Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 269–274

DOI: 10.15439/2015F276
ACSIS, Vol. 5

Computer based quantification of normal and pathological vocal folds phonatory processes from laryngovideostroboscopy

Bartosz Kopczyński
Institute of Electronics, Lodz
University of Technology,
ul. Wolczanska 211/215, 90-924
Lodz, Poland, e-mail:
bartosz.michal.k@gmail.com

Paweł Strumiłło
Institute of Electronics, Lodz
University of Technology,
ul. Wolczanska 211/215, 90-924
Lodz, Poland,
e-mail: pawel.strumillo@p.lodz.pl

Ewa Niebudek-Bogusz
Department of Audiology and
Phoniatrics, The Nofer Institute
of Occupational Medicine,
ul. Teresy 8, 91-348 Lodz, Poland
e-mail: ebogusz@imp.lodz.pl

***Abstract* — Medical imaging techniques offer novel visualization and analysis methods of the vocal folds during phonation and automatic computation of indices aiding the phoniatrist in a more precise diagnosis of voice disorders. The aim of this study is to apply *computer vision algorithms for qualitative and quantitative analysis of vocal folds' vibrations. Videostroboscopic examinations of the larynx were carried out for 30 patients. Image pre-processing and image segmentation algorithms were applied to compute the glottis area during phonation.* The glottovibrograms which are spatio-temporal visualizations of the vibrating vocal folds were also built. The proposed indices allow for a quantitative and comparative analysis of normal and disordered phonatory processes. The conducted pilot study has confirmed the validity of the computer aided imaging methods for the qualitative and quantitative analysis of the videostroboscopic images of the phonatory motions of the vocal folds.**

I. Introduction

Early diagnosis of occupational voice disorders is becoming one of the priorities in public health in Poland and in other countries of the European Union. Currently, European standards emphasize the need for a comprehensive assessment of voice disorders, including the assessment of the larynx function during the phonatory tests [1]. The test that allows the specialist to accurately assess the condition of the voice organ and is recognized as the gold standard is the laryngovideostroboscopy (LVS) [2]. The aim of our study is to apply computer image processing and analysis methods for quantification of vocal folds' phonatory activity by examining sequences of LVS images. In this communication we focus on examining videostroboscopic images of normophonic individuals with healthy vocal folds and patients with diagnosed nodules.

II. Related Works

The standardization of automatic segmentation of vocal folds videostroboscopic images remains an unsolved problem since 1995 [3]. There have been several automatic segmentation methods of the vocal folds images proposed [4]-[6]. Many of the developed methods and algorithms are designed for specific image recording conditions and work properly only for local databases containing videos collected in a particular institute, hospital or health center. It has turned out to be a very difficult problem to work out an algorithm which would give satisfactory results for every given video presenting the vibrating vocal folds. Authors in [7] proposed a method based on supervised thresholding methods by applying the Fourier descriptors. This method is additionally combined with a glottal neighborhood descriptor which specifies distance–weighted color differences between the glottis and the surroundings tissues of the vocal folds. The additional knowledge of local color distributions increases the recognition quality.

The key task preceding the image analysis methods of the glottis is segmenting out the space between the vocal folds, termed the glottal area. The most popular and straightforward image segmentation method is thresholding. The threshold value is selected on the basis of the image histogram. Local minima of the multimodal histogram designate threshold values. In the simplest case the first histogram minimum assigns a threshold value that distinguishes the image regions belonging to the space between the vocal folds from the regions representing the background (glottis environment). More sophisticated threshold methods utilize the neighborhood information, adaptively adjusting the threshold value [8] or including wavelet transformation [9]. There are several methods which work properly for certain image classes i.e. region based, model based and their combinations. More complex methods are based on the so called active contour models. Active contours are energy minimizing flexible splines guided by hypothetical external forces influenced by the image content and internal forces arising from the curvature and continuity of the nodes forming the contour. Note, however, that the contour must be properly initialized near the object of interest [Marendic et al. 2001].

