4th ed.; Pearson Education International: London, UK, 2017.
43. Bengio, Y.; Goodfellow, I.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
44. DiagNova Technologies Company. Available online: http://www.diagnova.pl (accessed on 30 July 2021).
45. Ahmad, K.; Yan, Y.; Bless, D.M. Vocal fold vibratory characteristics in normal female speakers from high-speed digital imaging. *J. Voice* **2012**, *26*, 239–253. [CrossRef] [PubMed]
46. Yamauchi, A.; Imagawa, H.; Yokonishi, H.; Nito, T.; Yamasoba, T.; Goto, T.; Takano, S.; Sakakibara, K.I.; Tayama, N. Evaluation of vocal fold vibration with an assessment form for high-speed digital imaging: Comparative study between healthy young and elderly subjects. *J. Voice* **2012**, *26*, 742–750. [CrossRef]
47. Wakita, H.; Fant, G. Toward a better vocal tract model. *Speech Transm. Lab. Q. Prog.* **1978**, *19*, 9–29.
48. Flanagan, J. *Speech Analysis Synthesis and Perception 1965*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1971.
49. Yamauchi, A.; Imagawa, H.; Sakakibara, K.-I.; Yokonishi, H.; Nito, T.; Yamasoba, T.; Tayama, N. Age- and gender-related difference of vocal fold vibration and glottal configuration in normal speakers: Analysis with glottal area waveform. *J. Voice* **2014**, *28*, 525–531. [CrossRef]
50. Yamauchi, A.; Yokonishi, H.; Imagawa, H.; Sakakibara, K.I.; Nito, T.; Tayama, N.; Yamasoba, T. Vocal Fold Vibration in Vocal Fold Atrophy: Quatitative Analysis with High Speed-Digital Imaging. *J. Voice* **2015**, *29*, 1–8. [CrossRef]
51. Calcinoni, O.; Niebudek-Bogusz, E. *Occupational Voice. Diagnosis and Treatment of Voice Disorders*, 4th ed.; Rubin, J., Sataloff, R., Korovin, G., Eds.; Plural Publishing: San Diego, CA, USA, 2014; pp. 735–762.
52. Schlegel, P.; Stingl, M.; Kunduk, M.; Kniesburges, S.; Bohr, C.; Döllinger, M. Dependencies and Ill-designed Parameters within High-speed Videoendoscopy and Acoustic Signal Analysis. *J. Voice* **2018**, *33*, 811.e1–811.e12. [CrossRef]
53. Kosztyła-Hojna, B.; Zdrojkowski, M.; Duchnowska, E. Application of the HRES 5562 Camera Using the HSDI Technique in the Diagnosis of Glottal Insufficiencies in Teachers. *J. Voice* **2020**. [CrossRef]
54. Powell, M.E.; Deliyski, D.D.; Hillman, R.E.; Zeitels, S.M.; Burns, J.A.; Mehta, D.D. Comparison of videostroboscopy to stroboscopy derived from high-speed videoendoscopy for evaluating patients with vocal fold mass lesions. *Am. J. Speech-Lang. Pathol.* **2016**, *25*, 576–589. [CrossRef]
55. Yamauchi, A.; Imagawa, H.; Yokonishi, H.; Sakakibara, K.-I.; Tayama, N. Multivariate Analysis of Vocal Fold Vibrations on Various Voice Disorders Using High-Speed Digital Imaging. *Appl. Sci.* **2021**, *11*, 6284. [CrossRef]
56. Henrich, N.; D'Alessandro, C.; Doval, B.; Castellengo, M. Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *J. Acoust. Soc. Am.* **2005**, *117*, 1417–1430. [CrossRef] [PubMed]
57. Ikuma, T.; Kunduk, M.; McWorther, A.J. Objective quantification of pre- and postphonosurgery vocal fold vibratory characteristics using high-speed videoendoscopy and a harmonic waveform model. *J. Speech Lang. Hear. Res.* **2014**, *57*, 743–757. [CrossRef] [PubMed]
58. Yousef, A.M.; Deliyski, D.D.; Zacharias, S.R.C.; de Alarcon, A.; Orlikoff, R.F.; Naghibolhosseini, M. Spatial Segmentation for Laryngeal High-Speed Videoendoscopy in Connected Speech. *J. Voice* **2020**. [CrossRef] [PubMed]