III. Materials and methods

The general block diagram of the developed image processing and analysis algorithms for analysis of laryngostroboscopic images is shown in Fig. 1.

978-83-60810-66-8/$25.00©2015, IEEE

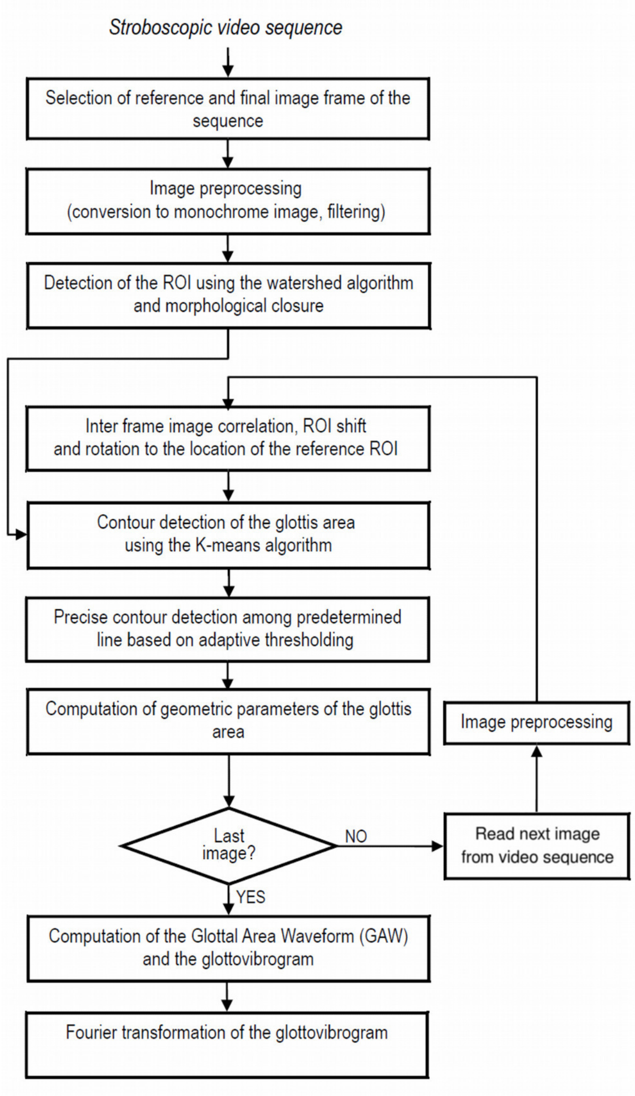
Stroboscopic video sequence
Selection of reference and final image frame of the sequence
Image preprocessing (conversion to monochrome image, filtering)
Detection of the ROI using the watershed algorithm and morphological closure
Inter frame image correlation, ROI shift and rotation to the location of the reference ROI
Contour detection of the glottis area using the K-means algorithm
Precise contour detection among predetermined line based on adaptive thresholding
Computation of geometric parameters of the glottis area
Image preprocessing
Last image?
NO
Read next image from video sequence
YES
Computation of the Glottal Area Waveform (GAW) and the glottovibrogram
Fourier transformation of the glottovibrogram

Fig 1. Schematic of the algorithm for qualitative and quantitative analysis of the vocal folds' phonation from videostroboscopy

The algorithm was developed after numerous consultations with phoniatrists working in the Department of Audiology and Phoniatrics, The Nofer Institute of Occupational Medicine in Lodz, Poland. The algorithm allows the user to select the analyzed film and indicate the phonation region of the glottis. Then the algorithm proceeds with automatic computations of the following quantities and representations:

- fundamental frequency of the vocal folds' vibration and the vibration frequency at each level of the total glottis length,
- geometric and kinetic parameters of the glottal area (the space between the vocal folds) during phonation,
- relations between the opening and closing time of the glottis,
- the glottal area waveform (GAW), illustrating time variations of the glottal area,
- the glottovibrogram which is a time-space representation of the glottal gap (local distances between the vocal folds at different levels of the glottis) imaged as a grayscale image (see Figs. 4, 5),
- the Fourier transform of the glottovibrogram.

Videostroboscopic examinations of the larynx were carried out for 30 patients, i.e. for 15 individuals with no voice disorders (see an example in Fig. 2) and 15 patients with diagnosed nodules (see Fig. 3). At the outset of the analysis the phoniatrist is asked to select the reference sequence of images for analysis i.e. the starting and ending frame from the videostroboscopic film. This is the only user-dependent step in the proposed image analysis procedure. After defining the reference sequence the program determines the region of interest and adjusts processing parameters to the selected regions of the analyzed video and reduces the distortions due to movements of the camera versus the glottis.

The aim of the second step is to apply general image processing methods for image enhancement. Color images in the LVS sequence are converted from an RGB to a monochrome image format, by applying the following weighted average of the color components:

$$I=0.299\,R+0.587\,G+0.114\,B \qquad (1)$$

in which I is the level of intensity in the monochrome image and R, G, B are the values of the red, green and blue components of the color image correspondingly. Digital image filtering is then applied to reduce noise and artifacts due to unwanted reflections and compression distortions. This step allows the residual adaptation of parameters for the varying shape and location of the region of interest (ROI) which undergoes the following distortions: non-uniform and time-varying illumination, loss of sharpness due to the vapor covering the camera lens, varying distance relative to the vocal folds and loss of camera focus. Preprocessing of the region of interest additionally improves results of the thresholding step.

Then, an automatic analysis of the LVS images starts with the detection of the ROI, i.e. a rectangular region containing the examined part of the glottis. Similarly to the approach adopted in [10] this was improved with morphological operations (erosion and closure) and basic filtering methods comprising median and lowpass filtering. Finally, the watershed image segmentation algorithm is employed to roughly determine the glottal area ROI. The watershed transformation enables extraction of areas in which there is continuity in terms of image features such as brightness or color. This transformation method is considered as one of the most effective methods of image segmentation [11]. It yields well designated areas of the image with clearly defined outlines. The behavior of this method lies in the distribution of the image gradient which separates areas of the image assigned to the seeds along the thresholded value of the gradient.

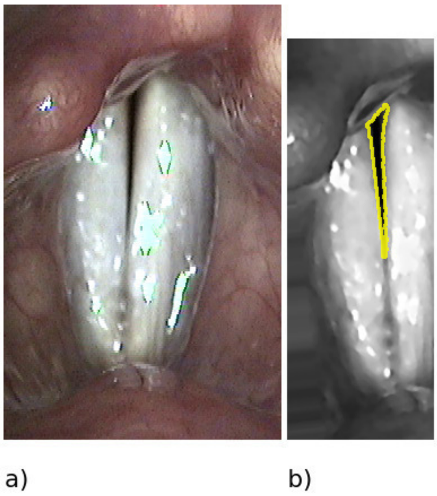

Fig 2. Stroboscopic images of the normophonic vocal folds (i.e. no diagnosed abnormalities): input image a), segmentation result b)

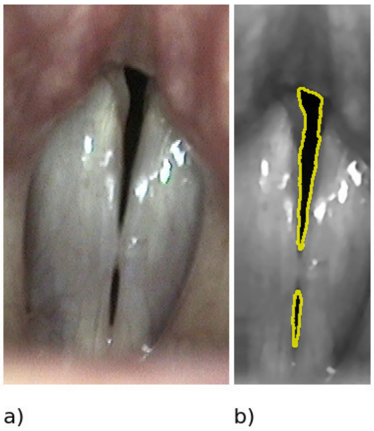

Fig 3. Stroboscopic images of the vocal folds with nodules: input image a), segmentation result b)

While being a robust segmentation method, the watershed algorithm features poor precision in determination of the glottis boundary. Thus, this segmentation method is adopted for the first frame of the image sequence only and serves as a reference location of the ROI (see block diagram in Fig. 1)

After defining the coarse position of the ROI obtained by the watersheds, the precise contour, i.e. the boundaries of the vocal folds, is being extracted with the use of the *K*-means algorithm. The *K*-means algorithm is an iterative data clustering method in which the image content is partitioned into *K*-separate regions, in which each pixel of the image is assigned to the region with the nearest distance to the centroid of the region, i.e. according to the criterion:

$$minimum(S)=\sum_{i=1}^{K}\sum_{x\in S_i}||x-u_i||^2 \qquad (2)$$

where: K – is the number of segmentation regions, x – is grey level value of a pixel, S_i – is the i-th segmentation region, and u_i – is the centroid of pixels within S_i [12]. The main advantage of the *K*-means algorithm is its computational simplicity yielding such a partitioning result that the shapes of the clusters are maximally compact.

In our application, at the first step we define two ($K = 2$) clusters of which the one with a lower mean value (darker) is assigned to the glottal area and the other to the remaining regions of the larynx image.

In videos taken from patients with vocal folds nodules there may occur a circumstance in which the glottal area is separated into two regions (see an image of vocal fold with nodules in Fig. 3b).

Further, the method is improved by applying an adaptive thresholding method for subregions of the ROI that contain the glottal area. Firstly an ellipse is fitted to the so far detected area that is based on a method which automatically sets a reference line along the hitherto designated area. Ellipse fitting is achieved by applying a built in OpenCV library function named "fitEllipse" [13]. After defining the center point of the fitted ellipse and its angle, a reference semi-major axis line is determined.

During the videostroboscopic examination of the larynx there is a continuous movement of the camera versus the glottis. An efficient way to eliminate unwanted motion from the video is to move each ROI of the image by a vector calculated for each video frame. ROI displacement vector can be determined by computing the matrix of correlation coefficients between adjacent image frames in the videostroboscopic film. In all subsequent image frames the detected glottal area is shifted to this reference location by the estimated displacement vector. The criterion of the maximum of the two-dimensional correlation function is applied for the best positioning of the consecutive image frames versus the first reference frame. By these means, in the preprocessed image sequence, the glottal area is positioned in a fixed location which significantly facilitates further image analysis procedures. The cross-correlation of adjacent images was determined by using a built-in OpenCV library function which calculates the matrix of correlation coefficients [14]:

$$r_{x,y}=\frac{\sum_{x,y}[P(x',y')-\overline{P}]\cdot[N(x+x',y+y')-\overline{N}]}{\sqrt{\sum_{x,y}[P(x',y')-\overline{P}]^2\cdot\sum_{x,y}[N(x+x',y+y')-\overline{N}]^2}}$$

(3)

where:

- $r_{x,y}$ - correlation matrix,
- $P(x,y)$ - ROI from the current video frame,
- $\overline{P}$ - mean value of pixel intensities in an image,
- $N(x,y)$ - next video frame image,
- $\overline{N}$ - mean value of pixel intensities in an image.

The correlation matrix indicates the level of the linear relationship between P corresponding pixels in images and N.

After applying the multi-stage segmentation methods (the watershed algorithm followed by the *K*-means and adaptive thresholding) and removing image distortions due to movement of vocal folds the GAW and the glottovibrogram can be built (see Figs. 4, 5 and a more detailed explanation in Fig. 8).

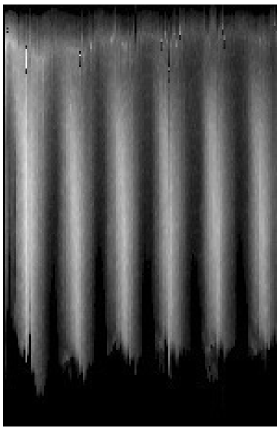

Fig 5. Glottovibrogram of the vibrating vocal folds of a healthy individual

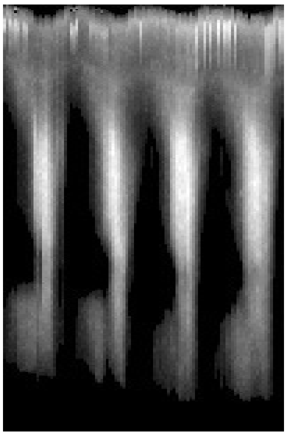

Fig 4. Glottovibrogram of the vibrating vocal folds of the patient with diagnosed vocal nodules

Such a representation termed, the glottovibrogram is a space-time image of the video sequence presenting vocal folds movements. The horizontal axis in the glottovibrogram is time (*t*) and the vertical axis represents the level (*l*) along the glottis (from the anterior commissure – bottom of the image to the posterior commissure – top of the image); each pixel value *g*(*t, l*), i.e. its brightness in the glottovibrogram, represents the value of width *g* of the space between the vocal folds (glottal gap) computed at level *l* of the glottis for time instance *t*. The GAW illustrates time changes of the instantaneous values of the glottal area.

A single phonatory cycle shown in the glottovibrogram or the associated GAW corresponds on average from 20 to 30 images taken by the stroboscopic camera. It is essential to gather at least 20 images for one cycle to determine the kinetic parameters of the vocal folds (opening, closing time and their ratios). Approximately 4 cycles are analyzed for each video sequence, so the glottovibrogram has an average length of 100 pixels. The stroboscopic films are recorded with a sampling rate of 25 frames per second, so the average length of the recording is about 4 seconds. The fundamental frequency of the vocal fold oscillation is provided by the videostroboscopic apparatus. This information enables to correctly scale the frequency axis of the Fourier transform of the glottovibrogram and visualize oscillation frequencies of the vibrating vocal folds (Fig. 6-7).

Thanks to the strobing frequency the information obtained from the videostroboscopic camera we can quantitatively estimate the frequencies taking part in the phonation. It is noteworthy that we analyze “virtual frequencies” which are aliases of the true frequencies of vocal folds' vibration, typically in the range of 150Hz (for men) and 250Hz (for women), reaching peak values of 450 Hz for singers.

The Fourier transformation of the glottovibrogram was calculated by applying the following equation:

$$G(s,l)=\frac{1}{N}\sum_{l=0}^{N-1} g(t,l)e^{-j\frac{2\pi l}{N}s} \tag{4}$$

$$|G(s,l)|=\sqrt{\Re^2(G(s,l))+\Im^2(G(s,l))} \tag{5}$$

where:

- $g(t,l)$ is the $l-th$ row of the glottovibrogram (y-axis, i.e. glottis level), t determines the column (x-axis, i.e. discrete time),

- $G(s,l)$ is the Fourier transformed glottovibrogram, determining the frequencies,

- N is the length of the glottovibrogram,

- j is the imaginary unit.

After calculating the magnitudes of the transformation Eq. (5) the phoniatrist is able to gain access to the distribution of frequencies taking part of the vocal folds phonation. In case of the vocal nodules the pathology manifests itself in the form of a second harmonic frequency, significantly blurring the first amplitude peak.

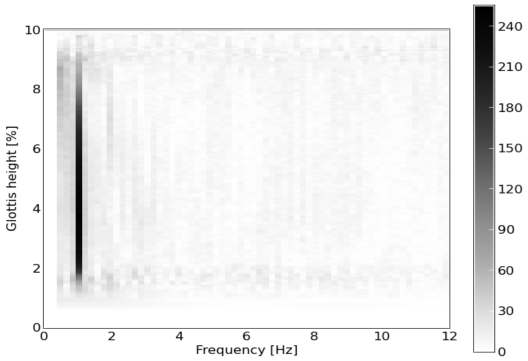

Fig 6. Fourier amplitude spectrum of the glottovibrogram of a healthy patient

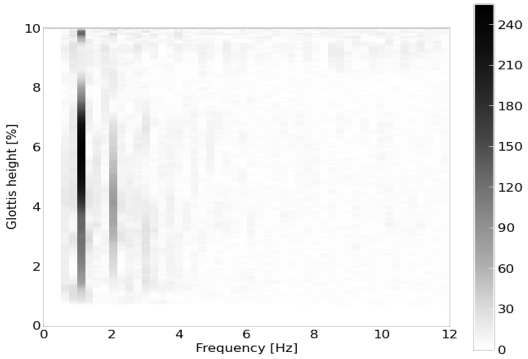

Fig 7. Fourier amplitude spectrum of the glottovibrogram of a patient with vocal nodules

IV. Phonatory indices

Even after the correct segmentation and transformation process there is still an ambiguity in the relevance of different visualizations representing the phonation process. In this paper we propose a number of indices for quantifying the phonatory processes of the healthy vocal folds and folds with diagnosed nodules.

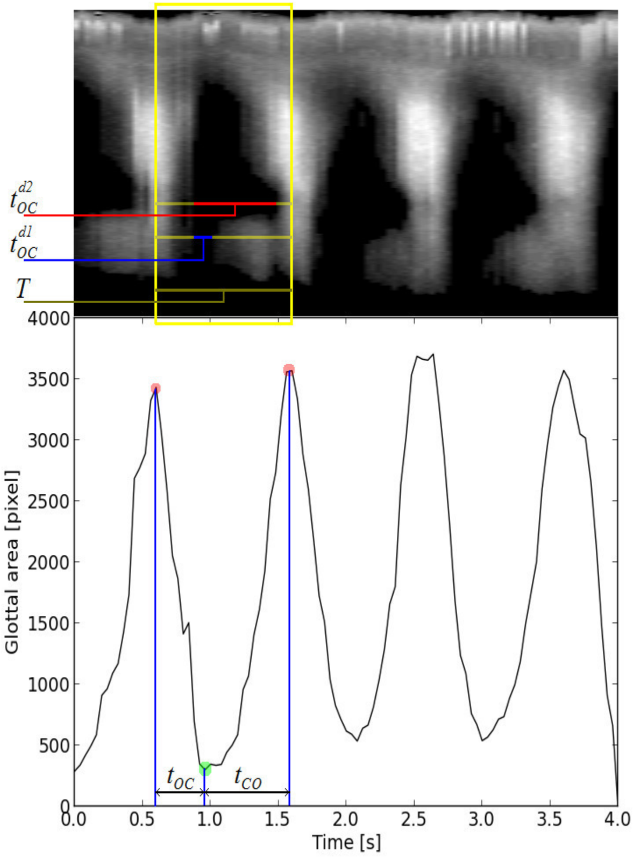

Fig 8. The glottovibrogram (upper panel) and the glottal area waveform (lower panel) with the indicated time intervals used in calculations of the phonatory indices (eq. (6, 7))

Table I.
Comparison of phonatory indices computed for the two studied groups of patients

	d1	d2	d3	CI	SI
Control group	0.00 ± 0.00	0.88 ± 0.99	2.25 ± 2.65	0.10 ± 0.09	-0.08 ± 0.30
Vocal nodules	2.10 ± 2.10	1.85 ± 1.96	4.42 ± 2.72	-0.48 ± 0.26	-0.14 ± 0.13
Significance P (ANOVA)	0.000	0.000	0.026	0.000	0.390

The *CI* (the closing index) is defined as follows (see also Fig. 8):

$$CI=\frac{t_{OC}^{d2}-t_{OC}^{d1}}{T} \tag{6}$$

where:

T – total cycle time,

t_{OC}^{d2} – is the closed time for the glottis gap measured at the level equal to 50% of the total length in case of normophonic patients, in case of patients with vocal nodules the value is measured for the level of the maximal closure (apparently the place where the vocal nodules occur),

t_{OC}^{d1} – is the closed time of the glottis gap measured at 25% level of the glottis length for normophonic patients, and at the level placed between the maximal closure point and the posterior part of the glottis.

The speed index *SI* is defined by the following equation:

$$SI=\frac{t_{CO}-t_{OC}}{t_O} \tag{7}$$

where:

t_O – opened interval,

t_{CO} – vocal fold closed phase,

t_{OC} – vocal fold closing phase.

In Table I the computed values of the indices proposed in this study are listed, where *d1*, *d2*, *d3* are the levels along the glottis length at which the indices were computed. Location of these levels are at the particular percentages of the total glottis length, correspondingly: 25%, 50% and 75%. The *SI* value in most patients assumes negative values indicating that the opening time of the folds is shorter than the closing time. However, this index yielded poor statistical significance values ($p < 0.36$ from the ANOVA analysis of variance) to use it as a differentiation index for the pathology in question. On the other hand, the computed values of the *CI* index clearly differentiate the patients with diagnosed nodules from the healthy individuals ($p < 0.026$) The *SI* value in most patients reaches negative values which indicates that the opening time is slower than the closing time. Comparison of the calculated parameters reveal that there are pronounced between-group differences and comparable within-group values.

V. Conclusion

The main purpose of this work is to support the phoniatrist in diagnosis of the vocal folds. The computer image processing methods applied to laryngovideostroboscopic images allow for quantitative analysis of the vocal folds vibrations during phonation. It was shown that the proposed *CI* parameter proved to be viable in differentiation and quantification of the studied glottis pathology.

The program, although, in an early development stage allows for automation of the analysis process of the laryngovideostroboscopic films. Larger scale trials are required before a wider introduction of these computed image analysis techniques into the clinical practice.

References

[1] B. Kopczyński, P. Strumiłło, E. Niebudek-Bogusz. Assessment of vocal folds phonation by means of computer analysis of laryngovideostroboscopic images – a pilot study. *Otorynolaryngologia – przegląd kliniczny* (in Polish), vol. 13, no. 3 2014 pp. 139–146.

[2] P. Woo. Stroboscopy. Plural Publishing, United Kingdom 2010.

[3] T. Wittenberg, U. Eysholdt. Estimation of Vocal Fold Vibrations Using Image Segmentation, Mustererkennung 1995, pp. 145-152.

[4] S-Z. Karakozoglou, N. Henrich, C. d'Alessandro, Y Stylianou. A Segmentation Scheme Based on Rayleigh Distribution Model for Extracting Glottal Waveform from High-speed Laryngeal Images.

[5] S. Z. Karakozoglou, N. Henrich, C. d'Alessandro, Y. Stylianou. Automatic glottal segmentation using local-based active contours and application to glottovibrography, Speech Communication 2012, 54: pp. 641-654.

[6] A. Méndez, E.M.Ismaili Alaoui, B. Garcia, E. Ibn-Elhaj, I. Ruiz, Glottal space segmentation from motion estimation and gabor filtering. 31st Annual International Conference of the IEEE EMBS. Minneapolis, Minnesota, USA, September 2-6, 2009.

[7] O. Gloger, B. Lehnert, A. Schrade, H. Volzke. Fully Automated Glottis Segmentation in Endoscopic Videos Using Local Color and Shape Features of Glottal Regions. IEEE Trans. Biomed. Eng. 2015, pp. 795-806.

[8] Information and Communication Technologies (WICT), 2011 World Congress on 11-14 Dec. 2011, pp. 313 - 318.

[9] Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on 30-31 March 2012 Nagapattinam, Tamil Nadu 161 - 166.

[10] V. Osma-Ruiz, JI Godino-Llorente, N. Saenz-Lechon, R. Fraile Segmentation of the glottal space from laryngeal images using the watershed transform. Comput Med Imaging Graph. 2008; 32(3): pp. 193-201.

[11] J.M. Gutierrez-Arriola, V. Osma-Ruiz, N. Saenz-Lechon, J.I. Godino-Llorente, R. Fraile, J.D. Arias-Londono, Segmentation of the glottal space from laryngeal images using the watershed transform. *Computerized Medical Imaging and Graphics*, 2008 pp.193-201.

[12] L. Dongju Liu, J. Yu. Otsu Method and K-means. Hybrid Intelligent Systems, 2009. HIS '09. Ninth International Conference, 12-14 Aug. 2009, pp. 344-349.

[13] A. W. Fitzgibbon, R.B. Fisher. A Buyer's Guide to Conic Fitting. Proc.5th British Machine Vision Conference, Birmingham, 1995, pp. 513-522.

[14] OpenCV image processing library: http://docs.opencv.org, accessed May 2015.